
jle

ISSN 2411-7390

JOURNAL
OF LANGUAGE
& EDUCATION

Volume 10 Issue 4, 2024



HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY



EDITOR-IN-CHIEF

Valery D. Solov'yev

Kazan Federal University, Russia

INVITED EDITORS

Marina I. Solnyshkina

Kazan Federal University, Kazan, Russia

Ildar Batyrshin

Instituto Politecnico Nacional, Mexico City, Mexico

EDITORIAL BOARD

Cem Balçıkanoğlu

Gazi Üniversitesi, Ankara, Turkey

Tatiana A. Baranovskaya

National Research University Higher School of Economics, Russia

Christine Coombe

Dubai Men's College, Higher Colleges of Technology, Al Ruwayyah, United Arab Emirates

Tariq Elyas

King Abdulaziz University, Jeddah, Saudi Arabia

Fan (Gabriel) Fang

Shantou University, Shantou, China

Charles Forceville

University of Amsterdam, Amsterdam, The Netherlands

Shima Ghahari

Shahid Bahonar University of Kerman, Kerman, Iran

Irina Golubeva

University of Maryland Baltimore County, Baltimore, USA

Bui Phu Hung

Ho Chi Minh City University of Food Industry, Vietnam

Raphiq Ibrahim

University of Haifa, Haifa, Israel

Andy Kirkpatrick

Griffith University, Brisbane, Australia

Iryna Lenchuk

Dhofar University, Salalah, Oman

Theresa Lillis

the Open University, Milton Keynes, United Kingdom

Callie W. Little

University of New England, Armidale, Australia,

Irshat Madyarov

American University, Washington, United States

Elena Makarova

University of Basel, Muttensz, Switzerland

Lynn Mastellotto

Free University of Bozen-Bolzano, Bolzano, Italy

Zoya G. Proshina

Lomonosov Moscow State University, Moscow, Russia

Lilia K. Raitskaya

Moscow State Institute of International Relations (MGIMO University), Moscow, Russia

Wayne Rimmer

Cambridge Active Grammar, Cambridge, United Kingdom

Grisel Sonia Salmaso

National University of Cuyo, Mendoza, Argentina

Jane Setter

University of Reading, Reading, United Kingdom

Vladimir D. Shadrinov

National Research University Higher School of Economics, Moscow, Russia

Prithvi Narayan Shrestha

the Open University, Milton Keynes, United Kingdom

Ashley Squires

New Economic School, Moscow, Russia

Dušan Stamenković

University of Niš, Niš, Serbia

Svetlana G. Ter-Minasova

Lomonosov Moscow State University Moscow, Russia

Svetlana V. Titova

Lomonosov Moscow State University, Moscow, Russia

Anatoliy N. Voronin

Russian Academy of Sciences, Institute of Psychology, Moscow, Russia

Shlomo Weber

Emeritus Professor, New Economic School, Moscow, Russia

Ha Xuan Van

Macquarie University, North Ryde, Australia

Søren Wichmann

Leiden University Centre for Linguistics, Leiden, Netherlands

Irina V. Melik-Gaykazyan

Beijing Language University, Beijing, China

Roman V. Svetlov

Tomsk State Pedagogical University, Tomsk, Russia

Galina V. Sorina

Immanuel Kant Baltic Federal University, Kaliningrad, Russia

Galina A. Suvorova

Lomonosov Moscow State University Moscow, Russia

Moscow Pedagogical State University, Moscow, Russia

EDITORIAL TEAM

Elena V. Tikhonova

Vice Editor-in-Chief, MGIMO University, Moscow, Russia

Lilia K. Raitskaya

Head of the Editorial Board, HSE University, Moscow, Russia

Armen I. Hakobyan

Book Review and Social Media Editor, MGIMO University, Russia

Marina A. Kosycheva

Website Editor, HSE University, Moscow, Russia

Lilit Beganyan

Executive Secretary, HSE University, Moscow, Russia

Alexey Iakovlev

Assistant Editor, University of California, USA

Assistant Editor, Dresden University of Technology, Germany

CONTENT

EDITORIAL

Lilia Raitetskaya, Elena Tikhonova
Appliances of Generative AI-Powered Language Tools in Academic Writing: A Scoping Review..... 5-30

RESEARCH PAPERS

Muhammad Ahmad, Usman Sardar, Farid Humaira, Ameer Iqra , Muhammad Muzzamil, Ameer Hmaza, Grigori Sidorov, Ildar Batyrshin
Hope Speech Detection Using Social Media Discourse (Posi-Vox-2024): A Transfer Learning Approach31-43

Vladimir Bochkarev, Anna Shevlyakova, Andrey Achkeev
Synchronic and Diachronic Predictors of Socialness Ratings of Words44-55

Nikita Login
Wrong Answers Only: Distractor Generation for Russian Reading Comprehension Questions Using a Translated Dataset 56-70

Dmitry Morozov, Timur Garipov, Olga Lyashevskaya, Svetlana Savchuk, Boris Iomdin, Anna Glazkova
Automatic Morpheme Segmentation for Russian: Can an Algorithm Replace Experts?..... 71-84

Sergey Pletenev
Probing the Pitfalls: Understanding SVD’s Shortcomings in Language Model Compression85-97

Valery Solovyev, Marina Solnyshkina, Andrey Ten, Nikolai Prokopyev
A BERT-Based Classification Model: The Case of Russian Fairy Tales98-111

Vladimir Starchenko, Darya Kharlamova, Elizaveta Klykova, Anastasia Shavrina, Aleksey Starchenko, Olga Vinogradova, Olga Lyashevskaya
Fighting Evaluation Inflation: Concentrated Datasets for Grammatical Error Correction Task 112-129

Mikhail Tikhomirov, Daniil Chernyshov
Facilitating Large Language Model Russian Adaptation with Learned Embedding Propagation..... 130-145

Zhengye Xu, Yixun Li, Duo Liu
Predictions of Multilevel Linguistic Features to Readability of Hong Kong Primary School Textbooks: A Machine Learning Based Exploration 146-158

THANKING OUR REVIEWERS

Our Reviewers 159-161



<https://doi.org/10.17323/jle.2024.24181>

Appliances of Generative AI-Powered Language Tools in Academic Writing: A Scoping Review

Lilia Raitskaya¹, Elena Tikhonova²

¹ Moscow State Institute of International Relations (MGIMO University)

² Peoples' Friendship University of Russia (RUDN University)

ABSTRACT

Introduction: Academic writing is getting through a transformative shift with the advent of the generative AI-powered tools in 2022. It spurred research in the emerging field that focus on appliances of AI-powered tools in academic writing. As the AI technologies are changing fast, a regular synthesis of new knowledge needs revisiting.

Purpose: Though there are scoping and systematic reviews of some sub-fields, the present review aims to set the scope of the research field of research on GenAI appliances in academic writing.

Method: The review adhered to the PRISMA extension for scoping reviews, and the PPC framework. The eligibility criteria include problem, concept, context, language, subject area, types of sources, database (Scopus), and period (2023-2024).

Results: The three clusters set for the reviewed 44 publications included (1) AI in enhancing academic writing; (2) AI challenges in academic writing; (3) authorship and integrity. The potential of AI language tools embraces many functions (text generation, proofreading, editing, text annotation, paraphrasing and translation) and provides for assistance in research and academic writing, offers strategies for hybrid AI-powered writing of various assignments and genres and improvements in writing quality. Language GenAI-powered tools are also studied as a feedback tool. The challenges and concerns related to the appliances of such tools range from authorship and integrity to overreliance on such tools, misleading or false generated content, inaccurate referencing, inability to generate author's voice. The review findings are in compliance with the emerging trends outlined in the previous publications, though more publications focus on the mechanisms of integrating the tools in AI-hybrid writing in various contexts. The discourse on challenges is migrating to the revisiting the concepts of authorship and originality of Gen AI-generated content.

Conclusion: The directions of research have shown some re-focusing, with new inputs and new focuses in the field. The transformation of academic writing is accelerating, with new strategies wrought in the academia to face the challenges and rethinking of the basic concepts to meet the shift. Further regular syntheses of knowledge are essential, including more reviews of all already existent and emerging sub-fields.

KEYWORDS

academic writing, artificial intelligence (AI), generative artificial intelligence (GenAI), AI-powered language tools, authorship

Citation: Raitskaya, L., & Tikhonova, E. (2024). Appliances of generative AI-powered language tools in academic writing: A scoping review. *Journal of Language and Education*, 10(4), 5-30. <https://doi.org/10.17323/jle.2024.24181>

Correspondence:
Elena Tikhonova,
etikhonova@hse.ru

Received: November 1, 2024

Accepted: December 16, 2024

Published: December 30, 2024

INTRODUCTION

Academic writing is an important component of knowledge production as well as scientific and academic communication (Tusting et al., 2019). It serves as an internationally admitted convention embracing activity, cognitive processes, "language rules, communication norms"

(Nguyen et al., 2024), "structured expression of ideas, data-driven arguments, and logical reasoning" (Khalifa & Albadawy, 2024) and reporting of contributions to science and synthesis of knowledge. Higher education and science are essentially dependent of academic writing (Coffin et al., 2003). The latter embodies written communication within the



academia. The recent two years academic writing is going through an essential transformation in view of generative artificial intelligence tools widely applied to generate texts mimicking human writings (Mondal, 2023).

Artificial intelligence encompasses an array of technologies such as machine learning, natural language processing, large language models and others (Ou et al., 2024) that are successfully applied in academic writing. Natural language models have been developing since the 1960s when the first computer programme called “Eliza” was offered to explore the human-computer communication. In late 2010s, large language models actively began their development, resulting in ChatGPT breakthroughs. AI-based appliances for enhancing academic writing appeared prior to ChatGPT and new-generation generative AI-powered tools¹, including Grammarly (2009), QuillBot (2017), DeepL (2017), Dimensions (2018) and others. But it was ChatGPT 3.5 that offered a real advance in functions and high-quality generated text (Raitetskaya & Lambovska, 2024).

With the advent of the new generation of generative artificial intelligence (GenAI) in 2022, especially with the breakthrough technologies of ChatGPT 3.5 and 4.0, the academia received smart tools that can perform numerous functions related to academic writing (Williams, 2024; Kohnke, 2024; Gunawan et al., 2024): text generation, grammar- and spelling-checking, citation- and reference-management, translation, editing, proofreading, feedback on writing, extraction of data, paraphrasing, reviewing articles, collaborative writing, annotation, and text coherence. Though some of them had been in use before 2022, the potential in all spheres has risen ever since. The technological shift signifies an AI-dominated age².

The impact of the GenAI-powered tools on academic writing is on the rise. Writings of various genres constitute an essential part of many professional activities, including those of writers, researchers, journalists, doctors, and teachers (Tikhonova & Raitetskaya, 2023). From today’s perspective, we might only guess an ultimate picture of GenAI spread. So far, it is obvious that these technologies would primarily be grasped within many professions and occupations, with education, science, and journalism as the frontrunners (Raitetskaya & Lambovska, 2024). The emerging field of research is changing very fast. Though, following the advances needs regular reviewing to fix the new shifts and adjustments, the proficiency of GenAI tools in academic writing requires the academia “to critically reconsider concepts such as cocreation, ownership, and authorship” (Borkurt, 2023) as the new tools “disrupt both the ontology and epistemology of academia, science and teaching” (Borkurt, 2024).

Generative AI-powered tools implemented into academic writing give rise to numerous challenges and concerns (Yao et al., 2024; Kim, 2024). First, the issues of authorship and integrity have become a highly disputed interdisciplinary field. The arguments of opponents and proponents of granting authorship to AI inspire a general revision and transformation of the authorship concept on the ground that it embraces ownership, accountability, and the integrity of ideas (Amirjalili et al., 2024). One more aspect connected with authorship is author’s voice that can be blurred by overreliance on GenAI language tools (Amirjalili et al., 2024). Second, to introduce progressive hybrid patterns of AI-human writing, researchers will have to study the subject in progress and find the optimal algorithms that require researching, teaching, academic writing practice, forging AI literacy and AI competence. Third, the progress in GenAI technologies is accelerating (Yao et al., 2024). It exacerbates potential negative effects that may be set off later with more research and empirical data on hand.

Starting with early 2023, the Scopus data base has been indexing publications on appliances of ChatGPT and GenAI-powered tools in academic writing. The emerging field requires regular revisiting for researchers to realize how new contributions may transform the research area and what new directions of research are forming. The reviews published in 2023 and 2024 focus on either a wider perspective, including academic writing as part of education, medicine, etc. (Ahn, 2024; Khalifa & Albadawy, 2024; Shorley et al., 2024) or on a narrowed context, e.g. optimizing the systematic reviewing process (Fabiano et al., 2024) and ethical dilemmas in using AI for academic writing (Miao et al., 2024). Having found a gap – a synthesis of knowledge related to appliances of GenAI-powered tools in academic writing, we aspire to make a review to add to the understanding of the field.

This scoping review aimed to explore the prevalent topics of the emerging research field of research on GenAI appliances in academic writing. To attain the objective, we were to reply to the following review questions:

- RQ#1: What are the prevailing directions of research of the potential appliances of generative artificial intelligence (GenAI) tools in academic writing at university and in university science?
- RQ#2: What are the major challenges and concerns related to GenAI appliances in academic writing?
- RQ#3: What are the key approaches in research towards authorship and academic integrity in the context of academic writing?

¹ AI-powered language tools are software programmes/ applications that use AI methods to analyse or generate human language, including but not limited to writing assistants, machine translators, speech-to-text transcribers, and text generators (chatbots) (Ou et al., 2024)

² Gates, B. (2023). The Age of AI has begun. Gates Notes. <https://www.gatesnotes.com/The-Age-of-AI-Has-Begun>

METHOD

Protocol

Prior to starting the present scoping review, a research protocol was meticulously developed. The authors hereby certify that this review report constitutes a faithful, precise, and transparent description of the conducted review. No deviations from the protocol were registered. Any departures from the original study design were appropriately described. This scoping review is based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) extension for Scoping Reviews (Tricco et al., 2018), and the framework proposed by Arksey and O'Malley (2005).

Search Eligibility Criteria

In the present review, the problem, concept, and context (PCC) framework was applied to state the eligibility criteria and structure the review (Table 1).

Table 1
Eligibility Criteria

Criterion	Inclusion	Exclusion	Rationale
Problem	GenAI appliances in academic writing	All publications going beyond	The review focuses on appliances of GenAI-based tools in academic writing. The problem is defined by the scope of such appliances
Concept	Academic writing	Other concepts	The aim of the review is to determine the trends of research on enhancing academic writing via GenAI
Context	Higher education and science	Other contexts	The review dwells upon the appliances of GenAI in academic writing in higher education and university science
Language	English	Other languages	The object of all research in focus is scholarly publications in English. The language choice is identified by its status as a lingua franca of international science
Time span	2023-2024	Previous years	The introduction of ChatGPT in 2023 started a new era of generative artificial intelligence that was widely spreading all over education and science
Types of sources	In the Scopus database all types of indexed publications relating to the theme	Unavailable sources, unavailable full texts	This review aims to get a comprehensive understanding of the field
Geographical location	Any location	None	Getting international perspective
Database	Scopus	Other bases than Scopus	Scopus was selected as it is widely recognized as the preferred source for scoping and systematic reviews, it has a reliable citation tracking and an impressive coverage of literature
Areas of Research	Social Sciences Arts & Humanities Medicine	Other areas	As the review focuses on the higher education and science contexts, publications rarely go beyond social sciences and arts & humanities. Though medicine is also under scrutiny as GenAI is widely introduced in research and academic writing within the field

Search Strategy

The search to attain the aim and to reply to the review questions was conducted as of October 24, 2024. The Scopus as one of the world's biggest high-quality databases was thoroughly searched to identify relevant publications subject to the eligibility criteria. The review questions, objective, and existing literature were studied to select the most appropriate keywords to achieve a search. The search was conducted using the keywords, i.e. "academic writing" AND "AI-based tools", "academic writing" AND "AI-powered tools", and "scholarly writing" AND "AI-based writing". Other potential keywords were applied in pre-protocol searches but failed to bring any relevant results. The full-text publications eligible for the review were identified after screening of the titles, keywords, and abstracts. All relevant documents with full texts were included in the review.

Study Selection

First, both authors identified research publications sticking to the eligibility criteria. After applying the Scopus filters

(time span, subject area, language), each reviewer independently screened the titles, and then the abstracts and keywords of the identified documents. Second, each reviewer tagged the documents with “to include” or “to exclude” marks. In case of disagreement, the authors arrived at a mutual consent. No disputed issue required lateral expertise. The full texts were found via the publishers. Each full text was thoroughly read and independently analysed by each reviewer. Eligible publications were identified.

Data Extraction

The title and review questions were determined under the PCC framework. Pre-protocol searches made us identify the basic structure of the extracted data for the review:

- 1. data from the reviewed documents relating to the potential of GenAI-powered language tools for academic writing at university and in science;
- 2. data from the reviewed publications regarding challenges and concerns arising out of the appliances of such tools in academic writing;
- 3. data from the articles under review containing information on authorship and integrity issues relating to the

use of the GenAI-powered language tool in academic writing at large.

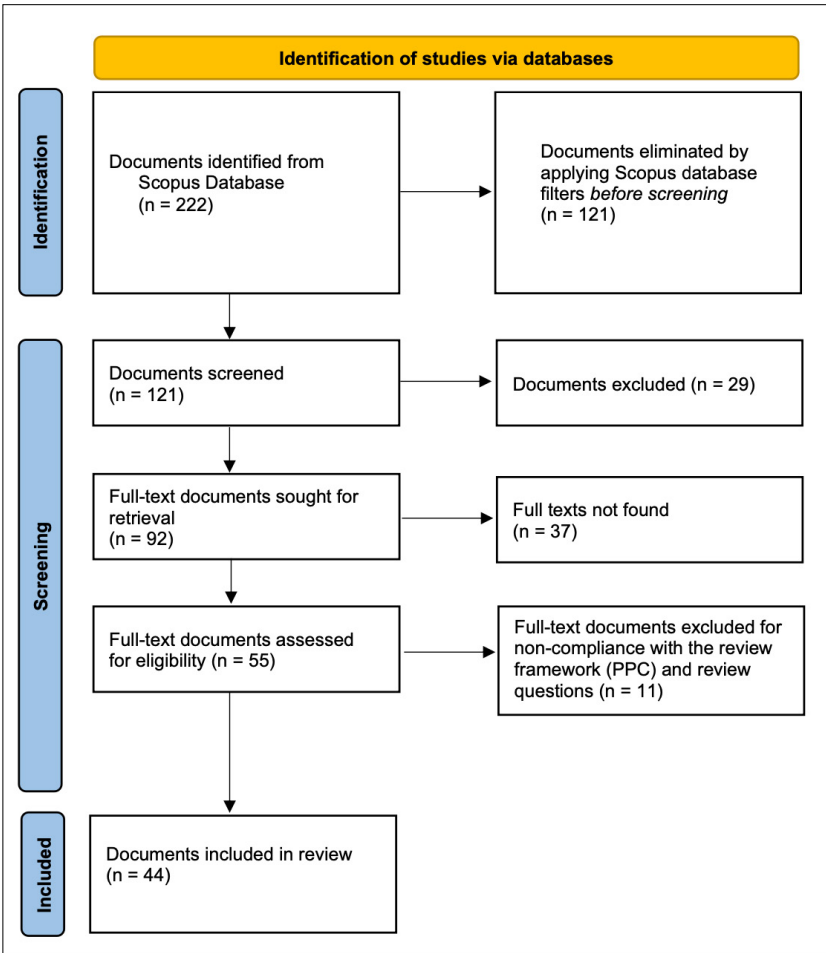
All raw data were double-checked by the authors. The extracted data were classified and formalized in the corresponding exhibits (Table 5 and Appendixes 2-3).

RESULTS

Search and Selection Results

The search results were finalised as of October 24, 2024. Initially, we found a total of 222 documents in the Scopus database. After we had applied the selected Scopus filters (language; social sciences, time span), the total decreased from 222 to 121 studies. Then we screened the titles and abstracts. 92 documents were deemed irrelevant and excluded from the review. As not all full texts were found, we analysed only 55 publications, with another 11 publications eliminated as non-eligible. The PRISMA flow-chart (Figure 1) depicts the whole identification and screening procedure.

Figure 1
Selection of Publications for the Review



Documents Ultimately Included in the Review

Bibliometric Characteristics of the Research Field

The review yielded 33 articles, two editorials, and nine reviews meeting the objective and eligibility criteria (Table 2). For the complete metadata on the included documents, see Appendix 1.

The 44 documents included into the present review were analysed on the following aspects: yearly distribution; types of documents; authors; countries of affiliation; journals; or-

Table 2
Documents included in the review

Reference		Publication Title
Articles & Editorials		
1	Williams, 2024	Comparison of generative AI performance on undergraduate and postgraduate written assessments in the biomedical sciences
2	Kohnke, 2024	Exploring EAP students' perceptions of GenAI and traditional grammar-checking tools for language learning
3	Li et al, 2024	Exploring the potential of artificial intelligence to enhance the writing of English academic papers by non-native English-speaking medical students - the educational application of ChatGPT
4	Liu et al., 2024	The great detectives: humans versus AI detectors in catching large language model-generated medical writing
5	Johnston et al., 2024	Student perspectives on the use of generative artificial intelligence technologies in higher education
6	Mahapatra, 2024	Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study
7	Gralha & Pimentel, 2024	Gotcha GPT: Ensuring the Integrity in Academic Writing
8	Rafida et al., 2024	EFL students' perception in Indonesia and Taiwan on using artificial intelligence to enhance writing skills
9	Bolaños et al., 2024	Artificial intelligence for literature reviews: opportunities and challenges
10	Kraika & Olszak, 2024a	"AI, will you help?" How learners use Artificial Intelligence when writing
11	Rababah et al., 2024	Graduate Students' ChatGPT Experience and Perspectives during Thesis Writing
12	Ou et al., 2024	Academic communication with AI-powered language tools in higher education: From a post-humanist perspective
13	Yao et al., 2024	A Qualitative Inquiry into Metacognitive Strategies of Postgraduate Students in Employing ChatGPT for English Academic Writing
14	Morreale et al., 2024	Artificial Intelligence and Medical Education, Academic Writing, and Journal Policies: A Focus on Large Language Models
15	Kurt & Kurt, 2024	Enhancing L2 writing skills: ChatGPT as an automated feedback tool
16	Krajka & Olszak, 2024b	Artificial intelligence tools in academic writing instruction: exploring the potential of on-demand AI assistance in the writing process
17	Parker et al., 2024	Negotiating Meaning with Machines: AI's Role in Doctoral Writing Pedagogy
18	Kim, 2024a	Research ethics and issues regarding the use of ChatGPT-like artificial intelligence platforms by authors and reviewers: a narrative review
19	Alkamel & Alwagieh, 2024	Utilizing an adaptable artificial intelligence writing tool (ChatGPT) to enhance academic writing skills among Yemeni university EFL students
20	Kim et al., 2024	Exploring students' perspectives on Generative AI-assisted academic writing
21	Tarchi et al., 2024	The Use of ChatGPT in Source-Based Writing Tasks
22	Maphoto et al., 2024	Students' Academic Excellence in Distance Education: Exploring the Potential of Generative AI Integration to Improve Academic Writing Skills
23	Alea Albada & Woods, 2024	Giving Credit Where Credit is Due: An Artificial Intelligence Contribution Statement for Research Methods Writing Assignments

Reference		Publication Title
24	Mohammad et al., 2024	Paraphrasing Prowess: Unveiling the Insights of EFL Students and Teachers on QuillBot Mastery
25	Amirjalili et al., 2024	Exploring the boundaries of authorship: a comparative analysis of AI-generated text and human academic writing in English literature
26	Bozkurt, 2024	GenAI et al.: Cocreation, Authorship, Ownership, Academic Ethics and Integrity in a Time of Generative AI
27	Nguyen et al., 2024	Human-AI collaboration patterns in AI-assisted academic writing
28	Kanddel & Eldakak	Legal dangers of using ChatGPT as a co-author according to academic research regulations
29	Malik et al., 2023	Exploring Artificial Intelligence in Academic Essay: Higher Education Student’s Perspective
30	Utami et al., 2023	Utilization of artificial intelligence technology in an academic writing class: How do Indonesian students perceive?
31	Jarrah et al., 2023	Using ChatGPT in academic writing is (not) a form of plagiarism: What does the literature say?
32	Alberth, 2023	The use of ChatGPT in writing: a blessing or a curse in disguise?
33	Khaif et al., 2023	The Potential and Concerns of Using AI in Scientific Research: ChatGPT Performance Evaluation
34	Teng, 2023	Scientific Writing, Reviewing, and Editing for Open-access TESOL Journals: The Role of ChatGPT
35	Mahyoob et al., 2023	A Proposed Framework for Human-like Language Processing of ChatGPT in Academic Writing
Reviews		
36	Ahn, 2024	The transformative impact of large language models on medical writing and publishing: current applications, challenges and future directions
37	Fabiano et al., 2024	How to optimize the systematic review process using AI tools
38	Khalifa & Albadawy, 2024	Using artificial intelligence in academic writing and research: An essential productivity tool
39	Miao et al., 2024	Ethical Dilemmas in Using AI for Academic Writing and an Example Framework for Peer Review in Nephrology Academia: A Narrative Review
40	Shorey et al., 2024	A scoping review of ChatGPT’s role in healthcare education and research
41	Gunawan et al., 2024	ChatGPT integration within nursing education and its implications for nursing students: A systematic review and text network analysis
42	Imran & Almusharraf, 2023	Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature
43	Tikhonova & Raitskaya, 2023	ChatGPT: Where Is a Silver Lining? Exploring the realm of GPT and large language models
44	Mondal & Mondal, 2023	ChatGPT in academic writing: Maximizing its benefits and minimizing the risks

ganisations (affiliations); research areas (though we limited this aspect to Social Sciences, Arts & Humanities and Medicine, documents tend to be classified in more than one area).

As ChatGPT 3.5 was introduced in late 2022, and the first research publications came into being as early as January 2023, we set the timespan covering 2023 and 2024. 2023 brought 10 documents, whereas 2024 accounted for 34 publications (though the annual statistics for 2024 are not complete yet). By type, the 44 documents broke down as follows: 33 articles, 2 editorials, and 9 reviews. The following journals brought out two publications each: *Contemporary Educational Technology*, *International Journal for Educational Integrity*, *International Journal of Artificial Intelligence in Education*, *Nurse Edu-*

cation Today, and *Open Praxis*. The other 34 documents were published in 34 journals, with one per journal.

The total number of authors was 156. The most prolific researchers were J. Krajka and I. Olszak who co-authored two out of 44 articles. The remaining 154 researchers participated in one article each either as an author or co-author. Every publication in the review had an average of 3.5 authors. Only six documents were written by a single author. The authors had 102 affiliations, including three authors affiliated with the Hong Kong Polytechnic University, two authors from each of the three universities – Uniwersytet Marii Curie-Skłodowskiej w Lublinie, University of Liverpool, and Uniwersytet Marii Curie-Skłodowskiej w Lublinie. The remaining 98 affil-

iations were represented by one author each, though some authors had more than one affiliation.

Geographically, the breakdown of the publications under review embraced the USA with three documents, China, Hong Kong, and Indonesia, with four documents each, and Saudi Arabia with two publications. The remaining 10 countries had one publication each. Though 39 (48.1 %) out of 44 publications belonged to Social Sciences, they and the remaining five documents were simultaneously attributed to other areas: Computer Science (11 documents or 13.6 %), Arts & Humanities (7 documents or 8.6 %), Medicine (6 documents or 7.4 %), Engineering (4 documents or 4.9 %), Business, Management and Accounting (3 documents or 3.7 %), Psychology (3 documents or 3.7 %), Nursing (2 documents or 2.5 %), and Other areas (6 documents or 7.4 %).

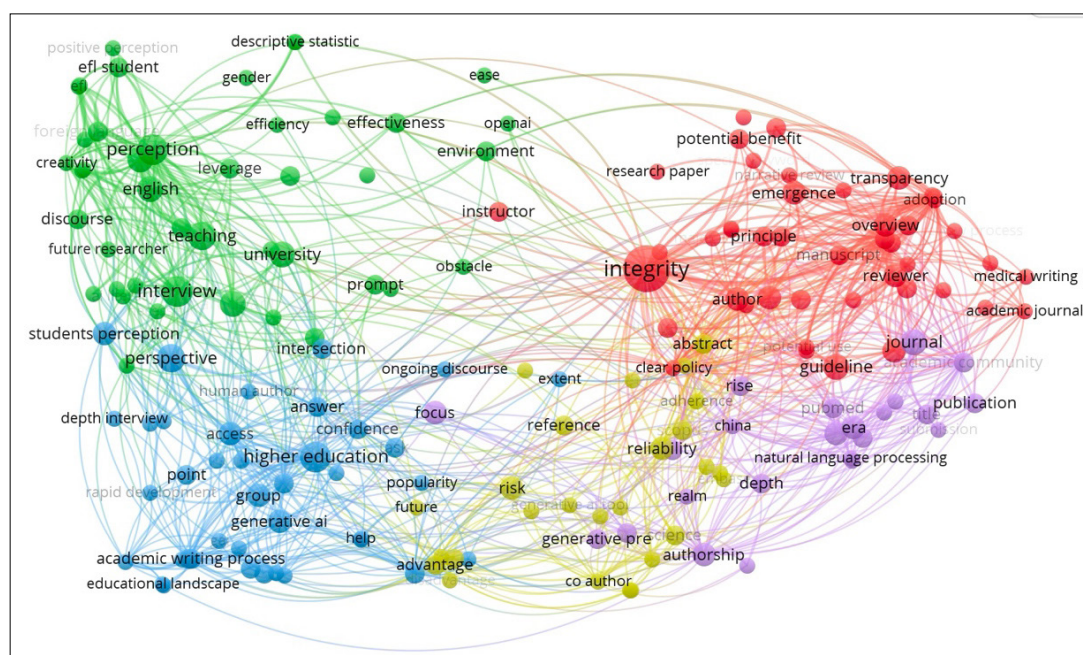
Hypothetical Thematic Clusters

While pilot-searching the Internet, we outlined probable thematic clusters that were tested and explored during the search, identification, screening and eliminations of the documents subject to the eligibility criteria. After we had iteratively revised the clusters, the following thematic clusters were finalised for the present review: (1) AI in enhancing academic writing; (2) AI challenges in academic writing, and (2) Authorship and integrity related to GenAI appliances (Table 3). Clusters 2 and 3 cover all aspects of challenges and concerns arising out of applications of the language tools in academic writing. The clusters fully conformed to the review questions and covered the main findings of the 44 selected documents.

Table 3
Hypothetical Thematic Clusters

Thematic cluster		Cluster Description
1.	Cluster One	
	AI in enhancing academic writing	In the context of generative AI appliances: general issues of enhancing academic writing; strategies followed in academic writing; writing articles and literature reviews; GenAI-based feedback on academic writing at university courses
2	Cluster Two	
	AI challenges in academic writing	Challenges and pitfalls brought about by GenAI appliances in academic writing
3	Cluster Three	
	Authorship and integrity related to GenAI appliances	Issues of authorship and plagiarism of texts produced by GenAI

Figure 2
VOSviewer Visualization of the Review Thematic Clusters



Some of the 44 publications focus on AI aspects other than academic writing, being complex in their scope (Johnston et al., 2024; Ou et al., 2024; Morreale et al., 2024; Maphoto et al., 2024; Shorley et al., 2024; Tikhonova & Raitskaya, 2023). Though they contain some findings that go beyond the present review, the extracted data were in compliance with the research objective and review questions.

The VOSviewer software’s analysis of the metadata from the 44 selected publications mapped out a structured landscape of thematic clusters, each colour-coded to denote a specific domain of the review (Figure 2). The density of terms start-

ed from 4. The software forked out five clusters. The clusters partially overlapped. The purple cluster covered integrity and authorship issues of appliances of AI-powered language tools. It also included journal practices related to AI-powered text. The red cluster focused on integrity, research papers, medical writing, and academic journals. The green cluster mainly represented perceptions, teaching, university issues, obstacles, open IA, creativity. The blue cluster comprised higher education, generative AI, academic writing process, advantages, and perspectives. The yellow cluster is densely interrelated with the purple, blue and red clusters, focusing on generative AI, risks, reference and reliability aspects.

Table 4
Mapping the publications to the clusters

SN	Authors and Year	Cluster 1	Cluster 2	Cluster 3
		AI in enhancing academic writing	AI Challenges in academic writing	Authorship and Integrity
1	Williams, 2024	✓	✓	
2	Kohnke, 2024		✓	
3	Li et al, 2024	✓		
4	Liu et al., 2024		✓	✓
5	Johnston et al., 2024	✓		
6	Mahapatra, 2024	✓		
7	Gralha & Pimentel, 2024			✓
8	Rafida et al., 2024	✓		
9	Bolaños et al., 2024	✓	✓	
10	Kraika & Olszak, 2024a	✓		
11	Rababah et al., 2024	✓	✓	
12	Ou et al., 2024	✓	✓	
13	Yao et al., 2024	✓	✓	
14	Morreale et al., 2024		✓	✓
15	Kurt & Kurt, 2024	✓	✓	
16	Krajka & Olszak, 2024b	✓	✓	
17	Parker et al., 2024	✓		
18	Kim, 2024			✓
19	Alkamel & Alwagieh, 2024	✓		
20	Kim et al., 2024	✓		✓
21	Tarchi et al., 2024	✓		
22	Maphoto et al., 2024	✓		
23	Alea Albada & Woods, 2024	✓		✓
24	Mohammad et al., 2024	✓		
25	Amirjalili et al., 2024		✓	✓
26	Bozkurt, 2024			✓
27	Nguyen et al., 2024	✓		

SN	Authors and Year	Cluster 1 AI in enhancing academic writing	Cluster 2 AI Challenges in academic writing	Cluster 3 Authorship and Integrity
28	Kanddel & Eldakak			✓
29	Malik et al., 2023	✓		
30	Utami et al., 2023	✓		
31	Jarrah et al., 2023			✓
32	Alberth, 2023	✓	✓	
33	Khaif et al., 2023	✓		
34	Teng, 2023	✓		
35	Mahyoob et al., 2023	✓		
36	Ahn, 2024	✓		
37	Fabiano et al., 2024	✓		
38	Khalifa & Albadawy, 2024	✓	✓	
39	Miao et al., 2024			✓
40	Shorey et al., 2024	✓	✓	
41	Gunawan et al., 2024	✓		
42	Imran & Almusharraf, 2023	✓		
43	Tikhonova & Raitskaya, 2023	✓	✓	✓
44	Mondal & Mondal, 2023	✓	✓	
TOTAL		35	16	12

Given the difference of the initial inputs, the hypothetical clusters differ from the software clusters. The VOSviewer was limited to the meta-data of the publications (titles, abstract, authors' keywords), whereas the reviewers analysed full-text publications. Overlapping of clusters also underlay the variances in clusters by the reviewers and the software. The reviewers enlarged the clusters as compared to the VOSviewer analysis.

Potential of GenAI-Powered Language Tools for Academic Writing at University and in Science

The raw data on GenAI-powered language tools in academic writing extracted from the reviewed publications are stated in Table 5. In describing this direction of study, we boiled down the most prominent features and characteristics articulated in the reviewed publications to the following:

General Issues of GenAI-Powered Language Tools

Gunawan et al. (2024) mark that academic writing is the pre-dominant cluster when it comes to GenAI applications in higher education. In many contexts, researcher focus on AI-powered language tools as they are easily accessible and

user-friendly for students and researchers (Ou et al., 2024; Kurt & Kurt, 2024; Krajka & Olszak, 2024a). A wide integration of such tools in academic writing presents a paradigm shift (Nguyen et al., 2024). The whole writing process is being transformed and reinforced (Ou et al., 2024). Researchers pointed out that AI provides "dynamic, responsive learning environments and bespoke educational experiences" (Malik et al., 2023). Unfortunately, no research on AI-powered translation was found and included in the review. This sub-field is evolving rather successfully, but this review considered only translation as function of GenAI-powered language tools (Amirjalili et al., 2024; Alberth, 2023; Imran & Almusharraf, 2023; Li et al., 2023).

Strategies in Academic Writing and Hybrid Writing

To be better equipped with "competencies sufficient to navigate this new terrain" (Parker et al., 2024), students or users should follow prudent strategies that were suggested in several research papers (Kraika & Olszak, 2024a; Yao et al., 2024; Mohammad et al., 2024; Nguyen et al., 2024; Mahyoob et al., 2024). Those articles add to the bulk of the reviewed publications on enhancing academic writing with GenAI tools. Some of the documents in this cluster dwelt upon hybrid or entirely AI-powered writing of specific assignments, articles and reviews (Li et al., 2024; Bolaños et al., 2024; Williams, 2024; Alea

Albada & Woods, 2024; Tarchi et al., 2024). Hybrid forms of writing offer appliances of the language AI-powered tools not in text generation but in “writing structure, relevant sources, and new insights about the topic” (Alberth, 2023). Some problems in academic writing may be overcome via AI-powered tools, including typos, spelling errors, and grammar mistakes (Kim et al., 2024b), proper referencing practices (Jarrah et al., 2023), data summarization (Shorley et al., 2024), morphological analyzers, speech recognizers, text classifiers (Mahyoob et al., 2023) and others. Studies of specific GenAI-powered language tools are rare (Krajka & Olszak, 2024b; Mahyoob et al., 2023; Mohammad et al., 2024) as the technology terrain is evolving with information getting outdated fast.

Providing Feedback to the Users

Mahapatra (2024) notices that AI-driven tools that had been introduced before ChatGPT 3.5 were successful at “providing immediate feedback” to students, language learners and authors. The potential of the new technologies is higher because GenAI-powered tools are trained on big data corpora and are capable “to identify complex language patterns” (Kurt & Kurt, 2024). The learners may manipulate by wording prompts and look for a feedback they personally need (Kurt & Kurt, 2024).

Literature Review Generated with GenAI-Powered Language Tools

In this review, we are limited to the appliances of GenAI-powered language tools in academic writing, we still suppose that there is an essential aspect for both academic writing and research. It is a literature overview that constitutes an integral part of any research paper. According to the reviewed documents, we are “moving towards semi-automatic creation of literature reviews” (Bolaños et al., 2024). Kim et al. (2024) admit using AI in literature review at several stages (identifying relevant publications, supplying background information, summarizing texts and others).

GenAI-Powered Tools as Assistants

Another important issue in enhancing academic writing is a potential of GenAI-powered tools that may assist in writing an article in compliance with the best standards of the academia. Today, such a task is not possible (Nguyen et al., 2024). Liu et al. (2024) doubt that a credible academic article can be created by GenAI as there are no well-established discipline-specific large language models. They also note that AI-generated articles offer “superficial discussion” without evidence and suffer from redundancy (Liu et al., 2024).

Major Challenges and Concerns Related to GenAI Appliances in Academic Writing

Khalifa and Albadaawy (2024) indicated the major ethical challenges arising out of the AI-human interrelation in research. They are “the importance of human intelligence in research and the limitations ...[of] AI tools ... in guiding research ideas and design” (Khalifa & Albadaawy, 2024). The challenges associated with the appliances of GenAI language tools in academic writing (Appendix 2) range from basic and profound in nature (authorship and integrity, overreliance on AI, equity issues, lack of transparency, absence of long-term memory in dialogue in GenAI-powered language tools, problematic identification of AI-generated text) to more specific (propensity to generate inaccurate references, false or biased content, lack of author’s voice in AI-generated texts, inability to create credible academic texts).

Key Approaches Towards Authorship and Academic Integrity in the Context of Academic Writing

The issue that is the core of many challenges and concerns is the characteristics of the AI-generated text. Researchers make attempts to define or evaluate its originality and authenticity (Yao et al., 2024). The multiple stances regarding authenticity and related issues result from the lack of transparency in content generation (Shorey et al., 2024). Rephrasing of various concepts may be perceived as new ideas as AI-generated text tends to be deficient in references. AI-generated texts, including articles, abound in “information repetition, nonfactual inferences, illogical reasoning, fake references, hallucination, and lack of pragmatic interpretation” (Mahyoob et al., 2023).

Some researchers link authorship with the author’s voice. In academic writing the concept “voice” is constructed by “genre and community constraints in academic writing” (Amirjalili et al., 2024). Amirjalili et al. (2024) see “the notion of voice as an extension of the human author”. The author’s voice is determined by the use of lexical selections, syntactic structures, hedges, boosters, and personal pronouns as measurable indicators of an author’s presence and position. Amirjalili et al. (2024) offered the voice intensity rating scale that included assertiveness, self-identification, “reiteration of the central point”, authorial presence and autonomy of thought. In their research, a deep analysis of human-like writings by ChatGPT showed that there were problems with “register, clichéd language, and a lack of nuance” (Amirjalili et al., 2024).

Table 5*Enhancing academic writing with Generative Artificial Intelligence: Raw Data from the documents under review*

GenAI Potential	Reference	Raw Data
Predominant cluster in literature reviews of GenAI applications in education	Gunawan et al., 2024	"Academic Writing" is the predominant cluster, constituting 39 % of total nodes, and signifies a novel contribution to the literature. The study emphasizes the strong influence of ChatGPT in enhancing academic writing skills among nursing students
AI-powered language tools (AILT)	Ou et al., 2024	AILTs (i.e., ChatGPT, Grammarly and Google Translate) in academic writing, showing their utility as part of a spatial repertoire in enhancing students' academic communication performance and facilitating personal language development The proliferation of AILTs has transformed students' everyday academic writing process into an additional learning space and bestowed upon students a novel identity of spatially advised learners, empowering them to acknowledge AI's facilitating role for personal competence enhancement while remaining aware of its inherent limitations
	Yao et al., 2024	... the students effectively employed ChatGPT to generate ideas, create outlines, revise the content and proofread their manuscripts... The students recognised several strengths of ChatGPT in the context of academic writing, including its efficient responsiveness to human instructions and proficiency in language revision
	Kurt & Kurt, 2024	ChatGPT can generate grammatically correct essays, suggest essay topics, create outlines (Barrot, 2023), help generate ideas (Lingard, 2023),T adjust text difficulty to learners' proficiency levels (Bonner et al., 2023), and facilitate guided writing (Kohnke et al., 2023)
	Krajka & Olszak, 2024	A range of intelligent CALL tools, supported by artificial intelligence, can be used to assist foreign language writing teaching and learning. Pokrivcakova (2019) [offers] a comprehensive overview of such applications, enumerating a) personalised learning materials, b) machine translation tools, c) AI writing assistants, d) chatbots, e) AI-powered language learning software (platforms and apps), f) intelligent tutoring systems (ITS), g) intelligent virtual reality (IVR) applications
	Nguyen et al., 2024	The integration of state-of-the-art AI-assisted writing assistants into the academic writing process represents a paradigm shift. These tools not only provide assistance in drafting and revising text but also in conducting literature reviews and synthesising information, which are critical components of scholarly writing
	Alkamel & Alwagieh, 2024	The development of artificial intelligence (AI) has facilitated the creation of highly advanced language and writing tools that possess enhanced capabilities and effectiveness. (Geitgey, 2018; Brown et al., 2020)
	Mondal & Mondal, 2023	Its [ChatGPT] ability to generate human-like text, answer questions, and summarize information has made it a valuable resource for researchers and academics across a wide range of disciplines. ChatGPT can assist in tasks such as literature review, data analysis, and even writing entire sections of academic papers
	Parker et al., 2024	...the goal is to equip doctoral students with competencies sufficient to navigate this new terrain confidently and responsibly... ...a prudent strategy that (a) recognizes the potential synergy of human-AI interactions, (b) values the potential innovative partnerships, and (c) maintains ethical academic standards... Given the hybrid human-AI writing process that evolved through students' collaboration with AI, there is an urgent need for institutions to develop clear and prominently displayed policies regarding ethical AI use and academic integrity...

GenAI Potential	Reference	Raw Data
Strategies for applying GenAI	Kim et al., 2024	... this study found a strong need to develop students' capacity for prompt engineering, the process of crafting, optimizing, and employing text that can be interpreted and understood by GenAI. This would enable improved communication with GenAI to harness its capability to perform tasks (e.g., generating educational content) as intended, and ensure accurate, relevant, and quality outcomes
	Tarchi et al., 2024	Eager and Brunton (2023) suggested how the efficacy of AI in education may depend on the ability to write effective prompts for use with conversation-style AI models
	Nguyen et al., 2024	Our study shows that the higher-performing doctoral students' engagement in AI-assisted writing is multifaceted, suggesting a higher familiarity with the tool. The observed sequence of actions, starting with prompting the GAI-powered tool for content and subsequently searching articles, reflects a proactive approach to information gathering. This tactic, contrasting with merely waiting for generated responses, optimises productivity and stimulates cognitive processes. The subsequent sequence of reading, copying, pasting, and editing or integrating content indicates a methodical approach whereby the students critically assess, adapt, and incorporate the AI-generated material into their writing
	Alberth, 2023	The idea is not to rely solely on the application to write an entire research paper, but rather to use it as a tool for gathering necessary information such as the writing structure, relevant sources, and new insights about the topic. Authors may also ask ChatGPT to provide feedback on their draft papers
Enhancing academic writing	Kohnke, 2024	Research indicates that tools such as Grammarly, which provides AWCF, enhance the accuracy of student writing, metalinguistic awareness and self-directed learning (Barrot, 2023b)
	Kim et al., 2024	Expanding beyond an automated evaluation and correction, AI writing systems facilitate students' metacognition by allowing them to identify and correct language errors (Fitria, 2021), notice dissonance in their writing (Gayed et al., 2022), and improve their manuscript's overall clarity and coherence (Liu et al., 2023)...
	Mahapatra, 2024	Yan (2023) has reported benefits to students' writing skills through its use, he has also warned that its use can threaten academic honesty and ethicality in writing...
	Rafida et al., 2024	AI also improves academic writing among EFL students, including task completion, citation accuracy, and sentence construction (Pitychoutis, 2024; Setyowati et al., 2023)
	Parker et al., 2024	This evolution in writing practices signifies a shift towards a more integrated, collaborative approach to academic writing, where AI tools are not mere aids but partners in the creative process
	Kim, 2024	In writing articles, AI can be utilized for accurate translation, grammatical correctness, and idea generation, as well as for summarizing content, and crafting conclusions (Kim, 2024a)
Enhancing academic writing	Malik et al., 2023	AI, with its groundbreaking technologies and adaptive learning mechanisms, enriches academic writing by providing dynamic, responsive learning environments, and bespoke educational experiences. It delves into the intricacies of language acquisition and offers tailored solutions, making the processes inherent in academic writing more streamlined and intuitive The data also reveals that many students appreciate AI's role in suggesting appropriate essay ideas (67 %), extracting meaningful data from large datasets (69 %), and analyzing data for data-driven writings (70 %). Furthermore, a considerable percentage of respondents acknowledge AI's contribution to ensuring uniqueness and avoiding accidental plagiarism (73 %), improving language by providing sentence recommendations (75 %), and enhancing article quality by spotting flaws (83 %)...
	Alberth, 2023	McFarlane (2023), the current version of the chatbot can assist with academic writing in two ways. The first way is that when the author conducts a literature review and takes brief notes or bullet points for each reference, they can request ChatGPT to arrange and convert these notes into a well-structured text. The second way is that ChatGPT can be useful for sorting and managing references and citations

GenAI Potential	Reference	Raw Data
Enhancing academic writing	Tikhonova & Raitskaya, 2023	...the technologies are advantageous for non-native English-speaking authors or even native speakers as they may avoid weaknesses in their submissions related to the language quality...
Assistive tools that improve writing skills	Li et al., 2024	The results from the participants' two-week unrestricted usage of the AI model ChatGPT to enhance their assignments indicated a noticeable improvement in the quality of student papers. This suggests that large language models could serve as assistive tools in medical education by potentially improving the English writing skills of medical students
	Alkamel & Alwagieh, 2024	While ChatGPT can provide valuable support in academic writing, it is important for students to view it as a tool to enhance their skills rather than a replacement for their own efforts. Students should use ChatGPT to gain insights, learn from its suggestions, and improve their writing, but they also should strive to develop their own critical thinking and writing abilities
	Tarchi et al., 2024	The potential of ChatGPT Firstly, it enhances efficiency by significantly reducing the time and effort required for content creation, benefiting both students and educators (Lund et al., 2023; Yan, 2023). Secondly, it provides ideation support by suggesting new ideas and perspectives for writing assignments (Kasnezi et al., 2023; Taecharungroj, 2023). Additionally, it offers invaluable language translation assistance, helping non-native language students ensure accuracy and grammatical correctness in their writing (Lametti, 2022; Lund & Wang, 2023; Stacey, 2022; Stock, 2023)
	Imran & Almusharraf, 2023	These points would help in understanding its [ChatGPT] use in writing as an assistant and AI tool. 1. Increased efficiency: ChatGPT's invention can reduce the time and effort required to generate written content... 2. Idea generation: ChatGPT can help students generate new ideas for their writing assignments by suggesting topics, themes, and perspectives that they might not have considered otherwise (Kasnezi et al., 2023; Taecharungroj, 2023). 3. Language translation: ChatGPT can translate text from one language to another, which can be useful for students who are writing papers in a language that is not their native tongue. This can help students ensure that their writing is accurate and grammatically correct (Lametti, 2022; Lund & Wang, 2023; Stock, 2023). 4. More accurate and consistent content: With the ChatGPT invention, there is a higher likelihood of producing accurate and consistent content... 5. Improved collaboration: ChatGPT can also facilitate collaboration among students and educators
	Mahapatra, 2024	ChatGPT ...offers advice regarding various structural aspects of a text and translate it (Imran & Almusharraf, 2023), and facilitate guided writing (Kohnke et al., 2023)...
Quality of the writing	Rababah et al., 2024	[The] findings suggest that postgraduate students at Jadara University hold favorable views regarding ChatGPT's utility, ease of use, impact on thesis completion speed, and the quality of work it produces ...using this tool reduces the time spent on literature review and referencing, improves readability, enhances the quality of the thesis, and provides valuable research ideas The findings suggest that students view ChatGPT as a beneficial tool that enhances the writing process, writing quality, knowledge retrieval, and the generation of new ideas.
	Nguyen et al., 2024	While AI, particularly in its generative form, does not possess the ability to fully synthesise literature or independently engage in critical writing, it has shown considerable proficiency in aiding these processes. Specifically, generative AI can assist by aggregating and summarising relevant literature and generating written content based on specific prompts

GenAI Potential	Reference	Raw Data
Text generation – Proofreading and Editing – Summarizing the texts – Translation and inter- pretation Paraphrasing	Shorey et al., 2024	In healthcare research and academic writing, ChatGPT's value is evident in aiding manuscript drafting/composition, data summarization and citation management (Lund et al., 2023; Sallam, 2023)
	Williams, 2024	...the AI tools were able to generate essays that generally met the scientific accuracy criteria for both undergraduate and postgraduate levels
	Gunawan et al., 2024	ChatGPT has...potential as a tool for generating written content, reviewing articles, and collaborative writing exercises (Sun and Hoelscher, 2023)
	Kim, 2024	ChatGPT could also proofread and edit sentences, identify grammatical errors, paraphrase, improve writing quality, and summarize the texts (Castonguay et al., 2023; Sun and Hoelscher, 2023)
	Kim et al., 2024	... ChatGPT has been noted for producing more refined sentences more quickly than traditional English proofreading services, making it a valuable tool for language editing (Kim, 2023)...
	Kim et al., 2024	An evaluation of Wordvice AI, a proofreading tool, highlighted that the tool could outperform the built-in proofreading abilities of Google Docs or Microsoft Word, but still only managed to identify 77% of what was identified by a human proofreader (Heintz et al., 2022)
	Mohammad et al., 2024	...AI may be able to support typos, spelling errors, and grammar mistakes...
	Jarrah et al., 2023	Numerous studies suggest that online paraphrasing tools such as paraphrase-tool.com, QuillBot.com, prepotseo.com, and spinbot.com can be beneficial in addressing students' challenges in academic writing
	Kim et al., 2024	AI tools can streamline the process of citation and referencing by automatically generating accurate citations based on given referencing styles. This reduces the likelihood of citation errors and helps students maintain consistency and adhere to proper referencing practices
	Kim et al., 2024	AI writing systems also offer real-time translation and interpretation services. This enables students to overcome language barriers to access and assimilate content in multiple languages and learn diverse perspectives (Salvagno et al., 2023)
Feedback on writing	Kohnke, 2024	By providing immediate and clear feedback, GenAI tools reduce extraneous cognitive load, allowing students to focus more on content and higher-order writing skills. This can lead to more efficient learning and better retention of writing strategies (Paas et al., 2003)
	Kohnke, 2024	Compared to traditional grammar-checking tools, GenAI tools go beyond simple error correction to provide detailed explanations of linguistic rules, potentially enhancing students' overall language proficiency (Dizon & Gayed, 2021; Tan, 2023)
	Kohnke, 2024	These tools support self-regulated learning (SRL) by providing immediate feedback that helps students monitor and control their learning (Chiu, 2024; Zimmerman, 2000)
Feedback on writing	Mahapatra, 2024	With the proliferation of AI-driven tools such as Grammarly, QuillBot, Copy.ai, Word-Tune, ChatGPT, and others, it has become easier for students to obtain feedback on their writing (Marzuki et al., 2023; Zhao, 2022)...
	Rafida et al., 2024	These tools offer real-time feedback on grammar, structure, and style, systematically improving skills (Ahmad et al., 2023; Khotimah et al., 2024)
	Kurt & Kurt, 2024	Trained on extensive text corpora, these LLMs can identify complex language patterns and offer more detailed and contextually relevant feedback. In contrast to traditional AWE systems, LLMs use a natural language interface that simplifies and enhances the feedback process (Kasneci et al., 2023)...
	Kurt & Kurt, 2024	Its interactive and adaptable nature allows users to gain expertise in manipulating and accessing the kind of feedback they are seeking... Given the critical role of feedback in L2 writing and ChatGPT's potential to offer high-quality feedback on mechanics, styling, content, and organization, its integration as an automated evaluation tool is seen as promising for L2 learners' writing development (Guo & Wang, 2024)

GenAI Potential	Reference	Raw Data
The use of AI in Literature Reviews (SLRs) and other scholarly publications	Bolaños et al., 2024	The increasing role of AI in this field shows great potential in providing more effective support for researchers, moving towards the semi-automatic creation of literature reviews... AI [is applied] in the screening and extraction phases The other four tools (Covidence, PICOPortal, and EPPI-Reviewer, Colandr) undertake two AI-related tasks. They all classify papers as relevant/irrelevant, but also execute an additional task, such as identifying a specific type of paper (e.g., economic evaluation, randomised controlled trials, etc.) or categorising papers according to a set of entities defined by the user ExaCT, Dextr, and Iris.ai perform Named Entity Recognition (NER) Nasar et al. (2021) to extract various types of information from the relevant articles. Only two tools offer support for post-screening: Iris.ai and Nested Knowledge. Specifically, Iris.ai generates summaries from either a single document, multiple abstracts, or multiple documents... the summary is formed by generating new sentences that encapsulate the core information of the original text
	Kim, 2024	...researchers can use AI to discover, translate, and summarize articles and research trends, identify experimental methods and scientific knowledge, and compile results and statistics... In a hybrid narrative review case involving collaboration between humans and ChatGPT, the results highlighted both the effectiveness and the concerns associated with ChatGPT (Temsah et al., 2023)
	Kim et al., 2024	AI writing systems assist students in literature review by identifying relevant research articles (Behrooz et al., 2023), supplying background information on writing topics (Chichekian & Benteux, 2022; Rowland, 2023), summarizing texts (Behrooz et al., 2023), and providing recommendations tailored to students' preferences and search patterns (Chichekian & Benteux, 2022; Rowland, 2023).
	Khalifa & Albadawy, 2024	The synthesis of literature through AI, producing summary tables and comparative analyses, represents a revolutionary stride in automated literature synthesis, offering a comprehensive and nuanced perspective of existing research... AI's capability to identify gaps in literature is invaluable. Through advanced natural language processing, it can scrutinize thousands of documents, revealing overlooked or under-researched areas

Identifying AI-Generated Text

AI-generated text is detected by some software with a high probability (Liu et al., 2024). But no detector may infallibly identify such a text (Morreale et al., 2024). We found that in the reviewed publications a special attention was paid to detecting AI-powered texts as compared with human-produced writings (Liu et al., 2024). Generative AI generates texts similar to human writings, but the difference potentially may be detected in their perplexity (unpredictable and diverse text) and burstiness (complexity of sentences and rare words) of the text (Krajka & Olszak, 2024). Judging by the progress ChatGPT has been making lately, the AI text generation will become quite soon more identical to the human-produced writings or will conform to any identified parameters. Already at present, efforts are made to distinguish AI-generated from human produced texts. The afore-mentioned concepts of burstiness and perplexity lay the foundation for detecting AI-generated texts. Alexander (2023) points out that AI-generated text may have lower

burstiness and lower perplexity than human writing as human writers may turn to rare words and more complex text due to “their complex thought processes and personal experiences” (Krajka & Olszak, 2024).

Inaccurate References and Non-Authentic Content

In many publications on ChatGPT and GenAI-powered language tools, authors write that AI generate “verbose overelaborate content and overused/repetitive phrases” (Shorey et al., 2024), inaccurate information (Yao et al., 2024), inaccurate citation (Rafida et al., 2024), hallucinate content, ... unquoted material (Kim, 2023).

Limitations in Creating Credible Academic Text

At present, large language models are not discipline specific. It is the key reason for their inability to create high-quality research articles at a level comparable with publications in reputed journals (Lui et al., 2024).

DISCUSSION

In attaining the objective to determine the prevailing directions of research on GenAI appliances in academic writing in tertiary education and science, the present review revisited the three aspects. First, enhancing academic writing. The findings of the review essentially follow the publications that have been brought out for the 2-3 years (Tewari et al., 2021; Thorp, 2023; Misra & Chandwar, 2023). Some new aspects either arise or manifest themselves more vividly. The updated AI technologies improved many features helpful for academic writing and essentially expanded beyond automated correction and simple text generation.

When it comes to hybrid writing that is seen as a strategy incorporating the advantages that GenAI-powered language tools may offer to their users, the authors are mainly unanimous (Parker et al., 2024). In this kind of collaboration, humans “remain accountable for fact-checking, verification procedures, and truth-telling” (Eaton, 2023). And this approach eases many risks associated with AI-powered academic writing, especially those connected with plagiarism. In the reviewed publications, we saw several approaches to plagiarism. Jarrah et al. (2023) cited the authors that saw ChatGPT as a source of information that should be properly cited (Perkins, 2023; Okaibedi, 2023). But the problem with ChatGPT long-term memory in dialogues prevents users from citing the AI-generated text. It cannot be reproduced or found after the generation.

Part of the academic community treats plagiarism differently as a concept. It remains a disputable issue. Eaton (2023) comments on discarding the term “plagiarism” as it is used. There is no universally accepted definition of plagiarism. Many practices are considered as plagiarism, including contract cheating, academic outsourcing, any misconduct in the academia (Kandeel & Eldakak, 2024). Eaton (2023) offers post-plagiarism as the new concept of plagiarism that will transcend the previous concept. It implies that the academia is on the verge of philosophical revision of the concept.

Kandeel & Eldakak (2024) refer to the Terms of Use by OpenAI³, an owner and creator of ChatGPT, regarding the services offered by ChatGPT, stating that the user is responsible for content that include both input and output. Moreover, the Terms of Use also incorporate a number of provisions that negate some ideas underlying research on authorship of AI-generated and hybrid-generated texts. The Terms of Use state:

1. You may not ...represent that Output was human-generated when it was not...

2. Output may not be unique and other users may receive similar output from our Services...
3. Output may not always be accurate...
4. Our Services may provide incomplete, incorrect, or offensive Output...
5. Any use of Outputs from our Services is at your sole risk...⁴

These clear-cut provisions eliminate partially arguments that ChatGPT may be approached as an author or co-author. Moreover, attempts to examine and prove that GenAI-powered language tools produce false or incorrect information and are not authors or vice versa are more about the scope and accuracy of wordings. Those attempts might precede the deeper and wider revision of the plagiarism concept.

CONCLUSION

The emerging field of generative AI-powered language tools is evolving, with the prevailing directions of studies: enhancing academic writing (functional aspects, content generation, assistance in writing, feedback on academic writing, learning environment for academic writing), hybrid writing as a form of the most efficient strategies in overcoming AI challenges and using the benefits of GenAI-powered language tools, challenged and concerns in the context of appliances of such tools, with a special focus on plagiarism issues of AI-generated content.

The findings proved that the previous directions of studies are still in place with some new accents, including perceptions of the strategies of appliances and hybrid academic writing as a new important concept.

The review has several limitations, including the search in one database and only for publications in English. Probably, some meaningful documents in other languages might have widened our perceptions of the research field. Further and regular (at least yearly) reviews are essential for the field that is evolving fast. New knowledge is added monthly, with an unclear though impressive perspective of the GenAI appliances in the long run.

DECLARATION OF COMPETING INTEREST

None declared.

³ OpenAI. Terms of Use. Updated Dec.11, 2024. <http://openai.com/policies/row-terms-of-use>

⁴ Ibid.

AUTHOR CONTRIBUTION

Lilia Raitskaya: conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation, visualization; writing – original draft; writing – review & editing.

Elena Tikhonova: conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing – original draft; writing – review & editing.

REFERENCES

- Ahn, S. (2024). The transformative impact of large language models on medical writing and publishing: Current applications, challenges and future directions. *The Korean Journal of Physiology & Pharmacology: Official Journal of the Korean Physiological Society and the Korean Society of Pharmacology*, 28(5), 393–401. <https://doi.org/10.4196/kjpp.2024.28.5.393>
- Alberth, A. (2023). The use of ChatGPT in writing: A blessing or a curse in disguise? *Teflin Journal*, 34(2), 337–352. <http://doi.org/10.15639/teflinjournal.v34i2/337-352>
- Alea Albada, N., & Woods, V.E. (2024). Giving credit where credit is due: An artificial intelligence contribution statement for research methods writing assignments. *Teaching of Psychology*. <http://doi.org/10.1177/00986283241259750>
- Alexander, C. (2023). *Best practices for using ChatGPT at the university of Nicosia* [Training delivered to University of Nicosia faculty]. Nicosia.
- Alkamel, M.A.A., & Alwagieh, N.A.S. (2024). Utilizing an adaptable artificial intelligence writing tool (ChatGPT) to enhance academic writing skills among Yemeni university EFL students. *Social Sciences and Humanities Open*, 10, Article 101095. <http://doi.org/10.1016/j.ssaho.2024.101095>
- Amirjalili, F., Neysani, M., & Nikbakht, A. (2024). Exploring the boundaries of authorship: A comparative analysis of AI-generated text and human academic writing in English literature. *Frontiers in Education*, 9, Article 1347421. <http://doi.org/10.3389/educ.2024.1347421>
- Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19–32. <https://doi.org/10.1080/1364557032000119616>
- Atkinson, & Bolter, J. D. (2001). *Writing space: Computers, hypertext, and the remediation of print*. Routledge.
- Biber, D. (2006). Stance in spoken and written university registers. *Journal for English for Academic Purposes*, 5, 97–116. <http://doi.org/10.1016/j.jeap.2006.05.001>
- Bolaños, F., Salatino, A., Osborne, F., & Motta, E. (2024). Artificial intelligence for literature reviews: Opportunities and challenges. *Artificial Intelligence Review*, 57(9), Article 259. <http://doi.org/10.1007/s10462-024-10902-3>
- Bozkurt, A. (2023). Generative AI, synthetic contents, open educational resources (OER), and open educational practices (OEP): A new front in the openness landscape. *Open Praxis*, 15(3), 1–7. <https://doi.org/10.55982/openpraxis.15.3.579>
- Bozkurt, A. (2024). GenAI et al.: Cocreation, authorship, ownership, academic ethics and integrity in a time of generative AI. *Open Praxis*, 16(1), 1–10. <http://doi.org/10.55982/openpraxis.16.1.654>
- Coffin, C., Curry, M.J., Goodman, S., Hewings, A., Lillis, M., & Swann, J. (2003). *Teaching academic writing; A toolkit for higher education*. Routledge.
- Eaton, S.E. (2023). Postplagiarism: transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology. *International Journal for Educational Integrity*, 19(1), Article 23. <http://doi.org/10.1007/s40979-023-00144-1>
- Fabiano, N., Gupta, A., Bhambra, N., Luu, B., Wong, S., Maaz, M., Fiedorowicz, J. G., Smith, A. L., & Solmi, M. (2024). How to optimize the systematic review process using AI tools. *JCPP Advances*, Article e12234. <http://doi.org/10.1002/jvc2.12234>
- Gralha, J.G., & Pimentel, A.S. (2024). Gotcha GPT: Ensuring the integrity in academic writing. *Journal of Chemical Information and Modeling*, 64(21), 8091–8097. <http://doi.org/10.1021/acs.jcim.4c01203>
- Gunawan, J., Aungsuroch, Y., & Montayre, J. (2024). ChatGPT integration within nursing education and its implications for nursing students: A systematic review and text network analysis. *Nurse Education Today*, 141, Article 106323. <http://doi.org/10.1016/j.nedt.2024.106323>
- Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, 15(4), Article ep464. <http://doi.org/10.30935/cedtech/13605>

- Jarrah, A.M., Wardat, Y., & Fidalgo, P. (2023). Using ChatGPT in academic writing is (not) a form of plagiarism: What does the literature say? *Online Journal of Communication and Media Technologies*, 13(4), Article e202346. <http://doi.org/10.30935/ojcm/13572>
- Johnston, H., Wells, R.F., Shanks, E.M., Boey, T., & Parsons, B.N. (2024). Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity*, 20(1), Article 2. <http://doi.org/10.1007/s40979-024-00149-4>
- Kandeel, M.E., & Eldakak, A. (2024). Legal dangers of using ChatGPT as a co-author according to academic research regulations. *Journal of Governance and Regulation*, 13(1 Special issue), 289-298. <http://doi.org/10.22495/jgrv13i1siart3>
- Khalifa, M., & Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update*, Article 100145. <https://doi.org/10.1016/j.cmpbup.2024.100145>
- Khlaif, Z.N., Mousa, A., Hattab, M.K., Itmazi, J., Hassan, A.A., Sanmugam, M., & Ayyoub, A. (2023). The Potential and Concerns of Using AI in Scientific Research: ChatGPT Performance Evaluation. *Medical Education*, 9(1), Article e47049. <http://doi.org/10.2196/47049>
- Kim, J., Yu, S., Detrick, R., & Li, N. (2024). Exploring students' perspectives on Generative AI-assisted academic writing. *Education and Information Technologies*. <http://doi.org/10.1007/s10639-024-12878-7>
- Kim, S.J. (2024a). Research ethics and issues regarding the use of ChatGPT-like artificial intelligence platforms by authors and reviewers: a narrative review. *Science Editing*, 11(2), 96-106. <http://doi.org/10.6087/kcse.343>
- Kim, S.G. (2023). Using ChatGPT for language editing in scientific articles. *Maxillofacial Plastic and Reconstructive Surgery*, 45, 13. <https://doi.org/10.1186/s40902-023-00381-x>
- Kim, S.J. (2024b). Trends in research on ChatGPT and adoption-related issues discussed in articles: A narrative review. *Science Editing*, 11, 3-11. <https://doi.org/10.6087/kcse.321>
- Kohnke, L. (2024). Exploring EAP students' perceptions of GenAI and traditional grammar-checking tools for language learning. *Computers and Education: Artificial Intelligence*, 7, Article 100279. <http://doi.org/10.1016/j.caeai.2024.100279>
- Krajka, J., & Olszak, I. (2024a) "AI, will you help?" How learners use Artificial Intelligence when writing. *XLinguae*, 17(1), 34-48. <http://doi.org/10.18355/XL.2024.17.01.03>
- Krajka, J., & Olszak, I. (2024b). Artificial intelligence tools in academic writing instruction: Exploring the potential of on-demand AI assistance in the writing process. *Roczniki Humanistyczne*, 72(6), 123-140. <http://doi.org/10.18290/rh247206.8>
- Kurt, G., & Kurt, Y. (2024). Enhancing L2 writing skills: ChatGPT as an automated feedback tool. *Journal of Information Technology Education: Research*, 23, Article 24. <http://doi.org/10.28945/5370>
- Li, J., Zong, H., Wu, E., Wu, R., Peng, Z., Zhao, J., Yang, L., Xie, H., & Shen, B. (2024). Exploring the potential of artificial intelligence to enhance the writing of English academic papers by non-native English-speaking medical students - the educational application of ChatGPT. *BMC Medical Education*, 24(1), Article 736. <http://doi.org/10.1186/s12909-024-05738-y>
- Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., Huang, X., Zhu, D., Cai, H., Li, Q., Liu, T., & Li, X. (2023). Differentiating ChatGPT-generated and human-written medical texts: Quantitative study. *JMIR Medical Education*, 9, Article e48904. <https://doi.org/10.2196/48904>
- Liu, J.Q.J., Hui, K.T.K., Al Zoubi, F., Zhou, Z.Z.X., Samartzis, D., Yu, C.C.H., Chang, J.R., & Wong, A.Y.L. (2024). The great detectives: Humans versus AI detectors in catching large language model-generated medical writing. *International Journal for Educational Integrity*, 20(1), Article 8. <http://doi.org/10.1007/s40979-024-00155-6>
- Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study. *Smart Learning Environments*, 11(1), Article 9. <http://doi.org/10.1186/s40561-024-00295-9>
- Mahyoob, M., Algaraady, J., & Alblwi, A. (2023). A proposed framework for human-like language processing of ChatGPT in academic writing. *International Journal of Emerging Technologies in Learning*, 18(14), 282-293. <http://doi.org/10.3991/ijet.v18i14.41725>
- Malik, A.R., Pratiwi, Y., Andajani, K., Numertayasa, I.W., Suharti, S., Darwis, A., Marzuki (2023). Exploring artificial intelligence in academic essay: Higher education student's perspective. *International Journal of Educational Research Open*, 5, Article 100296. <http://doi.org/10.1016/j.ijedro.2023.100296>
- Maphoto, K.B., Sevnarayan, K., Mohale, N.E., Suliman, Z., Ntsopi, T.J., & Mokoena, D. Advancing (2024). Students' academic excellence in distance education: Exploring the potential of generative AI integration to improve academic writing skills. *Open Praxis*, 16(2), 142-159. <http://doi.org/10.55982/openpraxis.16.2.649>

- Miao, J., Charat, T., Supawadee, S., Garcia Valencia, O.A., Qureshi, F., & Cheungpasitporn (2024). Ethical dilemmas in using AI for academic writing and an example framework for peer review in nephrology academia: A narrative review. *Clinics and Practice*, 14(10), 89-105. <http://doi.org/10.3390/clinpract14010008>
- Misra, D.P., & Chandwar, K. (2023). ChatGPT, artificial intelligence and scientific writing: What authors, peer reviewers and editors should know? *Journal of the Royal College of Physicians of Edinburgh*, 1-4. <http://doi.org/10.1177/14782715231181023>
- Mohammad, T., Alzubi, A.A.F., Nazim, M., & Khan, S.I. (2024). Paraphrasing prowess: Unveiling the insights of EFL students and teachers on QuillBot mastery. *International Journal of Information and Education Technology*, 14(5), 642-650. <http://doi.org/10.18178/ijiet.2024.14.5.2088>
- Mondal, H., & Mondal, S. (2023). ChatGPT in academic writing: Maximizing its benefits and minimizing the risks. *Indian Journal of Ophthalmology*, 71(12), 3600-366. <http://doi.org/10.1002/jcv2.12234>
- Morreale, M.K., Balon, R., Beresin, E.V., Seritan, A., Castillo, E.G., Thomas, L.A., Louie, A.K., Aggarwal, R., Guerrero, A.P.S., Coverdale, J., & Brenner, A.M. (2024). Artificial intelligence and medical education, academic writing, and journal policies: A focus on Large Language Models. *Academic Psychiatry*. <http://doi.org/10.1007/s40596-024-02071-w>
- Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, 49(5), 847-864. <http://doi.org/10.1080/03075079.2024.2323593>
- Ou, A.W., Stöhr, C., & Malmström, H. (2024). Academic communication with AI-powered language tools in higher education: From a post-humanist perspective. *System*, 121, Article 103225. <http://doi.org/10.1016/j.system.2024.103225>
- Okaibedi, D. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology*, 13, Article 100060. <https://doi.org/10.1016/j.jrt.2023.100060>
- Parker, J.L., Richard, V.M., Acabá, A., Escoffier, S., Flaherty, S., Jablonka, S., & Becker, K.P. (2024). Negotiating meaning with machines: AI's role in doctoral writing pedagogy. *International Journal of Artificial Intelligence in Education*. <http://doi.org/10.1007/s40593-024-00425-x>
- Perkins, M. (2023). Academic Integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, 20(2). <https://doi.org/10.53761/1.20.02.07>
- Rababah, L.M., Rababah, M.A., & Al-Khawaldeh, N.N. (2024). Graduate students' ChatGPT experience and perspectives during thesis writing. *International Journal of Engineering Pedagogy*, 14(3), 22-35. <http://doi.org/10.3991/ijep.v14i3.48395>
- Rafida, T., Suwandi, S., Ananda, R. (2024). EFL students' perception in Indonesia and Taiwan on using artificial intelligence to enhance writing skills. *Jurnal Ilmiah Peuradeun*, 12(3), 987-1016. <http://doi.org/10.26811/peuradeun.v12i3.1520>
- Raitskaya, L., & Lambovska, M. (2024). Prospects for ChatGPT application in higher education: A scoping review of international research. *Integration of Education*, 28(1), 10-21. <https://doi.org/10.15507/1991-9468.114.028.202401.010-021>
- Rao, K.N., Arora, R.D., Dange, P., & Nagarkar, N.M. (2023). NLP AI models for optimizing medical research: Demystifying the concerns. *Indian Journal of Surgical Oncology*, 14, 854-858. <https://doi.org/10.1007/s13193-023-01791-z>
- Shorey, S., Mattar, C., Pereira, T.L.-B., & Choolani, M. (2024). A scoping review of ChatGPT's role in healthcare education and research. *Nurse Education Today*, 135, 106121. <http://doi.org/10.1016/j.nedt.2024.106121>
- Tarchi, C., Zappoli, A., Casado Ledesma, L., & Brante, E.W. (2024). The use of ChatGPT in source-based writing tasks. *International Journal of Artificial Intelligence in Education*. <http://doi.org/10.1007/s40593-024-00413-1>
- Temsah, O., Khan, S.A., Chaiah, Y., Senjab, A., Alhasan, K., Jamal, A., Aljamaan, F., Malki, K.H., Halwani, R., Al-Tawfiq, J.A., Temsah, M.-H., & Al-Eyadhy, A. (2023). Overview of early ChatGPT's presence in medical literature: Insights from a hybrid literature review by ChatGPT and human experts. *Cureus*, 15, Article e37281. <https://doi.org/10.7759/cureus.37281>
- Teng, M.F. (2023). Scientific writing, reviewing, and editing for open-access TESOL journals: The role of ChatGPT. *International Journal of TESOL Studies*, 5(1), 87-91. <http://doi.org/10.58304/ijts.20230107>
- Tewari, S., Zabounidis, R., Kothari, A., Bailey, R., & Alm, C.O. (2021). Perceptions of human and machine-generated articles. *Digital Threats: Research and Practice*, 2(2), Article 12. <https://doi.org/10.1145/3428158>
- Thorp, H.H. (2023). ChatGPT is fun, but not an author. *Science*, 379(6630), Article 313. <http://doi.org/10.1126/science.adg7879>
- Tikhonova, E., & Raitskaya, L. (2023). ChatGPT: Where is a silver lining? Exploring the realm of GPT and Large Language Models. *Journal of Language and Education*, 9(3), 5-11. <https://doi.org/10.17323/jle.2023.18119>
- Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., Moher, D., Peters, M.D.J., Horsley, T., Weeks, L., Hempel, S., & Akl, E.A. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467-73. <https://doi.org/10.7326/M18-0850>

-
- Tusting, K., McCulloch, S., Bhatt, I., Hamilton, M., & Barton, D. (2019). *Academics writing. The dynamics of knowledge creation*. Routledge.
- Utami, S.P.T., Andayani, Winarni, R., Sumarwati (2023). Utilization of artificial intelligence technology in an academic writing class: How do Indonesian students perceive? *Contemporary Educational Technology*, 15(4), Article ep450. <http://doi.org/10.30935/cedtech/13419>
- Williams, A. (2024). Comparison of generative AI performance on undergraduate and postgraduate written assessments in the biomedical sciences. *International Journal of Educational Technology in Higher Education*, 21(1), Article 52. <http://doi.org/10.1186/s41239-024-00485-y>
- Yao, Y., Sun, Y., Zhu, S., & Zhu, X. (2024). A Qualitative inquiry into metacognitive strategies of postgraduate students in employing ChatGPT for English academic writing. *European Journal of Education*, Article e12824. <http://doi.org/10.1111/ejed.12824>

APPENDIX 1

The publications under review

- Ahn, S. (2024). The transformative impact of large language models on medical writing and publishing: current applications, challenges and future directions. *The Korean Journal of Physiology & Pharmacology: Official Journal of the Korean Physiological Society and the Korean Society of Pharmacology*, 28(5), 393–401. <https://doi.org/10.4196/kjpp.2024.28.5.393>
- Alberth, A. (2023). The use of ChatGPT in writing: A blessing or a curse in disguise? *Teflin Journal*, 34 (2), 337-352. <http://doi.org/10.15639/teflinjournal.v34i2/337-352>
- Alea Albada, N., & Woods, V.E. (2024). Giving credit where credit is due: An artificial intelligence contribution statement for research methods writing assignments. *Teaching of Psychology*. <https://doi.org/10.1177/00986283241259750>
- Alkamel, M.A.A., & Alwagieh, N.A.S. (2024). Utilizing an adaptable artificial intelligence writing tool (ChatGPT) to enhance academic writing skills among Yemeni university EFL students. *Social Sciences and Humanities Open*, 10, Article 101095. <http://doi.org/10.1016/j.ssaho.2024.101095>
- Amirjalili, F., Neysani, M., & Nikbakht, A. (2024). Exploring the boundaries of authorship: A comparative analysis of AI-generated text and human academic writing in English literature. *Frontiers in Education*, 9, Article 1347421. <http://doi.org/10.3389/educ.2024.1347421>
- Bolaños, F., Salatino, A., Osborne, F., & Motta, E. (2024). Artificial intelligence for literature reviews: Opportunities and challenges. *Artificial Intelligence Review*, 57(9), Article 259. <http://doi.org/10.1007/s10462-024-10902-3>
- Bozkurt, A. (2024). GenAI et al.: Cocreation, authorship, ownership, academic ethics and integrity in a time of generative AI. *Open Praxis*, 16(1), 1-10. <http://doi.org/10.55982/openpraxis.16.1.654>
- Fabiano, N., Gupta, A., Bhambra, N., Luu, B., Wong, S., Maaz, M., Fiedorowicz, J.G., Smith, A.L., & Solmi, M. (2024). How to optimize the systematic review process using AI tools. *JCPP Advances*, Article e12234. <http://doi.org/10.1002/jvc2.12234>
- Gralha, J.G., & Pimentel, A.S. (2024). Gotcha GPT: Ensuring the Integrity in academic writing. *Journal of Chemical Information and Modeling*, 64(21), 8091-8097. <http://doi.org/10.1021/acs.jcim.4c01203>
- Gunawan, J., Aunguroch, Y., & Montayre, J. (2024). ChatGPT integration within nursing education and its implications for nursing students: A systematic review and text network analysis. *Nurse Education Today*, 141, Article 106323. <http://doi.org/10.1016/j.nedt.2024.106323>
- Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, 15(4), Article ep464. <http://doi.org/10.30935/cedtech/13605>
- Jarrah, A.M., Wardat, Y., & Fidalgo, P. (2023). Using ChatGPT in academic writing is (not) a form of plagiarism: What does the literature say? *Online Journal of Communication and Media Technologies*, 13(4), Article e202346. <http://doi.org/10.30935/ojcm/13572>
- Johnston, H., Wells, R.F., Shanks, E.M., Boey, T., & Parsons, B.N. (2024). Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity*, 20(1), Article 2. <http://doi.org/10.1007/s40979-024-00149-4>
- Kandeel, M.E., & Eldakak, A. (2024). Legal dangers of using ChatGPT as a co-author according to academic research regulations. *Journal of Governance and Regulation*, 13(1 Special issue), 289-298. <http://doi.org/10.22495/jgrv13i1siart3>
- Khlaif, Z.N., Mousa, A., Hattab, M.K., Itmazi, J., Hassan, A.A., Sanmugam, M., & Ayyoub, A. (2023). The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. *Medical Education*, 9(1), Article e47049. <http://doi.org/10.2196/47049>
- Khalifa, M., & Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update*, Article 100145. <https://doi.org/10.1016/j.cmpbup.2024.100145>
- Kim, J., Yu, S., Detrick, R., & Li, N. (2024). Exploring students' perspectives on Generative AI-assisted academic writing. *Education and Information Technologies*. <http://doi.org/10.1007/s10639-024-12878-7>
- Kim, S.-J. (2024). Research ethics and issues regarding the use of ChatGPT-like artificial intelligence platforms by authors and reviewers: A narrative review. *Science Editing*, 11(2), 96-106. <http://doi.org/10.6087/kcse.343>
- Kohnke, L. (2024). Exploring EAP students' perceptions of GenAI and traditional grammar-checking tools for language learning. *Computers and Education: Artificial Intelligence*, 7, Article 100279. <http://doi.org/10.1016/j.caeai.2024.100279>

- Krajka, J., & Olszak, I. (2024). "AI, will you help?" How learners use Artificial Intelligence when writing. *XLinguae*, 17(1), 34-48. <http://doi.org/10.18355/XL.2024.17.01.03>
- Krajka, J., & Olszak, I. (2024). Artificial intelligence tools in academic writing instruction: Exploring the potential of on-demand AI assistance in the writing process. *Roczniki Humanistyczne*, 72(6), 123-140. <http://doi.org/10.18290/rh247206.8>
- Kurt, G., & Kurt, Y. (2024). Enhancing L2 writing skills: ChatGPT as an automated feedback tool. *Journal of Information Technology Education: Research*, 23, Article 24. <http://doi.org/10.28945/5370>
- Li, J., Zong, H., Wu, E., Wu, R., Peng, Z., Zhao, J., Yang, L., Xie, H., & Shen, B. (2024). Exploring the potential of artificial intelligence to enhance the writing of English academic papers by non-native English-speaking medical students - The educational application of ChatGPT. *BMC Medical Education*, 24(1), Article 736. <http://doi.org/10.1186/s12909-024-05738-y>
- Liu, J.Q.J., Hui, K.T.K., Al Zoubi, F., Zhou, Z.Z.X., Samartzis, D., Yu, C.C.H., Chang, J.R., & Wong, A.Y.L. (2024). The great detectives: humans versus AI detectors in catching large language model-generated medical writing. *International Journal for Educational Integrity*, 20(1), Article 8. <http://doi.org/10.1007/s40979-024-00155-6>
- Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, 11(1), Article 9. <http://doi.org/10.1186/s40561-024-00295-9>
- Mahyoob, M., Algaraady, J., & Alblwi, A. (2023). A proposed framework for human-like language processing of ChatGPT in academic writing. *International Journal of Emerging Technologies in Learning*, 18(14), 282-293. <http://doi.org/10.3991/ijet.v18i14.41725>
- Malik, A.R., Pratiwi, Y., Andajani, K., Numertayasa, I.W., Suharti, S., Darwis, A., Marzuki (2023). Exploring artificial intelligence in academic essay: Higher education student's perspective. *International Journal of Educational Research Open*, 5, Article 100296. <http://doi.org/10.1016/j.ijedro.2023.100296>
- Maphoto, K.B., Sevnarayan, K., Mohale, N.E., Suliman, Z., Ntsopi, T.J., & Mokoena, D. Advancing (2024). Students' academic excellence in distance education: Exploring the potential of generative AI integration to improve academic writing skills. *Open Praxis*, 16(2), 142-159. <https://doi.org/10.55982/openpraxis.16.2.649>
- Miao, J., Charat, T., Supawadee, S., Garcia Valencia, O.A., Qureshi, F., & Cheungpasitporn (2024). Ethical dilemmas in using AI for academic writing and an example framework for peer review in nephrology academia: A narrative review. *Clinics and Practice*, 14(10), 89-105. <http://doi.org/10.3390/clinpract14010008>
- Mohammad, T., Alzubi, A.A.F., Nazim, M., & Khan, S.I. (2024). Paraphrasing prowess: Unveiling the insights of EFL students and teachers on QuillBot mastery. *International Journal of Information and Education Technology*, 14(5), 642-650. <http://doi.org/10.18178/ijet.2024.14.5.2088>
- Mondal, H., & Mondal, S. (2023). ChatGPT in academic writing: Maximizing its benefits and minimizing the risks. *Indian Journal of Ophthalmology*, 71(12), 3600-366. <http://doi.org/10.1002/jcv2.12234>
- Morreale, M.K., Balon, R., Beresin, E.V., Seritan, A., Castillo, E.G., Thomas, L.A., Louie, A.K., Aggarwal, R., Guerrero, A.P.S., Coverdale, J., & Brenner, A.M. (2024). Artificial intelligence and medical education, academic writing, and journal policies: A focus on Large Language Models. *Academic Psychiatry*. <http://doi.org/10.1007/s40596-024-02071-w>
- Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, 49(5), 847-864. <http://doi.org/10.1080/03075079.2024.2323593>
- Ou, A.W., Stöhr, C., & Malmström, H. (2024). Academic communication with AI-powered language tools in higher education: From a post-humanist perspective. *System*, 121, Article 103225. <http://doi.org/10.1016/j.system.2024.103225>
- Parker, J.L., Richard, V.M., Acabá, A., Escoffier, S., Flaherty, S., Jablonka, S., & Becker, K.P. (2024). Negotiating meaning with machines: AI's role in doctoral writing pedagogy. *International Journal of Artificial Intelligence in Education*. <http://doi.org/10.1007/s40593-024-00425-x>
- Rababah, L.M., Rababah, M.A., & Al-Khawaldeh, N.N. (2024). Graduate students' ChatGPT experience and perspectives during thesis writing. *International Journal of Engineering Pedagogy*, 14(3), 22-35. <http://doi.org/10.3991/ijep.v14i3.48395>
- Rafida, T., Suwandi, S., Ananda, R. (2024). EFL students' perception in Indonesia and Taiwan on using artificial intelligence to enhance writing skills. *Jurnal Ilmiah Peuradeun*, 12(3), 987-1016. <http://doi.org/10.26811/peuradeun.v12i3.1520>
- Shorey, S., Mattar, C., Pereira, T.L.-B., & Choolani, M. (2024). A scoping review of ChatGPT's role in healthcare education and research. *Nurse Education Today*, 135, 106121. <http://doi.org/10.1016/j.nedt.2024.106121>
- Tarchi, C., Zappoli, A., Casado Ledesma, L., & Brante, E.W. (2024). The use of ChatGPT in source-based writing tasks. *International Journal of Artificial Intelligence in Education*. <http://doi.org/10.1007/s40593-024-00413-1>
- Teng, M.F. (2023). Scientific writing, reviewing, and editing for open-access TESOL journals: The role of ChatGPT. *International Journal of TESOL Studies*, 5(1), 87-91. <http://doi.org/10.58304/ijts.20230107>

-
- Tikhonova, E., & Raitskaya, L. (2023). ChatGPT: Where is a silver lining? Exploring the realm of GPT and Large Language Models. *Journal of Language and Education*, 9(3), 5-11. <https://doi.org/10.17323/jle.2023.18119>
- Utami, S.P.T., Andayani, Winarni, R., Sumarwati (2023). Utilization of artificial intelligence technology in an academic writing class: How do Indonesian students perceive? *Contemporary Educational Technology*, 15(4), Article ep450. <http://doi.org/10.30935/cedtech/13419>
- Williams, A. (2024). Comparison of generative AI performance on undergraduate and postgraduate written assessments in the biomedical sciences. *International Journal of Educational Technology in Higher Education*, 21(1), Article 52. <http://doi.org/10.1186/s41239-024-00485-y>
- Yao, Y., Sun, Y., Zhu, S., & Zhu, X. (2024). A qualitative inquiry into metacognitive strategies of postgraduate students in employing ChatGPT for English academic writing. *European Journal of Education*, Article e12824. <http://doi.org/10.1111/ejed.12824>

APPENDIX 2

Challenges and concerns relating to applying GenAI-powered tools in academic writing: Raw Data from the documents under review

Issues	Reference	Raw Data
Major ethical considerations	Khalifa & Albadawy, 2024	...major ethical considerations are identified, such as the importance of human intelligence in research and the limitations that AI tools might enforce in guiding research ideas and design...
Potential violation of academic integrity	Yao et al., 2024	...caution should be taken to employ GenAI technology in academic writing, as concerns regarding the violation of academic integrity have been repeatedly raised by scholars (Barrot 2023; Hosseini, Rasmussen, and Resnik 2023; Yan 2023)...
Overreliance Equity	Kohnke, 2024	...there are concerns about the reliability of GenAI tools and potential overreliance on them, particularly for complex writing tasks that require critical thinking and creativity (Wang, 2024)... the cost of premium features raises significant concerns about equity. ...Institutional support and modified teaching methods can ensure equity and accessibility ...
	Rafida et al., 2024	Over-reliance on AI can lead to outputs lacking variety (Aisiy, 2024; Malik et al., 2023).
Lack of transparency in content generation	Shorey et al., 2024	ChatGPT's lack of transparency in content generation, high-recall retrieval inadequacies, and need for technical proficiency to formulate precise prompts that yield organized responses
Rephrasing and dishonesty	Rafida et al., 2024	AI use in rephrasing without acknowledgment can lead to academic dishonesty (Marzuki et al., 2023).
No detectors to infallibly identify an AI-generated text	Morreale et al., 2024	To date, there is no software available to reliably identify that an LLM has been used to create a document. Given this limitation, educators need to decide if assessment methods should change
	Krajka & Olszak, 2024	Human writers, with their complex thought processes and personal experiences, can produce more diverse and less predictable text...
Inability to create credible academic articles	Lui et al., 2024	...generative AI is less likely to successfully create credible academic articles without the development of discipline-specific LLMs
Propensity to generate inaccurate references	Shorey et al., 2024	ChatGPT's propensity to generate inaccurate or inadequate references, or cite non-existent sources, further challenges the overall credibility and veracity of the generated content
Misleading content	Lui et al., 2024	Its [ChatGPT's] tendency to generate plausible but non-rigorous or misleading content has raised doubts about the reliability of its outputs (Sallam 2023; Manohar & Prasad 2023). This poses a risk of disseminating unsubstantiated information...
Generation of non-authentic (i.e. sourced-based plagiarism) and verbose content	Shorey et al., 2024	ChatGPT's generation of 'robotic', non-authentic (i.e. sourced-based plagiarism), verbose overelaborate content and overused/repetitive phrases places an additional burden for researchers who must carefully review the output to render it more 'natural' or 'human-like'
	Yao et al., 2024	The presence of inaccurate information within the output generated by GenAI diminishes students' trust in technological reliability (Nugroho et al. 2024; Zou and Huang 2024)
	Morreale et al., 2024	[LLMs] ... can misattribute information and create or "hallucinate" false statements and references
	Kim, 2024	...the quality of training data, bias in learning outcomes, the accuracy of generated content, and potential issues with unquoted material and plagiarism (Rao et al., 2023)

Issues	Reference	Raw Data
Inaccurate responses and content	Kim, 2024	GenAI has been widely reported to hallucinate content or provide incorrect guidance, which refers to when a GenAI tool generates inaccurate responses that seem realistic (Alkaissi & McFarlane, 2023)
	Amirjalili et al., 2024	Despite its [ChatGPT] ability to generate impressive outputs, the model may produce inaccuracies or nonsensical responses, demonstrating awareness of context and wording in input (Bender et al., 2021)...
Detrimental to creative and critical thinking	Yao et al., 2024	...excessive dependence on GenAI technology can be detrimental to 'creative and critical thinking and the ability to make independent judgments about the quality of writing' (Huang and Tan 2023, 1151)
Lack of author's voice	Amirjalili et al., 2024	Conventional aspects of authorship will unavoidably change as AI becomes more prevalent in writing tasks, challenging the notion of voice as an extension of the human author. As ChatGPT generates text, the representation of voice undergoes a significant transformation, raising questions about the role of individual identity in AI-authored content Authorship, as emphasized by Charmaz and Mitchell (1996), encapsulates the core of the writer's voice and presence in written works. Ivanič (1998) extends this notion, positing that writing serves as a socio-political medium for expressing identity. The academic realm, however, introduces complexities, notably around the contested concept of "voice." Tardy (2012) acknowledges the broad spectrum of meanings attributed to "voice," while Atkinson (2001) and Bibber (2006) consider it a critical language aspect shaped by genre and community constraints in academic writing...
Other concerns	Yao et al., 2024	...students also identified several weaknesses of ChatGPT that are less frequently reported in the literature, including the use of esoteric vocabulary and an impersonal writing style, which may be attributed to the distinctive context of academic writing...
	Kurt & Kurt, 2024	Occasional inconsistencies, dependance on the quality of prompts, absence of human-like voice, risk of over-dependence on automated feedback
	Krajka & Olszak, 2024	Large language models cannot store new experiences in long-term memory in dialogue, so without finetuning they have to start dialoguing "afresh" with each person they meet Sejnowski (2023). Since they are huge, they find it difficult to maintain continuity during long dialogues
	Kim et al., 2024	GenAI can reflect any bias contained in training data or held by its developers

APPENDIX 3

Authorship and integrity-related issues of Generative Artificial Intelligence in academic writing: Raw Data from the documents under review

Reference	Raw Data
Williams, 2024	Many in the higher education (HE) sector are concerned that students will use generative AI to produce written assignments and therefore as a tool for plagiarism (Perkins, 2023)
Liu et al., 2024	...scientists did not support granting ChatGPT authorship in academic publishing because it could not be held accountable for the ethics of the content (Stokel-Walker, 2023)...
Kim, 2024	The guidelines for AI use in most journals, except Science, are summarized as follows: AI cannot be credited with authorship; human authors must assume full responsibility for the accuracy of the results; and there must be clear disclosure of the AI tools utilized
Kim et al., 2024	Within the academic community, there is genuine concern that use of AI will result in cases of plagiarism, particularly if academic writers do not think critically about the suggestions made by an AI and merely adopt and use whatever it recommends (Salvagno et al., 2023)
Morreale et al., 2024	...the authors pointed out that "LLMs do not enable scientists to cheat; scientific fraud has existed long before their advent, but LLMs simply make it easier"...
Kim, 2024	Given the distinct characteristics of text written by humans and AI, algorithms can effectively detect AI-generated content (Liao et al., 2023). In evaluating the AI detection performance of tools like GPTZero (GPTZero Inc), ZeroGPT, Writer AI content detector (Writer Inc), and Originality (Originality.AI Inc), Originality excelled in distinguishing between AI-generated and human texts... When reviewers were given AI-generated abstracts along with the full text of the journal, they achieved a 93% accuracy rate in identifying ChatGPT abstracts
Amirjalili et al., 2024	Authorship, extending beyond the act of writing, encompasses ownership, accountability, and the integrity of ideas...
Borkurt, 2024	The debate extends to whether generative AI can be acknowledged as a co-author. Some have credited generative AI as a co-author (See O'Connor & ChatGPT, 2023; O'Connor, 2023). Some others argue that using generative AI does not diminish human responsibility (Dien, 2023) and point to the overlooked contributions of the unnamed/invisible authors who trained these AI algorithms (Dwivedi et al., 2023; Lund et al., 2023). In some instances, generative AI is treated as a ghost contributor, acknowledging a passive contribution in content creation (Rahimi & Talebi Bezzmin Abadi, 2023; Teixeira da Silva & Tsigaris, 2023) ...this paper proposes that a final human approval statement should be articulated. In this context, this paper suggests Academic Integrity and Transparency in AI-assisted Research and Specification (aiTARAS) Framework for acknowledging and disclosing the use of generative AI in scholarly writing, to maintain academic integrity, transparency and ethics: Direct Contribution... General Assistance... Specific Sections... Idea Development... Editing and Reviewing... Language Translation and Localization... Data Analysis... Data Visualization... Code or Algorithms...
Jarrah et al., 2023	Ethical considerations in training data: When creating custom AI models or fine-tuning ChatGPT, be mindful of the data used for training. Ensure that the data is ethically sourced, does not perpetuate biases, and adheres to the principles of responsible AI development.

<https://doi.org/10.17323/jle.2024.22443>

Hope Speech Detection Using Social Media Discourse (Posi-Vox-2024): A Transfer Learning Approach

Muhammad Ahmad ¹, Sardar Usman ², Humaira Farid ³, Iqra Ameer ⁴,
Muhammad Muzammil ⁵, Ameer Hamza ⁵, Grigori Sidorov ¹, Ildar Batyrshin ¹

¹ Instituto Politecnico Nacional (CIC-IPN), Mexico City, Mexico

² Institute of Arts and Culture, Lahore, Pakistan

³ Independent Researcher, California, USA

⁴ Pennsylvania State University at Abington, PA, USA

⁵ Islamia University of Bahawalpur, Pakistan

ABSTRACT

Background: The notion of hope is characterized as an optimistic expectation or anticipation of favorable outcomes. In the age of extensive social media usage, research has primarily focused on monolingual techniques, and the Urdu and Arabic languages have not been addressed.

Purpose: This study addresses joint multilingual hope speech detection in the Urdu, English, and Arabic languages using a transfer learning paradigm. We developed a new multilingual dataset named Posi-Vox-2024 and employed a joint multilingual technique to design a universal classifier for multilingual dataset. We explored the fine-tuned BERT model, which demonstrated a remarkable performance in capturing semantic and contextual information.

Method: The framework includes (1) preprocessing, (2) data representation using BERT, (3) fine-tuning, and (4) classification of hope speech into binary ('hope' and 'not hope') and multi-class (realistic, unrealistic, and generalized hope) categories.

Results: Our proposed model (BERT) demonstrated benchmark performance to our dataset, achieving 0.78 accuracy in binary classification and 0.66 in multi-class classification, with a 0.04 and 0.08 performance improvement over the baselines (Logistic Regression, in binary class 0.75 and multi class 0.61), respectively.

Conclusion: Our findings will be applied to improve automated systems for detecting and promoting supportive content in English, Arabic and Urdu on social media platforms, fostering positive online discourse. This work sets new benchmarks for multilingual hope speech detection, advancing existing knowledge and enabling future research in underrepresented languages.

KEYWORDS

hope speech, BERT, machine learning, twitter analysis, social media, transfer learning, NLP

Citation: Ahmad, M., Sardar, U., Humaira, F., Iqra A., Muzzamil, M., Hmaza, A., Sidorov, G., & Batyrshin, I. (2024). Hope Speech detection using social media discourse (Posi-Vox-2024): A transfer learning approach. *Journal of Language and Education*, 10(4), 31-43. <https://doi.org/10.17323/jle.2024.22443>

Correspondence:
Ildar Batyrshin,
batyr1@cic.ipn.mx

Received: September 7, 2024

Accepted: December 16, 2024

Published: December 30, 2024



INTRODUCTION

Hope is defined as a positive emotional state that includes expectations or anticipation of beneficial outcomes in the future. Many online social media platforms have provided a space for millions of users to voice their thoughts and share their views. This opportunity not only generated negative content but also fostered the exchange of positive ideas (Alawadh et al., 2023) and promoted positivity. Recently, hope speech detection on social media has gained significant attention, with few studies addressing this

issue across both high- and low-resource languages (Arif et al., 2024; Balouchzahi et al., 2023; Chakravarthi, 2022). Hope speech detection is relatively new approach that focuses on identifying and amplifying positive online content to promote social harmony and encourage a more positive atmosphere within communities. Among the limited studies on hope speech detection, research has primarily focused on monolingual contexts, developing individual classification models tailored to each language, such as English (Balouchzahi et al., 2023), Spanish (Kumar et al., 2022), English, Tamil

and Malayalam (RamakrishnaIyer et al., 2023), and Bengali (Nath et al. 2023), while Arabic and Urdu languages have not been addressed in either monolingual or multilingual contexts.

For many, social media has become a vital platform for seeking support (Gowen et al., 2012; Yates et al., 2017; Wang & Jurgens, 2018). Social integration is essential for their overall well-being, particularly for those vulnerable to exclusion. By identifying and amplifying encouraging messages on social media, hope speech detection can contribute to a more equitable and inclusive digital landscape. Additionally, the methodology developed in this study has broad applications in psycholinguistics and natural language processing, where it can be used to identify positive sentiment, resilience, and constructive discourse across various contexts.

Social media platforms host numerous hateful or malicious posts (Louati, Ali, et al., 2024; Irfan, Asim, et al., 2024; Anjum, and Rahul Katarya., 2024), largely because of the lack of regulatory authority. Analyzing content on Twitter and other platforms has proven effective in curbing the spread of negativity through techniques like hate speech detection (Schmidt & Wiegand, 2017, Subramanian, Malliga, et al., 2023; Nagar, Barbhuiya, & Dey, 2023), offensive language identification (Anand et al., 2023; Kogilavani et al., 2023; Mnassri et al., 2024), and abusive language detection (Zampieri et al., 2019; Austin et al., 2020; Yenala et al., 2018). Nonetheless, as highlighted by recent research, existing technologies for detecting abusive language (Lee et al., 2018) often fail to account for the potential biases inherent in the datasets upon which they are trained. The presence of systematic racial biases within these datasets can render abusive language detection systems inherently biased, leading to discriminatory outcomes that disproportionately affect minority or marginalized groups. Such biases in language detection technology have the potential to perpetuate discrimination (Davidson et al., 2019). Therefore, we should prioritize promoting positive interactions rather than merely addressing individual negative posts. In this context, hope speech detection offers a novel approach, not only by counteracting negativity but also by contributing to a more positive and inclusive online environment across a wide range of linguistic and cultural contexts. To achieve this objective, we have created a comprehensive joint multilingual hope speech corpus for Urdu, Arabic, and English, using binary and multi classification. The process begins by collecting data related to hope speech tweets in English, Urdu, and Arabic from Twitter. After the collection of the dataset we pre-processed each sample to make it more robust for machine learning models. After pre-processing, the data goes through a joint multilingual process, where the English, Urdu, and Arabic datasets are combined. In the annotation phase, the data is labeled according to specific guidelines. The next step involves fine-tuning our proposed models and applying them to the dataset for classification tasks. Finally, the different machine learning and deep learning and trans-

former based models are evaluated for accuracy, F1-score, recall, and precision, and the results are analyzed for binary and multi-class classification tasks. This methodology provides a comprehensive approach to hope speech detection across different languages.

This study makes the following contributions:

1. To the best of our knowledge, joint multilingual hope speech detection for English, Urdu, and Arabic has not been developed earlier, and we have explored a comprehensive joint multilingual corpus with extensive guidelines for annotating the dataset;
2. We explored hope speech detection as a two-level text classification task for the first time in joint multilingual dataset (English, Arabic, and Urdu) languages and propose a multiclass classification approach for Urdu and Arabic languages;
3. A comprehensive series of experiments demonstrated that the proposed methodology achieved the best performance compared to the baseline;
4. The proposed framework demonstrated a 0.78 accuracy rate in binary class and a 0.66 accuracy rate in multi-class to our dataset. This represents improvements of 0.04 in accuracy in binary and 0.08 accuracy rate in multi-class compared to the baseline performance metrics.

LITERATURE REVIEW

Existing Datasets for Hope Speech Detection

The process of corpus generation for hope speech detection has become a major focus in this field, although these corpora are typically limited in terms of language coverage and sample size. For example, Balouchzahi et al. (2023) recently introduced a dataset for detecting hope speech in English and applied machine learning, deep learning, and transformer-based methods to benchmark the dataset. However, this dataset was limited to a single language, and the study did not address multilingual classification. Similarly, Chakravarthi (2022) introduced a CNN model for hope speech detection in English and Dravidian languages but did not address multi classification. These studies highlight the need for more diverse datasets, including multiple languages, to improve generalization. Furthermore, Chakravarthi (2022) created a joint multilingual dataset for English, Tamil and Malayalam language using YouTube comment to recognize and encourage positivity in the comments but the author did not use multi-classification task in hope speech detection.

Multilingual Hope Speech Detection

Several studies have explored multilingual hope speech detection, employed advanced machine learning models to handle linguistic and cultural differences across

languages. Ghanghor et al. (2021) applied pre-trained transformer models such as m-BERT-cased and XLM-RoBERTa for detecting hope speech in English, Tamil, and Malayalam. Their Results shows that m-BERT-cased perform better than all other models, achieving a highest F1-score of 0.93 for English, 0.83 for Malayalam, and 0.60 for Tamil. While this work contributes to multilingual detection but it does not explore multi classification task across diverse languages. Moreover, Chinnappa (2021) worked on detecting hope speech in Tamil, English, and Malayalam, highlighting the challenges posed by code-mixed data, which further complicates the classification task. Building on this, Malik et al. (2023) extended the scope by exploring a joint multilingual and translation-based approach, focusing on English and Russian languages, highlighting the potential of translation techniques in multilingual hope speech detection. They fine-tuned a pre-trained Russian-RoBERTa model and achieved impressive results, with an accuracy of 94% and an F1-score of 80.24%. This approach demonstrated the potential of leveraging translation for better model performance, but it did not address multiclass classification tasks, which remain an important area for further exploration.

Contribution of the Current Research

Our research presents a novel approach by focusing specifically on joint multilingual hope speech detection across English, Urdu, and Arabic languages using two level text classification. Unlike prior researches that have focused on individual languages, our methodology offers a comprehensive joint-multilingual dataset that comprises both binary and multiclass classification tasks. This study contributes valuable insights into the detection of hope speech across

three languages, offering new avenues for improving sentiment analysis and social media monitoring tools. Table 1 provides a summary of prior studies related to hope speech detection, highlighting the differences between these studies and the proposed study.

METHOD

Corpus Compilation Process

Dataset Collection and Integration

Our dataset consists of approximately 80,000 tweets from various disciplines in English, Urdu, and Arabic that were sourced from Twitter¹, as Twitter is the largest social media platform and microblogging service that enables users to post and interact with messages known as «tweets.» The collection process involved extracting tweets using Twitter’s API (Tweepy). In this study, we amassed a corpus of 80,000 recent and keywords-based tweets sourced from Twitter, employing a systematic approach centered on hope-related keywords, like in Urdu انشاء اللہ (In Sha Allah), خیر انشاء اللہ (Khair In Sha Allah), خواہش ہے (wish), کل (tomorrow), مستقبل (future), کامیابی (success), انتظار (waiting), امید سے بھرپور (hopeful), etc. while in English we used (aspiration, believe, coming soon, dreaming, expectation, feeling positive, I wish, looking forward to and joyful), etc. and while for Arabic we used التفاؤل (optimism), تشجيع (encouragement), ذهاب, موافقة (approval), يتمنى (wish), etc., with different variations. These keywords were used to capture a diverse spectrum of hopeful expressions and sentiments articulated across the

Table 1
Prior Studies Related to Hope Speech Detection vs. Proposed Study

References	Language	Joint Multilingual	Supervised Methods	Multi Classification
Balouchzahi et al. (2023)	English	No	LR, SVM, CNN, LSTM, BiLSTM, Transformer	Yes
Malik et al. (2024)	English, Russian	Yes	SVM, RF, CNN, RoBET base with classifier	No
Kumar et al. (2022)	English, Spanish, Tamil, Malayalam	No	SVM, LR, RF	No
Roy et al. (2022)	English	No		No
Chakravarthi et al. (2021)	English, Tamil, Malayalam, Kannada	No	SVM, DT, LR, KNN, RoBERTa Classifier	No
Ghanghor et al. (2021)	English	No		No
Proposed	English, Urdu, Arabic	Yes	DT, CatBoost, XGB, LR, BiLSTM, CNN, BGRU, BERT, DistilBERT	Yes

¹ Prohibited in Russian Federation.

platform. Data collection spanned from September 2023 to March 2024, offering a robust foundation for conducting in-depth analyses and investigations into the dynamics of hopeful communication within the digital landscape. After collecting the samples from Twitter, we combine our data from English, Arabic, and Urdu into a single CSV file. This combined dataset is called Posi-Vox-2024. «Posi Vox,» derived from «Posi» (positive) and «Vox» (voice), focuses on hope speech and aims to detect positive discourse across multilingual communities. Figure 1 shows the proposed methodology and design of the study, outlining the process of analyzing hope speech in mixed texts commonly found in social media discussions within multilingual communities. The term «multilingual hope speech detection» refers to this unified approach that processes and interprets mixed-language texts to enhance sentiment analysis across diverse online communities. Our proposed model captures linguistic nuances without translation, making it highly relevant for multilingual social media platforms. It offers a scalable solution for detecting hope speech in mixed-language content, providing greater flexibility and robustness compared to traditional monolingual models, thereby enhancing sentiment analysis and fostering positive discourse in diverse online communities.

Data Preprocessing

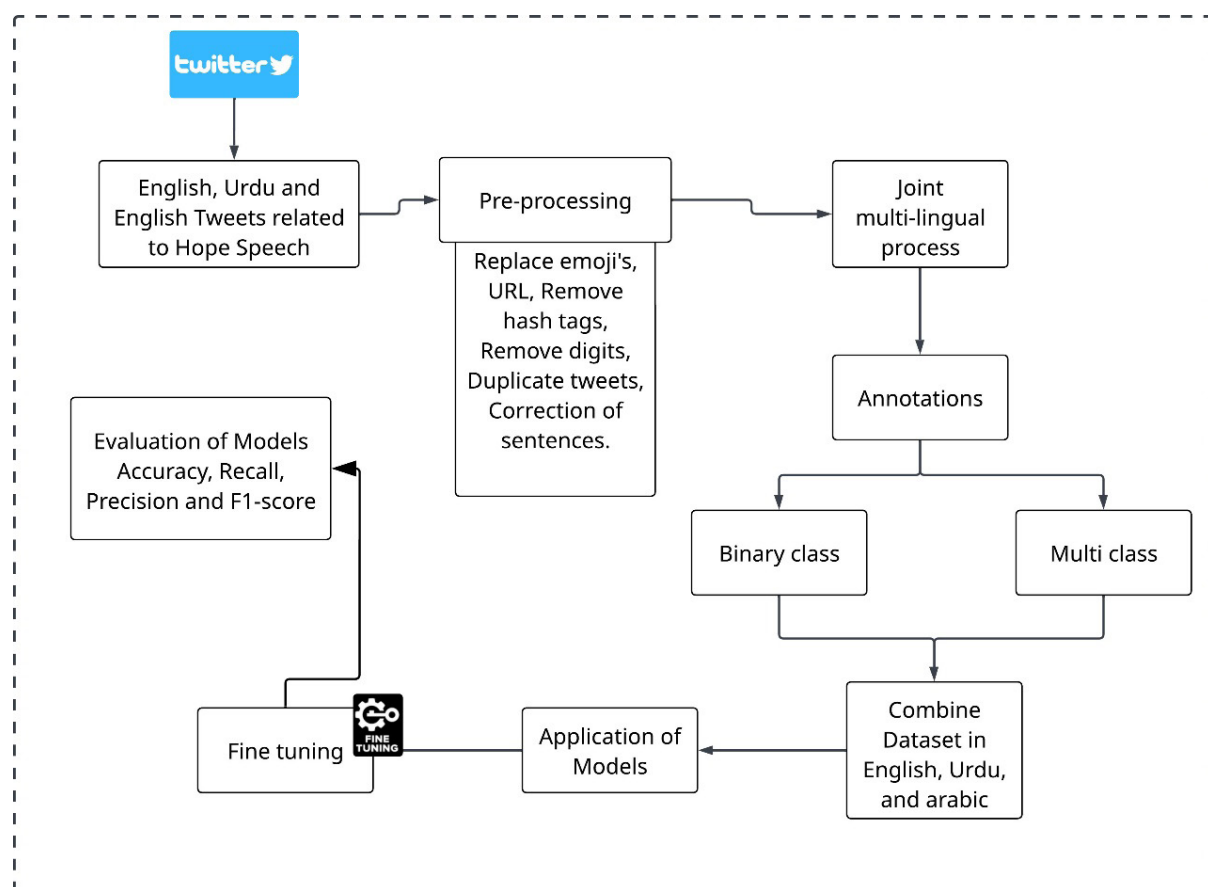
The Tweepy² API was developed for providing functionalities to filter tweets based on different criteria, such as date, location, language, and tweet id. Specifically, we utilized the date and language attributes to scrape tweets in the English, Urdu, and Arabic languages. Due to the extensive noise present in social media's textual content, we conducted various pre-processing procedures:

1. Eliminating URLs, user mentions in the form @use, and HTML Tags.
2. Removal of punctuation marks from the text.
3. Remove duplicates and less than 20-character tweets.
4. Uppercase Text is transformed to lowercase.
5. Replace the emoji with a corresponding text; as we know, emoji's play an essential role in detecting tweets.
6. Removal of the Digits in the tweets.
7. Decodes all the short text such as thnx to Thanks, plz to please, etc.

After processing 80,000 tweets, only 18,362 original tweets remained in Urdu, English, and Arabic to create a joint multilingual dataset.

Figure 1

Proposed Methodology and Design



² <https://www.tweepy.org/> Last visited: 11-10-2024. Prohibited in Russian Federation.

Annotation Process

Annotation Guidelines

Based on the definition of hope provided by psychologists, we categorized the tweets into two classes. The primary class encompassed tweets expressing hope, whereas the secondary class comprised tweets devoid of any sense of hopefulness. This classification methodology enables us to analyze and interpret the presence or absence of hope within tweet content, allowing for deeper insights into user sentiments and emotional expressions on social media platforms. In the next phase of analysis, we categorized tweets into various types of hope by examining the specific features and characteristics present in the content. We implemented specific guidelines for the primary and secondary categorization of tweets, which are detailed along with the examples in Tables 2 and 3.

NHS: The tweet does not convey any sense of hope, aspiration, desire, or anticipation of the future.

1. Generalized Hope: This form of hope is characterized by a general sense of optimism and hopefulness that is not tied to any particular event or outcome.
2. Unrealistic Hope: often manifests as a wish for something to materialize despite its likelihood of being remote or virtually non-existent. Occasionally,

individuals may harbor hope for irrational events or outcomes stemming from emotions such as anger, sadness, or depression.

3. Realistic Hope: This type of hope entails anticipating something that is reasonable, meaningful, and within the realm of possibility. There is a strong likelihood that anticipated events or outcomes will occur.

Annotator Selection

We explicitly avoided selecting annotators for the Posi-Vox-2024 dataset based on racial information, thereby demonstrating our unwavering dedication to promoting a culture of equity and diversity, while upholding the integrity of the dataset. We made a deliberate effort to record the nationality of annotators while avoiding the consideration of racial information. This approach allowed us to monitor the geographical diversity of our annotators in an unbiased manner, as shown in Table 4, without incorporating any biases related to racial characteristic.

Annotation Procedure

Selected annotators were provided with comprehensive annotation guidelines and sample annotations in the ‘Annota-

Table 2
Binary Class Hope Speech

No.	Tweets	Category
1	جب کسی کام کا ارادہ کر لیں تو مکمل اللہ پہ بھروسہ کریں اور ہر شک و شبہ کو دل سے نکال دیں کیونکہ جس کے سپرد آپ نے اپنے معاملات کیے ہیں وہ بہترین کارساز ہے (When you intend to do something, put your full trust in Allah and remove all doubts from your heart because He is the best doer to whom you have entrusted your affairs.)	Hope
2	Nobody cares, you are undesired, and these no standard black men are hyping u up.	Not hope

Table 3
Multi-Class Hope Speech

No.	Tweets	Category
1	Embracing each day with optimism, believing in brighter tomorrows, and trusting in the journey ahead.	Generalized Hope
2	Have faith that your life will improve and that everything will work out for the best. Trust that your health will improve and love will come your way.	Realistic Hope
3	I dream of flying without wings, soaring above the clouds, defying gravity’s hold.	Unrealistic Hope

Table 4
Annotators Based On Geographical Area

Language	Country	Male	Female	Undergraduate	Postgraduate
English	UK	2	0	2	0
Urdu	Pakistan	2	1	2	1
Arabic	UAE	2	0	2	0

Figure 4
Label Distribution of Binary and Multi class in the entire Dataset

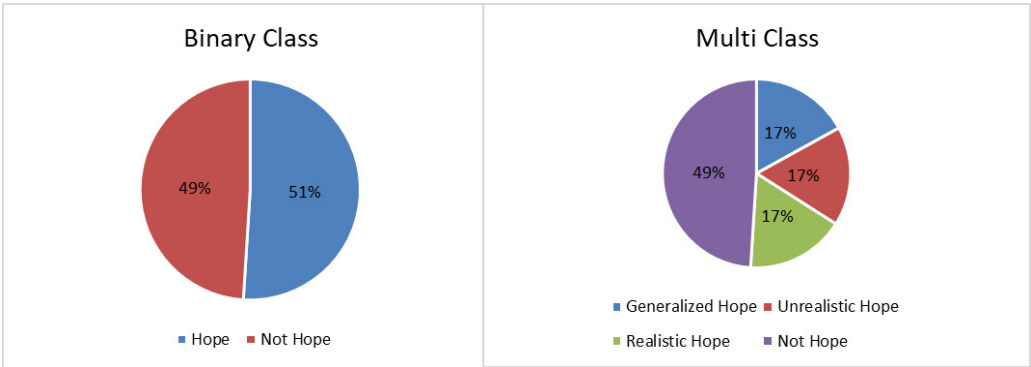


Table 5
Statistics of the Dataset

Class	Tweets	Words	Avg. Words	Characters Avg.	Characters	Vocabulary
Generalized hope	3082	82046	26.61	139.39	429603	19617
Realistic hope	3156	90355	28.63	150.12	473791	21286
Unrealistic hope	3074	83248	27.08	137.90	423933	20693
Not hope	9050	243837	26.94	138.83	1256442	44181
Total	18362	499486	27.32	141.56	2583769	105777

Ethical Concern

Data collected from Twitter is highly sensitive, and we highlight the privacy measures implemented in the data-annotation processes. The identities of the involved individuals remained hidden, and our annotator identified a name associated with a politician or celebrity. They adhered to a strict protocol of non-engagement, refraining from attempting to establish contact with such individuals.

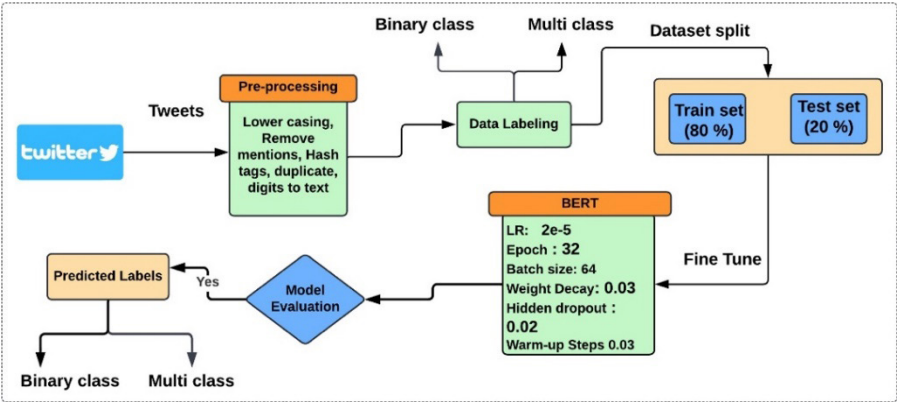
Methods for Hope Speech Detection

To demonstrate how our proposed Posi-Vox-2024 corpus can be used to develop, evaluate and compare methods

for hope speech detection task, we applied and compared state-of-the-art four machine learning models: (i) Decision Tree (DT), (ii) CatBoost (CB), (iii) Extreme Gradient boosting (XGB), and (iv) Logistic Regression (LR); three Deep learning models: (i) Bidirectional Long Short-Term Memory (BiLSTM), (ii) Convolutional Neural Network (CNN), and (iii) Bidirectional Gated Recurrent Unit (BGRU); and two transfer learning models: (i) pre-train BERT, and (ii) distilBERT.

Our best performing model is based on the BERT architecture, leveraging transformer layers to capture contextual relationships in multilingual text. After preprocessing with appropriate tokenizers for English, Urdu, and Arabic, we

Figure 5
BERT-Based Model Training Pipeline for Multilingual Text Classification



fine-tuned a pre-trained BERT model on our annotated dataset using cross-entropy loss and an Adam optimizer with a learning rate of 2e-5. The dataset was partitioned into 80% for training and 20% for testing. To ensure reproducibility, configurations, including batch size, number of epochs, and evaluation metrics, along with optimum hyper-parameter values, are presented in Table 9. A diagram illustrating the BERT architecture and data flow is provided in Figure 5 to clarify how the model processes multilingual input and predicts hope speech.

RESULTS

In this section, we present the results of various machine learning, deep learning, and transformer-based models applied to the task of multilingual hope speech detection. These models were evaluated on both binary and multi-class classification tasks using our proposed Posi-Vox-2024 corpus, which includes English, Urdu, and Arabic text. Tables 6, 7, 8, 10, and 11 present the Precision, Recall, F-1 score, and Accuracy results obtained by applying state-of-the-art machine learning algorithms such as Decision Tree (DT), Categorical Boosting (CatBoost), Extreme Gradient Boosting (XGB), and Logistic Regression (LR). For deep learning, we utilized Convolutional Neural Network (CNN), Bidirectional Gated Recurrent Unit (BGRU), and Bidirectional Long Short-

Term Memory (BiLSTM) models. In the transformer category, we employed Bidirectional Encoder Representations from Transformers (BERT) and Distilled BERT (DistilBERT) on our proposed Posi-Vox-2024 corpus. Our experiments focused on identifying the most suitable model for handling hope speech across languages, systematically tuning hyperparameters for each model and analyzing their performance based on metrics such as accuracy, precision, recall, and F1-score. The following subsections provide detailed results for each category of models.

Machine Learning

Table 6 shows the results attained by the various machine learning models using TF-IDF word embedding for hope speech detection, classified into binary and multi-class tasks. For binary class of hope speech detection, the DT, CatBoost, XGB, and LR models show F1-scores ranging from 0.70 to 0.73, with LR achieving the highest precision, recall, F1-score, and accuracy of 0.75. In the multi-classification hope speech detection task, again LR performed better than all other models, achieving F1-score of 0.61. CatBoost and XGB shows competitive performance with Accuracy rates of 0.58, hence LR outperforms the other models in both binary and multi-class tasks, achieving the highest Precision, recall, F1-score, and accuracy.

Table 6
Results of Machine Learning Models

Class	Model	Precision	Recall	F1-score	Accuracy
Binary class	DT	0.72	0.72	0.72	0.72
	Catboost	0.73	0.73	0.73	0.73
	XGB	0.72	0.71	0.7	0.71
	LR	0.75	0.75	0.75	0.75
Multi class	DT	0.59	0.59	0.59	0.59
	Catboost	0.57	0.58	0.53	0.58
	XGB	0.57	0.58	0.54	0.58
	LR	0.61	0.61	0.61	0.61

Deep Learning

Table 7 presents the performance metrics for three deep learning models such as CNN, BGRU, and BiLSTM on binary and multi-class classification tasks. For the binary classification task, all three models perform similarly, with the CNN and BiLSTM models achieving a precision, recall, and F1-score of 0.75, while the BGRU model has slightly lower values (0.74).

The accuracy for all three models is also consistent, at 0.75 for CNN and BiLSTM, and 0.74 for BGRU. In the multi-class classification task, the models show a decrease in performance across all metrics. The CNN and BGRU models have precision, recall, F1-score, and accuracy around 0.56, while BiLSTM performs slightly better with 0.62 for all metrics. This suggests that while the models perform well on binary classification, they face more challenges with multi-class classification.

Table 7
Results of Deep Learning Models

Class	Model	Precision	Recall	F1-score	Accuracy
Binary class	CNN	0.75	0.75	0.74	0.75
	BGRU	0.74	0.74	0.74	0.74
	BiLSTM	0.75	0.75	0.75	0.75
Multi class	CNN	0.56	0.56	0.56	0.56
	BGRU	0.55	0.55	0.55	0.55
	BiLSTM	0.62	0.62	0.62	0.62

Transformer Results

The table 8 summarizes the performance metrics Precision, Recall, F1-score, and Accuracy—for binary and multi-class classification tasks. In binary classification, BERT achieves higher precision, recall, F1-score, and accuracy (all at 0.78)

compared to DistilBERT, which has slightly lower scores (F1-score of 0.75 and accuracy of 0.76). In multi-class classification, both models show a performance drop; BERT attains an F1-score of 0.65 and accuracy of 0.66, while DistilBERT has a slightly lower F1-score and accuracy of 0.64. Overall, BERT outperforms DistilBERT across both tasks.

Table 8
Transformer Results

Class	Model	Precision	Recall	F1-score	Accuracy
Binary class	Bert	0.78	0.78	0.78	0.78
	DistilBert	0.76	0.76	0.75	0.76
Multi class	Bert	0.66	0.66	0.65	0.66
	DistilBert	0.64	0.64	0.64	0.64

Table 9 presents the optimal fine-tuning parameters for a pre-trained BERT model for both binary and multi-class classification tasks. The best hyper-parameters were recognized through grid search, considering the following ranges: learning rates are 1e-5, 1e-2, 2e-5, 3e-5, 3e-4, epochs are 9,

32, 64, batch sizes are 64, 128, 512, weight decay values from 0.01 to 0.1, hidden dropout rates of 0.02 and 0.1, and warm-up steps from 0.03 to 0.1. These settings ensure balanced training efficiency and robust model performance across various classification problems.

Table 9
Optimum Values Identified for the Hyper-Parameters of the Bert Model

Hyper-parameter	Grid search
Learning rate	1e-5,1e-2, 2e-5, 3e-5, 3e-4
Epoch	3, 9, 32
Batch size	32, 64, 128
Weight Decay	0.01-0.1
Hidden dropout	0.02, 0.1
Warm-up Steps	0.03-0.1

Error Analysis

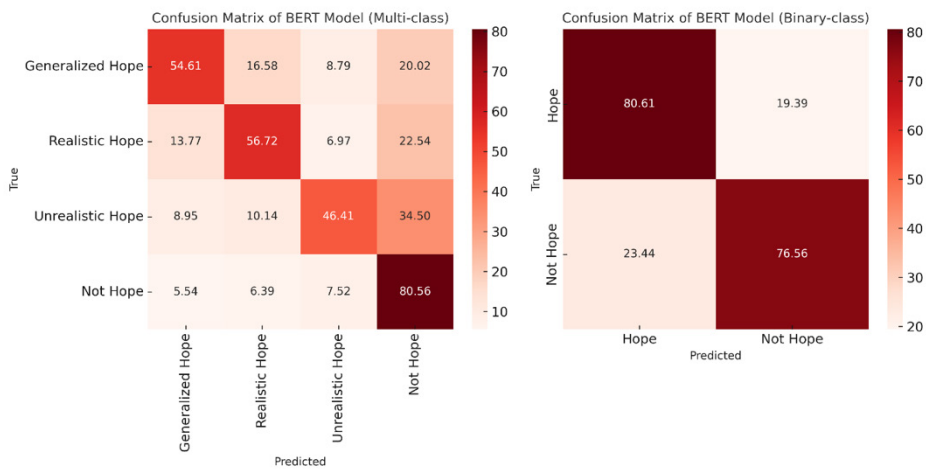
Table 10 shows class-wise scores, while Figure 6 shows the confusion matrix for both binary and multiclass classification in percentage achieved by our proposed model. Notably, our model demonstrated better performance in the not hope class in terms of precision. In classifying Unrealistic

hope, our proposed model performs better than all other labels of hope class while showing comparatively lower accuracy in distinguishing between generalized and realistic hope categories.

Table 10
Class Wise Score for the Proposed Methodology

Class	Categories	Precision	Recall	F1-Score	Support	Accuracy
Binary class	Hope	0.78	0.79	0.78	3666	0.78
	Not hope	0.79	0.78	0.78	3631	
Multi class	Generalized hope	0.50	0.63	0.55	1194	0.66
	Realistic hope	0.57	0.53	0.55	1220	
	Unrealistic hope	0.63	0.41	0.50	1252	
	Not hope	0.75	0.80	0.77	3631	

Figure 6
Confusion Matrix of Proposed Methodology



The pre-trained BERT model demonstrated a notable performance gain over traditional machine learning models, with a binary classification accuracy of 0.78 compared to traditional machine learning, such as LR 0.75, resulting in a performance improvement of approximately 4%. For multi-class tasks, BERT achieved 0.66, outperforming the LR 0.61, indicating a performance improvement of about 8.20%. This

suggests BERT’s superior contextual understanding and handling of nuanced language, especially in a multilingual setting. Thus, BERT’s advanced language modeling provides significant advantages in detecting hope speech across different languages. Table 11 shows the outcomes of the top performing models of each learning approach in binary and multi class.

Table 11
Top Performing Models in Each Learning Approach

Class	Model	Learning approach	Accuracy
Binary class	LR	Machine learning	0.75
	BiLSTM	Deep learning	0.75
	BERT	Transformer	0.78
Multi class	LR	Machine learning	0.61
	BiLSTM	Deep learning	0.62
	BERT	Transformer	0.66

DISCUSSION

This study explores a valuable set of features that effectively detect hope speech expressions in Twitter tweets. This research offers meaningful insights for online users and society, aiming to foster peace and positivity. Hope is often linked to providing encouragement, support, reassurance, suggestions, or inspiration to individuals during times of illness, stress, loneliness, or depression (Snyder et al., 2002). The literature review has primarily focused only binary class approaches to identifying hope speech detection on different social media platforms, with limited work on developing a multilingual framework. To address this research gap, we built a multilingual tool that combines joint multilingual methodology to tackle the task of hope speech detection in English, Urdu and Arabic languages. Our proposed tool was trained and tested on a multilingual dataset to uncover practical insights and ensure its applicability on real time. Our findings show that our proposed method is very effective and power tool to identify the hope speech detection in Twitter post. We utilized the power of transfer learning methods by fine-tuning the pre-trained BERT model and added a fresh contribution that attained 78% accuracy in binary class and 66% in multi classification. Furthermore, our proposed methodology outperformed the four baselines such as LR, XGB, CB and DT. Thus, based on these results, our proposed framework can be employed for other Multi-lingual Text Classification problems in similar fields.

There are several limitations in this study. Firstly, collecting and annotating hope speech data in English, Urdu, and Arabic pose several challenges. One of the main difficulties is identifying native speakers fluent in these languages who also possess knowledge in NLP and machine learning for accurate and reliable annotations. Secondly, during annotation process, we encountered numerous tweets that expressed hope but had a negative undertone. For instance, in Urdu, tweets such as "میری امید ہے کہ میرے دشمنوں کو تباہی کا سامنا کرنا" (My hope is that my enemies should suffer destruction, their destruction will be my joy. #USER») present a challenge. Although this tweet conveys hope, its primary sentiment is negative, further complicating the annotation process. Thirdly, Urdu and Arabic are considered low-resource languages in the context of machine learning and deep learning, making it more challenging to understand and process them, which in turn hinders the development of robust models for detecting hope speech. Fourthly, despite its notable performance in multilingual hope speech detection, the proposed work has limitations. The Posi-Vox-2024 dataset is limited in size and diversity, thereby affecting its generalizability. Language-specific nuances and code-switching have not been fully addressed, potentially impacting the classification accuracy. The model's complexity and resource require-

ments limit accessibility, and its performance may degrade over time owing to the dynamic nature of the social media discourse.

CONCLUSION

Social media has become a powerful space for public dialogue, influencing opinions and the emotional landscape of communities. Until now, most research has focused on addressing negativity in the English language, particularly hate speech detection. This study highlights the critical need for multilingual hope speech detection (MHSD) in social media discourse, particularly focusing on the Urdu and Arabic languages, which has been overlooked in existing research. To achieve this objective, we address the two level text classification and built a comprehensive dataset named as Posi-Vox-2024 based on three languages such as English, Urdu and Arabic to tackle the challenges of multilingualism and improve communication across different backgrounds. By creating a multilingual dataset and employing state-of-the-art transfer learning models with fine-tuning, we effectively addressed the challenges associated with identifying hope speech in English, Arabic and Urdu. The results indicate that our proposed framework, utilizing pre-trained BERT model, significantly outperformed four baseline models (DT, XGB, Catboost, and LR), achieving accuracies of 0.78 in binary class and 0.66 in multi class. These findings underscore the importance of promoting positive discourse online and demonstrate the potential of hope speech as a means to foster healthier and more constructive interactions within communities. Further exploration could focus on expanding the dataset and incorporating additional languages to enhance the generalizability and robustness of the proposed framework.

ACKNOWLEDGMENTS

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONAHCYT, Mexico, grants 20241816, 20241819, 20240936 and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONAHCYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

DATA AVAILABILITY

Data will be made available on request.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHOR CONTRIBUTIONS

Muhammad Ahmad: Conceptualization; Data curation; Methodology; Resources; Software; Visualization; Writing – original draft; Writing – review & editing.

Usman Sardar: Data curation; Formal analysis; Methodology.

Humaira Farid: Investigation; Visualization; Writing – review & editing.

Iqra Ameer: Data curation; Formal analysis; Methodology; Software; Writing – original draft; Writing – review & editing.

Muhammad Muzamil: Data curation; Methodology; Software.

Ameer Hmaza: Data curation; Methodology.

Grigori Sidorov: Conceptualization; Resources; Supervision; Validation.

Ildar Batyrshin: Conceptualization; Project administration; Resources; Supervision; Validation; Writing – original draft; Writing – review & editing.

REFERENCES

- Alawadh, H. M., Alabrah, A., Meraj, T., & Rauf, H. T. (2023). English language learning via YouTube: An NLP-based analysis of users' comments. *Computers*, 12(2), 24. <https://doi.org/10.3390/computers12020024>
- Anand, M., Sahay, K. B., Ahmed, M. A., Sultan, D., Chandan, R. R., & Singh, B. (2023). Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. *Theoretical Computer Science*, 943, 203-218. <https://doi.org/10.1016/j.tcs.2022.06.020>
- Anjum, & Katarya, R. (2024). Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. *International Journal of Information Security*, 23(1), 577-608. <https://doi.org/10.1007/s10207-023-00755-2>
- Arif, M., Shahiki Tash, M., Jamshidi, A., Ullah, F., Ameer, I., Kalita, J., ... & Balouchzahi, F. (2024). Analyzing hope speech from psycholinguistic and emotional perspectives. *Scientific Reports*, 14(1), 23548. <https://doi.org/10.1038/s41598-024-74630-y>
- Austin, D., Sanzgiri, A., Sankaran, K., Woodard, R., Lissack, A., & Seljan, S. (2020). Classifying sensitive content in online advertisements with deep learning. *International Journal of Data Science and Analytics*, 10(3), 265-276. <https://doi.org/10.1007/s41060-020-00212-6>
- Balouchzahi, F., Sidorov, G., & Gelbukh, A. (2023). Polyhope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, 225, 120078. <https://doi.org/10.1016/j.eswa.2023.120078>
- Chakravarthi, B. R. (2022). Hope speech detection in YouTube comments. *Social Network Analysis and Mining*, 12(1), 75. <https://doi.org/10.1007/s13278-022-00901-z>
- Chakravarthi, B. R. (2022). Multilingual hope speech detection in English and Dravidian languages. *International Journal of Data Science and Analytics*, 14(4), 389-406. <https://doi.org/10.1007/s41060-022-00341-0>
- Chinnappa, D. (2021). Dhivya-hope-detection@ LT-EDI-EACL2021: Multilingual hope speech detection for code-mixed and transliterated texts. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 73-78). Association for Computational Linguistics. <https://aclanthology.org/2021.ltedi-1.11>
- Davidson, T., Bhattacharya, D., & Weber, I. (2019). *Racial bias in hate speech and abusive language detection datasets*. arXiv preprint arXiv:1905.12516.
- Gowen, K., Deschaine, M., Gruttadara, D., & Markey, D. (2012). Young adults with mental health conditions and social networking websites: seeking tools to build community. *Psychiatric Rehabilitation Journal*, 35(3), 245. <https://doi.org/10.2975/35.3.2012.245.250>
- Ghanghor, N., Ponnusamy, R., Kumaresan, P. K., Priyadharshini, R., Thavareesan, S., & Chakravarthi, B. R. (2021). IIITK@ LT-EDI-EACL2021: Hope speech detection for equality, diversity, and inclusion in Tamil, Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 197-203). Association for Computational Linguistics.
- Irfan, A., Azeem, D., Narejo, S., & Kumar, N. (2024). Multi-Modal Hate Speech Recognition Through Machine Learning. In *2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC)* (pp. 1-6). IEEE. <https://doi.org/10.1109/KHI-HTC60760.2024.10482031>
- Kogilavani, S. V., Malliga, S., Jaiabinaya, K. R., Malini, M., & Kokila, M. M. (2023). Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*, 81, 630-633. <https://doi.org/10.1016/j.matpr.2021.04.102>

- Kumar, A. Saumya, S., & Roy, P. (2022). SOA_NLP@ LT-EDI-ACL2022: An ensemble model for hope speech detection from YouTube comments. In *Proceedings of the second workshop on language technology for equality, diversity and inclusion* (pp. 223-228). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.ltedi-1.31>
- Lee, Y., Yoon, S., & Jung, K. (2018). *Comparative studies of detecting abusive language on twitter*. arXiv preprint arXiv:1808.10245.
- Louati, A., Louati, H., Albanyan, A., Lahyani, R., Kariri, E., & Alabduljabbar, A. (2024). Harnessing machine learning to unveil emotional responses to hateful content on social media. *Computers*, 13(5), 114. <https://doi.org/10.3390/computers13050114>
- Malik, M. S. I., Nazarova, A., Jamjoom, M. M., & Ignatov, D. I. (2023). Multilingual hope speech detection: A Robust framework using transfer learning of fine-tuning RoBERTa model. *Journal of King Saud University-Computer and Information Sciences*, 35(8), 101736. <https://doi.org/10.1016/j.jksuci.2023.101736>
- Mnassri, Kh., Farahbakhsh, R., Chalehchaleh, R., Rajapaksha, P., Jafari, A.R., Li, G., & Crespi, N. (2024). A survey on multi-lingual offensive language detection. *PeerJ. Computer Science*, 10, e1934–e1934. <https://doi.org/10.7717/peerj-cs.1934>
- Nagar, S., Barbhuiya, F. A., & Dey, K. (2023). Towards more robust hate speech detection: Using social context and user data. *Social Network Analysis and Mining*, 13(1), 47. <https://doi.org/10.1007/s13278-023-01051-6>
- Nath, T., Singh, V. K., & Gupta, V. (2023). *BongHope: An annotated corpus for Bengali hope speech detection*. Research Square. <https://doi.org/10.21203/rs.3.rs-2819284/v1>
- Palakodety, S., KhudaBukhsh, A. R., & Carbonell, J. G. (2020). Hope speech detection: A computational analysis of the voice of peace. In *ECAI 2020* (pp. 1881-1889). IOS Press.
- RamakrishnaIyer LekshmiAmmal, H., Ravikiran, M., Nisha, G., Balamuralidhar, N., Madhusoodanan, A., Kumar Madasamy, A., & Chakravarthi, B. R. (2023). Overlapping word removal is all you need: Revisiting data imbalance in hope speech detection. *Journal of Experimental & Theoretical Artificial Intelligence*, 36(8), 1837–1859. <https://doi.org/10.1080/0952813X.2023.2166130>
- Roy, P., Bhawal, S., Kumar, A., & Chakravarthi, B. R. (2022, May). IIITSurat@ LT-EDI-ACL2022: Hope speech detection using machine learning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 120-126). Association for Computational Linguistics. <https://aclanthology.org/2022.ltedi-1.13>
- Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1-10). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W17-1101>
- Snyder, C. R., Rand, K. L., & Sigmon, D. R. (2002). Hope Theory: A Member of the Positive Psychology Family. In C. R. Snyder, & S. J. Lopez (Eds.), *Handbook of Positive Psychology* (pp. 257-276). Oxford University Press.
- Subramanian, M., Sathiskumar, V. E., Deepalakshmi, G., Cho, J., & Manikandan, G. (2023). A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, 80, 110-121. <https://doi.org/10.1016/j.aej.2023.08.038>
- Wang, Z., & Jurgens, D. (2018). It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 33-45). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1004>
- Yates, A., Cohan, A., & Goharian, N. (2017). *Depression and self-harm risk assessment in online forums*. arXiv preprint arXiv:1709.01848.
- Yenala, H., Jhanwar, A., Chinnakotla, M. K., & Goyal, J. (2018). Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics*, 6, 273-286. <https://doi.org/10.1007/s41060-017-0088-4>
- Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). *Predicting the type and target of offensive posts in social media*. arXiv preprint arXiv:1902.09666.

<https://doi.org/10.17323/jle.2024.22439>

Synchronic and Diachronic Predictors of Socialness Ratings of Words

Vladimir Bochkarev , Anna Shevlyakova , Andrey Achkeev 

Kazan Federal University, Kazan, Russia

ABSTRACT

Introduction: In recent works, a new psycholinguistic concept has been introduced and studied that is socialness of a word. A socialness rating reflects word social significance and dictionaries with socialness ratings have been compiled using either a survey or machine method. Unfortunately, the size of the dictionaries with word socialness ratings created by a survey method is relatively small.

Purpose: The study objective is to compile a large dictionary with English word socialness ratings by using machine extrapolation, transfer the rating estimations to other languages as well as to obtain diachronic models of socialness ratings.

Method: The socialness ratings of words are estimated using multilayer direct propagation neural networks. To obtain synchronic estimates, pre-trained fasttext vectors were fed to the input. To obtain diachronic estimates, word co-occurrence statistics in a large diachronic corpus was used.

Results: The obtained Spearman's correlation coefficient between human socialness ratings and machine ones is 0.869. The trained models allowed obtaining socialness ratings for 2 million English words, as well as a wide range of words in 43 other languages. An unexpected result is that the linear model provides highly accurate estimate of the socialness ratings, which can be hardly further improved. Apparently, this is due to the fact that in the space of vectors representing words there is a selected direction responsible for meanings associated with socialness driven by of social factors influencing word representation and use. The article also presents a diachronic neural network predictor of concreteness ratings using word co-occurrence vectors as input data. It is shown that using a one-year data from a large diachronic corpus Google Books Ngram one can obtain accuracy comparable to the accuracy of synchronic estimates.

Conclusion: The created large machine dictionary of socialness ratings can be used in psycholinguistic and cultural studies. Changes in socialness ratings can serve as a marker of word meaning change and be used in lexical semantic change detection.

KEYWORDS

Socialness, Psycholinguistics, Psycholinguistic data bases, Pre-trained word vectors, Neural networks, Lexical semantic change

Citation: Bochkarev, V., Shevlyakova, A., & Achkeev, A. (2024). Synchronic and diachronic predictors of socialness ratings of words. *Journal of Language and Education*, 10(4), 44-55. <https://doi.org/10.17323/jle.2024.22439>

Correspondence:
Vladimir Bochkarev,
vbochkarev@mail.ru

Received: September 6, 2024

Accepted: December 16, 2024

Published: December 30, 2024

INTRODUCTION

Semantic knowledge is represented in different ways including natural language and its means. Language is directly connected with human perception of reality and cultural context. Therefore, various psycholinguistic parameters of words have been introduced that serve as key features in concept representation and have been widely studied in modern science. Among the mentioned parameters are word concreteness, imageability, va-

lence etc. Social significance is also one of the key features in concept representation as socialness has a great impact on the concept structure and cognition. Socialness means the extent to which each word has social relevance by describing or referring to some socially relevant concept such as a social role, a social space, ideology etc. (Pexman et al., 2022). Recently, dictionaries of word psycholinguistic parameters, including word socialness, have been compiled, which can be employed for solving various practical tasks.



Survey and machine-based methods are used to study psycholinguistic word parameters. The size of the dictionaries compiled using the survey method is relatively small as this method is time- and labour-consuming. Creation of large text corpora and development of methods of natural language processing allowed creation of large dictionaries with word psycholinguistic ratings by machine extrapolation. In this case, the computational model is trained on a small number of words for which human ratings exist, and then the trained model is used to obtain machine ratings for a wide range of words. This approach has made it possible to obtain large machine dictionaries with concreteness ratings, affective ratings (Mohammad et al., 2013; Koper & Schulte im Walde, 2016), etc.

Some recent works have been devoted to estimation of word socialness ratings using the survey method. One of the works presented the first English dictionary of socialness ratings of 535 words was (Binder et al., 2016). Then, the work (Diveica et al., 2023) presented another dictionary which was compiled using the survey method and included 8838 words. It should be noted that the instructions used in (Diveica et al., 2023) were much more detailed than those used in (Binder et al., 2016). Similar study was conducted for the Chinese language (Wang et al., 2023). It was performed in several stages. First, a dictionary of socialness ratings for 17,940 Chinese words; was compiled using the survey method. Then, the ratings were extrapolated to 900 thousand Chinese words and the model was trained to obtain machine ratings for them. Finally, using the trained model and machine translation, a machine-based English dictionary was obtained by transferring Chinese ratings to English ones. A certain drawback of the dictionary presented in (Diveica et al., 2023) is its relatively small size, which may limit its use in practical tasks. This makes it relevant to create a computer model that would allow extrapolating socialness ratings to the largest possible range of words.

The purpose of this paper is to create a large English dictionary with word socialness ratings by using machine extrapolation, transfer the rating estimations to 43 other languages as well as to obtain diachronic models of socialness ratings. We use a model that allows predicting socialness ratings for 2 million English words. Unlike Wang et al. (2023), who transferred ratings from Chinese, we train our model on human judgments collected through surveys of native English speakers. This approach enhances the accuracy of socialness ratings for English words.

METHOD

Source Data

As a source of human ratings for training predictors, the dictionary described in (Diveica et al., 2023) is used. It con-

tains ratings for 8,388 English words. The ratings given in the dictionary range from 1 to 7. High rating values indicate that the word has great social relevance, while low values indicate that the word, on the contrary, is not socially significant. The distribution mode of the of rating values lies in the middle of the range. Thus, the socialness rating scale in the dictionary is essentially bipolar. For convenience, we transformed the rating scale to the range from -1 to +1.

Used Sets of Vectors

To estimate psycholinguistic parameters of words, word vector representations developed within the framework of distributional semantics are employed. The general idea of distributional semantics is that distributional similarity and meaning similarity correlate with each other (Harris, 1970; Rubenstein & Goodenough, 1965; Firth, 1957). Therefore, word meaning can be revealed and estimated by the analysis of its distribution. There are different algorithms of distributional meaning acquisition. In early works, mainly representations based on co-occurrence vectors were used (Weeds et al., 2004; Pantel, 2005; Bullinaria & Levy, 2007; Gulordava & Baroni, 2011). In (Bullinaria & Levy, 2012), it was proposed to employ vectors constructed from Point Mutual Information (PMI). One of the reasons that hindered the effective application of early word embeddings was the high dimensionality of the resulting vectors. Various options for reducing the dimensionality of vector representations were considered, for example, using SVD (Turney & Pantel, 2010; Bullinaria & Levy, 2012). In 2013, an improved word embeddings technique using neural network approaches was proposed in (Mikolov et al., 2013; Bojanowski et al., 2017) that opened new horizons in this field of research. Recent advances in this area involve the use of contextualized word embeddings (Peters et al., 2018; Devlin et al., 2019). There is an overview of modern usage of low-dimensional word embeddings presented in (Worth, 2023; Pilehvar & Camacho-Collados, 2020). Currently, methods based on vector neural network models are applied in most cases. However, simpler representations based on explicit word vectors are also employed because their use has some advantages: the obtained results are easily interpreted (Basile & McGillivray, 2018), as well as the diachronic models can be easily constructed (Bochkarev et al., 2022). In this paper, we test both types of word embeddings in relation to the problem of predicting word socialness ratings.

Firstly, we selected two sets of pre-trained vectors trained on the largest corpora. One of them is the fasttext pre-trained vectors trained on the CommonCrawl corpus (Grave et al., 2018) with a total size of 650 billion words. In accordance with the recommendations by (Charbonnier & Wartena, 2019) and the results of our experiments, we used vectors trained without using subword information. Besides, we employed the Glove-840B pre-trained vectors also trained on the CommonCrawl corpus that included 840 billion words at the time of creating the vector set (Pennington, 2014).

Secondly, to obtain dictionaries for various languages (besides English), we used two multilingual sets of vectors. The fasttext project page provides embeddings for 44 languages, trained on Wikipedia texts as of 2017, aligned in a single vector space. The algorithm presented in (Juolin, 2018) was employed to align the vectors. The MuSE project page provides aligned embeddings for 29 languages. To create this multilingual dataset, as in the previous case, the fasttext vectors trained on Wikipedia texts were chosen as the initial ones, however, a different algorithm was used for the alignment (Conneau et al, 2017). All the above vector sets belong to the class of context-free models. As mentioned above, contextualized word embeddings are more promising. However, it should be noted that in the existing dictionary presented in the work (Diveica et al., 2023), only one value of the socialness rating is given for each word form. Moreover, it is not indicated for polysemantic words to which of its meanings the rating refers. In this case, contextualized embeddings may not show advantages over context-free models. It is also worth mentioning that all the above vector sets were obtained by training on synchronic corpora, and thus cannot be used to obtain diachronic estimates of socialness ratings.

Therefore, besides the low-dimensional vector representations mentioned above, we also employed explicit word vectors built according to the CFW (co-occurrence with the most frequent words) method. A detailed description of the method is proposed in (Xu & Kemp, 2015; Khristoforov et al., 2020). According to the CFW method, the vectors were composed of the values of regularized pointwise mutual information (in the form proposed in Bochkarev et al., 2021) for bigrams of the form Wx and xW , where W is the target word and x is one of the most frequent words. The frequency data on words and phrases required for constructing the vectors were extracted from the large diachronic corpus Google Books Ngram (Lin et al., 2012). To train the neural network, we use the frequency data averaged over the period 1900-2019. In this paper, following (Khristoforov et al., 2020), we use a list of 20 thousand most frequent words. Thus, a word is described by a vector of dimension 40,000.

Neural Network Predictors Training

The socialness degree of words was estimated using multi-layer direct propagation neural networks. To maximize prediction accuracy, a number of network architectures with different numbers of layers and neurons per layer were tested. Each network was trained using several algorithms (adadelat, adagrad, adam, SGD) with different learning rate parameters. Tests were also conducted using L1 and L2 regularization and dropout regularization. Based on the tests, the following architecture of neural network predictors and learning parameters was selected for the case of low-dimensional vector representations:

- (1) 3 dense layers of 3072 neurons with the ReLU activation function, the output layer of dimension 1 with linear activation;
- (2) L2 regularization with coefficient $5 \cdot 10^{-4}$;
- (3) The MSE metric for early stopping (no improvement greater than $1 \cdot 10^{-6}$ during 100 epochs)

Similarly, the following parameters were chosen for predictors that use explicit word vectors:

- (1) 6 dense layers of 512 neurons with the ReLU activation function, the output layer of dimension 1 with linear activation;
- (2) dropout-regularization – between dense layers with coefficient 0.02;
- (3) The MSE metric for early stopping (no improvement greater than $1 \cdot 10^{-6}$ during 5 epochs).

In both cases, the best results were obtained using the MSE loss function and the SGD optimization algorithm.

Cross-Validation Procedure

Cross-validation has been used to improve reliability of the results and control the accuracy of the obtained estimates. Following (Bochkarev et al., 2024a), the list of words was divided into 6 groups including non-overlapping words. In each case, four groups out of six were used to train the model, and the remaining two groups were used as a test set. There are 15 different ways to select 4 groups out of 6, so for each word we get 15 independently trained models. In this case, for any word there are 5 models for which this word was in the test set, not the training set. Having several models allows us to further improve the accuracy by averaging the estimates, as well as to determine the standard deviation of the obtained estimate.

Training Linear Predictors

In addition to neural networks, we will also present the results for linear models for comparison. As for explicit word vectors, the relationship between individual vector components and meaningful characteristics of the word is obviously non-linear. Therefore, it makes sense to use linear predictors only for cases where low-dimensional vectors are fed to the input. Training a linear predictor is a linear regression task and is carried out using pseudo-inversion according to the L2 norm. Just as for neural network predictors, in this case a set of models is independently trained on 15 subsets of the sample. The estimates obtained from independently trained models can then be averaged in one way or another.

Transferring Ratings to Other Languages

The existence of freely available multilingual vector sets aligned in a single vector space makes it easy to transfer

ratings from one language to another. There is a dictionary with human estimates of the socialness ratings for the English language. We train a socialness predictor using vectors for English words from a multilingual vector set as input data. By feeding word vectors for another language from the same set, we obtain a dictionary with socialness ratings for this language. It should be taken into account that errors in rating estimates associated with the imperfection of the model will be summed up with errors in vector alignment in two languages. Therefore, to solve the problem of transferring ratings in this paper, we use linear predictors, since in this case it is easier to predict the error value of the output value if errors in the input data are known.

RESULTS

We trained models of socialness ratings of English words using four sets of pre-trained vectors and one set of explicit word vectors. Neural network predictors were trained for each of these five sets of vectors. Also, linear predictors were trained for the four low-dimensional sets of vectors.

The estimates were obtained for the 5 models for which this word was in the test sample, and therefore was not presented to the neural network at the training stage. These estimates were averaged for each word. The Pearson's and Spearman's correlation coefficients between the averaged estimates obtained in this way and human ratings for different word representations and predictor architectures are given in Table 1. Averaging over a set of independently trained models allows one to increase the estimation accuracy. For example, for a set of pre-trained fasttext-CommonCrawl vectors, the average value of the Pearson's and Spearman's correlation coefficients between human ratings and their machine estimates for 15 models was 0.8531 and 0.8566, respectively. Averaging over independently trained models allowed us to

increase the values of the correlation coefficients to the values of 0.8655 and 0.8688, respectively (Table 1).

First of all, it should be noted that the accuracy of linear predictors is very slightly inferior to the accuracy of neural network predictors using the same set of pre-trained vectors. At the same time, if linear predictors have a number of adjustable parameters equal to the dimension of the input vectors (in our case - 300 parameters), then neural network predictors using the same input data have 13.9 million weight coefficients (see Section 2). The model using explicit word vectors has even 21.8 million weight coefficients. Thus, this insignificant increase in the accuracy of neural network predictors is achieved by a colossal complication of the model, and a corresponding increase in training time.

It should also be noted that 15 linear models independently trained on different word subsets are highly consistent with each other. For a trained linear predictor, the gradient of the model output is constant throughout the vector space. Thus, the i -th model can be characterized by a unit direction vector v_i , the gradient normalized to unit length. For example, for the set of pre-trained fasttext-CommonCrawl vectors, the median value of the cosines of the angles between pairs of direction vectors of independently trained models was 0.9728. We can synthesize a single model from 15 independently trained models. To do this, we average the direction vectors of individual models, and normalizing the resulting vector to unit length, we obtain the direction vector of a single synthetic model V :

$$V = \frac{\sum_i v_i}{\|\sum_i v_i\|}$$

The median value of the projections of the direction vectors of the 15 models v_i onto the direction V for the set of

Table 1

Pearson's (r) and Spearman's (ρ) correlation coefficients between averaged estimates from independently trained models and human ratings

Used embeddings	Predictor type	r	ρ
fasttext-CommonCrawl	FNN	0.8655	0.8688
fasttext-CommonCrawl	linear	0.8411	0.8502
GloVe-840B	FNN	0.8541	0.8577
GloVe-840B	linear	0.8361	0.8418
fasttext-wiki	FNN	0.8390	0.8418
fasttext-wiki	linear	0.8179	0.8251
MuSE	FNN	0.8388	0.8414
MuSE	linear	0.8183	0.8255
Co-occurrence vectors (CFW)	FNN	0.8512	0.8540

pre-trained fasttext-CommonCrawl vectors is 0.9842. This proves the high degree of consistency of all 15 models.

Interestingly, using the synthetic model allows us to achieve a better accuracy compared to simple averaging of ratings. For example, the Spearman's correlation coefficient between human ratings and the average machine rating for a set of pre-trained fasttext-CommonCrawl vectors equals 0.8354. It was obtained using simple averaging. The Spearman correlation coefficient obtained using the synthetic model is 0.8502. Therefore, Table 1 provides values of the correlation coefficients for linear predictors obtained using the synthetic model.

Let us also compare the accuracy of predictors using different sets of pre-trained vectors. The fasttext-wiki and MuSE embeddings were obtained by training using the Wikipedia text corpus, which has a much smaller size compared to the CommonCrawl corpus, therefore, the predictors employing these vectors show lower accuracy. It should be noted that a slightly lower result was obtained using the Glove-840B pre-trained vectors compared to fasttext-CommonCrawl, despite a larger size of the training corpus. A similar observation was described in (Wang et al., 2023). Apparently, this is due to a higher quality of embedding training using the fasttext algorithm.

Diachronic Predictor of Socialness Ratings of Words

As can be seen from Table 1, the use of explicit word vectors built using the CFW method allows us to obtain almost the same prediction accuracy as employing low-dimensional vectors trained on large-volume corpora. Despite a slightly lower accuracy, the CFW method has the great advantage of

easily obtaining a diachronic model if we have an appropriate corpus (see, for example, (Bochkarev et al., 2022)). To do this, we only need to build explicit word vectors for the target time intervals, using co-occurrence data of a target word in a diachronic corpus. Then, the obtained vectors are fed to the predictor input; thus, we obtain diachronic estimates of the socialness rating of the target word.

The Google Books Ngram corpus provides annual frequency data on words and phrases. Accordingly, we built vectors for each word present in the (Diveica et al., 2023) dictionary using the GBN corpus data for each year from 1870 to 2019 and calculated the corresponding socialness rating estimates of these words for each year. The Pearson's and Spearman's correlation coefficients between human ratings and machine estimates calculated using annual data are shown in Figure 1.

The highest value of the Spearman's correlation coefficient is 0.8531 (in 2010), which is only a few ten-thousandths less than the value of the Spearman's correlation coefficient given in Table 1 obtained using data of the entire time interval 1900-2019. No year in the interval 2000-2019 shows drop of the Spearman's correlation coefficient value below 0.8480. The annual size of the English subcorpus of Google Books Ngram for these years is between 22.8 and 34.9 billion words. As can be seen from Figure 1 and the presented values such data size provides estimates that are no less accurate than those obtained using all available data. For earlier years, as can be seen from Figure 1, the correlation coefficients between human ratings and their machine estimates calculated from annual data are smaller. The main reason for this is the decrease in the annual corpus size for earlier years. In addition, language evolution can cause changes in the socialness ratings of words over time. Since human ratings are obtained as a result of surveys conducted in recent years, this phenomenon can also lead to a drop in the cor-

Figure 1

The Pearson's (r) and Spearman's (ρ) Correlation Coefficients between Human Ratings and Machine Estimates Calculated Using Annual Data

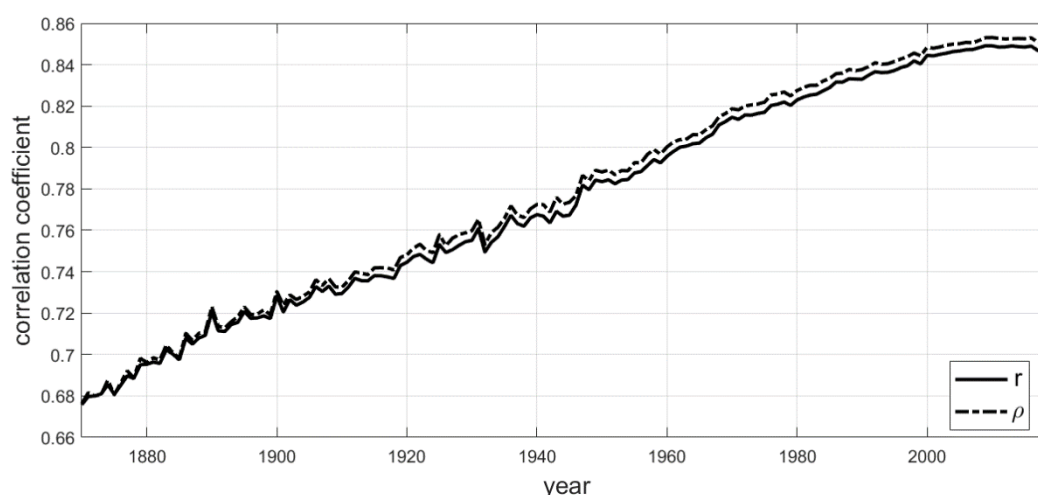
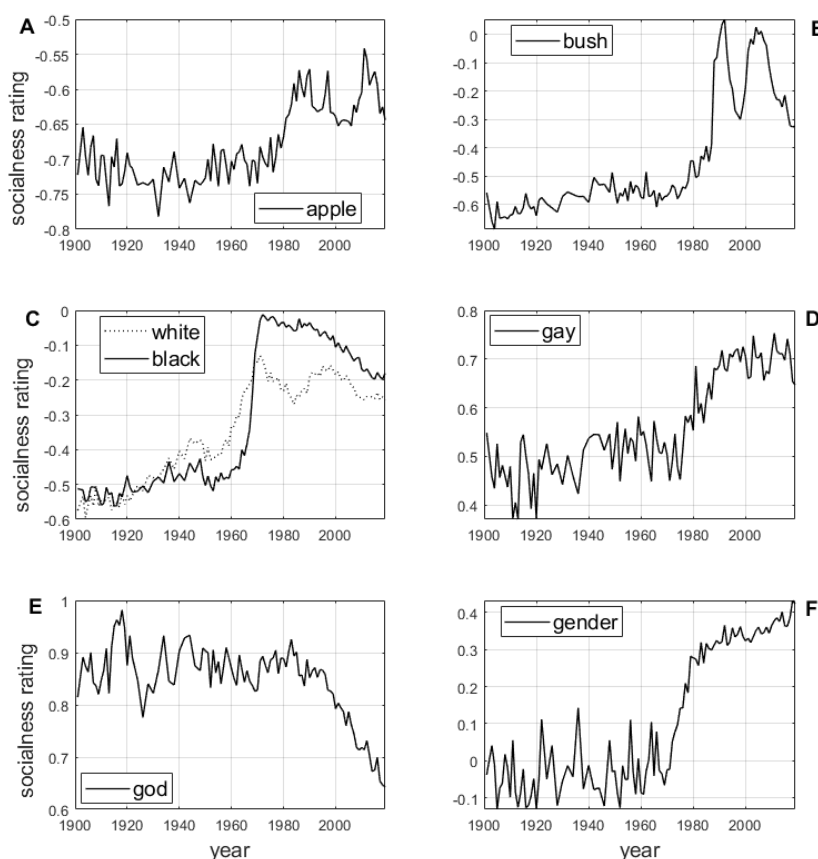


Figure 2

Examples of Words for which Socialness Ratings Change Over Time



relation coefficient between human socialness ratings and their machine estimates for earlier years.

Below we will analyse some examples of trends of socialness rating change trying to reveal possible regularities of rating change. The classification is not strict, just made for more convenient analysis.

The first example of socialness rating change is represented by a word *apple* (see Figure 2, A). Until the end of the 70s, the rating was negative and was fluctuating around -0.7. The graph shows a rapid growth of its ratings values since the end of 80s. This can be explained by the fact that *apple* traditionally denoting a kind of fruit gained a new denotata, which is a multinational corporation. The fastest growth in socialness ratings is observed at the time of the release of the Macintosh computer model, when Apple personal computers gained a wide popularity (Linzmayr, 2004). And now *Apple* is a world-famous brand associated with high social status. Thus, emergence of new meaning triggered changes of socialness ratings. The second burst on the graph is also not accidental; the growth of the socialness rating in

this case coincides with the launching of families of mobile devices by Apple.

Another example is a word *bush*, which shows a similar tendency as the previous word (see Figure 2, B). The main meaning of this word is “woody-stemmed plant that grows much shorter and wider than a tree”. Therefore, the socialness rating of this word was initially negative for a long time remaining around the value of -0.6. However, this word also means a surname of German origin. In particular, this is a surname of a famous political dynasty in the USA (Schweizer & Schweizer, 2004). Starting with the time of the 2nd World War, one can see a gradual increase in the socialness rating of the word *bush*, associated with an increase in the percentage of use of this word as a proper name. The growth accelerated in the 70s. Career of a famous American politician George H. W. Bush started in 60s and was widely discussed in press. Particularly significant jumps in socialness ratings are observed in 1988 (when George H. W. Bush won the US presidential election) and in 2000 (when his son George W. Bush won the election).

More examples of words that have changed its socialness ratings are *white* and *black* (see Figure 2,C). Originally they denote colours. However, due to metonymical shift they also denote individuals, social groups of people distinguished by complexion. The below graphs show how socialness ratings of these words evolved. For the 19th century, the socialness ratings of both words range from -0.6 to -0.55. Tendency to the ratings growth has been observed since 1920, after the 1st World War with rapid jump in 60s. The peak can be explained by activation of the African American civil rights movement. It is interesting that the greatest change in the ratings occurred in 1968, when Martin Luther King, a prominent leader in the African American civil rights movement, was assassinated. This tragic event caused a huge resonance in American society.

One more example of meaning change that caused word rating change is the word *gay* (see Figure 2,D), which changed both denotata and reference to a particular part of speech (POS). *Gay* (cheerful) as an adjective has been a word with moderately low socialness ratings (in the range from 0.45 to 0.50). However, since the beginning of 80s one can observe their rapid growth. Also, being a noun, *gay*, denotes a homosexual person; and in this sense it is a "more social" concept. Though *gay* as a homosexual person appeared long ago, however, until recent times it was not widely used. The increase in the frequency of the word in the second meaning in the 1980s was associated with the social processes in the USA and triggered growth of socialness ratings.

There are some words denoting abstract notions and social concepts which socialness ratings change mostly due to the change of cultural context and perception. Let us consider the words *god* and *gender* (see Figures 2, E, F). *God* has always been a word with a high socialness rating. However, its ratings also fluctuate. The highest peaks are observed in 1914 – 1921 (the time of the 1st World War and some years after it) and slightly lower peak is detected in 1940 -1945 (the time of the 2nd World War). The maximum rating value of 0.982 was reached in 1918, while the annual rating values were mostly in the range of 0.85-0.9. An extremely interesting phenomenon is the current trend towards a decrease in the socialness rating of the word *god*. The trend has emerged since the last years of the 20th century. This issue requires additional research, however, it can be assumed that this is due to a tendency towards a more personal perception of the idea of *god*.

One more interesting example is the word *gender* that basically means sex. Its socialness ratings were almost neutral until 1970s. However, starting from the beginning of 1970s one can observe a rapid upgoing trend. According to the etymological dictionary¹, no new meanings for the word *gender* have emerged in the 20th century. However, as a re-

sult of public debate, its perception has been changing, and this concept is being rethought as more socially significant, which caused growing of socialness ratings.

Transferring of Socialness Ratings to Other Languages

Using trained linear predictors for the fasttext-wiki and MuSE pre-trained vector sets for English, we obtained machine dictionaries with socialness ratings for 43 and 28 other languages, respectively. The main challenge is to check the quality of the obtained socialness rating estimates for each of these languages.

This can be done, firstly, by selective manual check of the obtained machine ratings (total manual check is practically impossible due to the large size of the obtained dictionaries). We checked Russian words with the highest and lowest ratings. The checking showed that the model coped with the task very well. Words with high socialness ratings were at the top of the list, among them are *obshchestvennost'*, *partnerstvo*, *druzhestvennost'*, *vezhliviy*, *demokratichnost'* (public, partnership, friendliness, politeness, democracy). Words with low ratings are at the bottom of the list and include such words as *cellofan*, *bryzhejka*, *nubuk*, *struchok*, *peschanik* (cellophane, mesentery, nubuck, pod, sandstone).

Secondly, when both fasttext-wiki and MuSE pre-trained vectors are available for a language, we can compare ratings obtained using each of these two sets, which (to some extent) makes it possible to judge the quality of the obtained ratings. Table 2 shows the values of the Pearson's and Spearman's correlation coefficients between the socialness ratings obtained using the two sets of aligned pre-trained vectors.

Higher correlation coefficient values indicate a greater degree of similarity between the machine ratings obtained using two different sets of vectors, and thus indirectly indicate a greater degree of reliability of the results for a given language.

As for Chinese, only vectors from the fasttext-wiki multilingual dataset are available for it. However, independently obtained human socialness ratings for 17,940 Chinese words are available in (Wang et al., 2023). The Pearson's and Spearman's correlation coefficients between the human socialness ratings from that work and our machine ratings are 0.6010 and 0.6382, respectively. For 64,791 Chinese words, both the machine ratings available in (Wang et al., 2023) and our machine ratings are available. The Pearson and Spearman correlation coefficients between these two sets of ratings are 0.5828 and 0.5827, respectively. It should be noted that among the 64.8 thousand words mentioned,

¹ Online Etymology Dictionary. (n.d.). Gender. In Online Etymology Dictionary. Retrieved July 15, 2024, from <https://www.etymonline.com/search?q=gender>

Table 2
The Pearson`s (r) and Spearman`s (p) Correlation Coefficients between Socialness Ratings Obtained Using the Fasttext-Wiki and MuSE Aligned Pre-Trained Vector Sets

Language	r	p	Language	r	p
Bulgarian	0,8106	0,7910	Indonesian	0,8763	0,8638
Catalan	0,9111	0,9005	Italian	0,9409	0,9316
Czech	0,8730	0,8617	Macedonian	0,8221	0,8067
Danish	0,8818	0,8728	Dutch	0,9300	0,9201
German	0,9062	0,8959	Norwegian	0,9206	0,9137
Greek	0,8256	0,8032	Polish	0,8842	0,8745
English	0,9997	0,9996	Portuguese	0,9436	0,9386
Spanish	0,9492	0,9427	Romanian	0,8889	0,8775
Estonian	0,8922	0,8859	Russian	0,8649	0,8512
Finnish	0,8540	0,8420	Slovak	0,8240	0,8107
French	0,9259	0,9149	Slovenian	0,8357	0,8216
Hebrew	0,6948	0,6656	Swedish	0,9100	0,8922
Croatian	0,8901	0,8833	Turkish	0,8328	0,8162
Hungarian	0,8628	0,8528			

there is a significant percentage of rare and low-frequency words for which machine ratings are less accurate. This influenced the decrease in the correlation level.

DISCUSSION

The survey method used to create dictionaries with psycholinguistic ratings is rather labour-intensive. Therefore, dictionaries with human ratings are relatively small size. Development of natural language processing technologies triggered appearance of a significant number of works devoted to extrapolation of human ratings to a wide range of words. In this way, large machine dictionaries have been created for many psycholinguistic parameters, such as dictionaries of affective, concreteness and imageability ratings (Mohammad et al., 2013; Koper & Schulte im Walde, 2016; Charbonnier & Wartena, 2019).

At the same time, only one work is devoted to the recently introduced socialness rating, which attempts to build large machine dictionaries for Chinese and English. However, in this work, the ratings for English were obtained by transferring ratings from Chinese; and social weight of words from different languages may not be similar.

The following results were achieved in the present paper. Firstly, synchronic models of socialness ratings were trained for English and a dictionary with socialness ratings for 2 million words was compiled by using the obtained ratings. Secondly, a diachronic model of socialness ratings was also trained for English and examples of changes in the percep-

tion of words as related or not related to social were considered. Finally, using the aligned sets of pre-trained vectors, the obtained rating estimates were transferred to 43 other languages.

Synchronic Models

The constructed models allow obtaining estimates of the socialness ratings of English words with a fairly high accuracy. The best value of the Spearman`s correlation coefficient between human ratings and their estimates was 0.8688. This value is close to the values of the correlation coefficients of human and machine ratings obtained when predicting affective ratings, concreteness ratings and other psycholinguistic characteristics of English words (Buechel & Hahn, 2018; Charbonnier & Wartena, 2019; Bochkarev et al., 2021). There are a number of independently obtained English dictionaries with affective and concreteness ratings by different groups of researchers. The comparisons carried out in (Charbonnier & Wartena, 2019) showed that the achieved level of correlation of human and machine ratings is already close to the level of correlation of human ratings presented by different groups.

It is problematic to conduct similar comparisons for the socialness rating because the dictionary presented in (Diveica et al., 2023) is still the only large English dictionary with socialness ratings.

The Pearson`s correlation coefficient between the human ratings presented in (Diveica et al., 2023) and (Binder et al., 2016) is 0.76, but it is calculated only for 258 words that are

included in both dictionaries. It is mentioned in (Wang et al., 2023) that the Pearson's correlation coefficient between the human ratings obtained by translating Chinese words from the dictionary proposed by the authors of this work and the ratings of the dictionary by (Diveica et al., 2023) is 0.724. Thus, the level of correlation between human and machine ratings obtained in our work (see Table 1) significantly exceeds the level of correlation between human ratings obtained by different researchers. We also calculated the Pearson's and Spearman's correlation coefficients between the machine ratings we obtained (neural network predictor, fasttext-CommonCrawl vectors) and the ratings of 535 words from the dictionary by (Binder et al., 2016). They were 0.7545 and 0.7944, respectively, which is no lower than the correlation coefficients between the human ratings given in the two dictionaries.

Besides, similarity of the obtained value of the correlation coefficients of human and machine socialness ratings with the correlation coefficients of human and machine ratings for other psycholinguistic parameters suggests that the obtained level of accuracy in predicting socialness ratings is also close to the maximum achievable.

Diachronic Models

The low-dimensional pre-trained vectors available in the public domain were obtained by training on synchronic text corpora. Thus, they are not suitable for obtaining diachronic estimates of word socialness ratings. In contrast, the model that employs explicit word vectors allows one to easily obtain diachronic estimates of word socialness ratings using any diachronic corpus of sufficient size. We a priori expect that the perception of socialness of many words may change over time. Indeed, as soon as we begin to consider specific examples, we immediately reveal cases of such changes. The examples considered show, firstly, that the socialness rating of a word may undergo abrupt changes when the word acquires new meaning, connotation or due to the change of cultural context. A complete classification of cases of abrupt word sociality ratings requires a separate large study; in this paper, we considered only a few examples illustrating various possible directions for further work. Nevertheless, the examples given show that a change in the concreteness rating may be a marker of lexical semantic change.

A large number of works are devoted to the task of lexical semantic change detection (Tang, 2018; Hengchen et al. 2021). In most cases, such works use one or another diachronic vector representation of words. A change in the vector representing the word or a change in the word direct context in the vector space is considered as an indicator of meaning change. An alternative approach is also possible, first described in (Ryzhova et al., 2021), when the statistics

of the use of words in the text in one grammatical form or another (grammatical profiles of the word) are considered, and a change in such statistics serves as a marker of lexical semantic change. In the work (Ryzhova et al., 2021), only grammatical features of words were considered, however, for lexical semantic change detection, such features as the use of a word as a proper name or a common noun (Bochkarev et al., 2022), as well as psycholinguistic characteristics of words (Bochkarev et al., 2024a) can also be used. Changes in the socialness rating of words can also serve as additional markers of semantic changes.

Also, in some of the cases considered, we encounter the fact that the meaning of the word does not change, however, its socialness rating changes due to some cultural reasons. Thus, the diachronic model of the socialness ratings of words can be useful in cultural studies.

Rating Transfer

The presence of word embeddings aligned in a single vector space made it easy to transfer socialness ratings from English to 43 other languages. The main problem is how to verify the obtained machine ratings for other languages. A spot check for Russian showed that words that received large positive or large negative machine ratings are usually estimated adequately by the model. A complete manual check of ratings for all languages is extremely labor-intensive and is currently beyond our capabilities. However, for 28 languages we have independent estimates of socialness ratings obtained using two sets of vectors - fasttext-wiki and MuSE. Comparison of the ratings obtained by two independent methods allows us to judge the quality of the resulting dictionaries.

Two factors should be considered while interpreting correlation coefficients presented in Table 2. Initial human ratings were obtained for English, therefore, languages that are more related to English shows better correlation. For example, rating correlation with German and Danish is high and Vietnamese shows the lowest one. Obviously, word social significance is similarly precepted in these languages. The second factor to be considered is that there are more Wikipedia texts used for training written in European languages than in the other ones.

A very important result is an unexpectedly high efficiency of linear predictors in predicting the socialness rating. It was shown in the previous section that the results of linear predictors can be just slightly improved by more complex neural network predictors, which have 5 orders of magnitude more fitting parameters. In this case, estimation of socialness ratings differs sharply from what is observed for other psycholinguistic parameters, such as affective ratings, concreteness ratings and imageability. In all the mentioned

cases, except for the socialness rating, linear predictors are very much inferior to neural network predictors in accuracy.

It should also be noted that the linear predictors independently trained on different subsets of words were found to be highly consistent with each other. This proves that, in the space of vectors representing words, there is a distinguished direction responsible for the perception of words as related or not related to the social sphere. Relationships in society play a vital role in human life, which cannot but be reflected in language. Apparently, the significant role of social factors is captured by existing language models, which leads to the appearance of a distinguished direction in the vector space responsible for the degree of perception of a word as related to the social. Thus, the socialness rating of a word in the vector space of vectors representing words can be characterized very simply. The socialness rating grows along the distinguished direction. Accordingly, rating estimates in the first approximation can be obtained as projections of word vectors onto this direction.

The limitations of the present study may be, firstly, related to the features of the dictionary with human ratings used for training the model. As was shown in (Bochkarev et al., 2024b), differences in the composition of the lexicon of affective dictionaries created by the survey method can lead to biases in the obtained machine ratings. It is not yet possible to conduct a similar study for word socialness ratings due to the above mentioned fact that, at the moment, the dictionary by (Diveica et al., 2023) is the only large dictionary of the English language with socialness ratings. The second obvious limitation is related to the fact that the existing models do not provide ratings for different meanings of polysemantic words. Progress in this direction can be achieved by using context-sensitive word embeddings.

CONCLUSION

This paper has solved the problem of compiling a large dictionary with socialness ratings of English words by using proposed computer models. The accuracy of the developed models is high: the best achieved value of the Spearman's correlation coefficient between human ratings and their machine estimates is 0.8688. The employed models allowed us to extrapolate human ratings to a very wide range of words and we managed to obtain machine ratings for two million words. Therefore, the resulted dictionary of word socialness ratings is several times larger than those created before. Also, ratings for a wide range of words from 43 other languages were obtained by using freely available word embeddings aligned in a single vector space. Besides, a diachronic predictor of socialness rating was constructed using explicit word vectors.

High efficiency of linear predictors in the task of predicting socialness ratings was unexpected. In fact, it is enough to simply find the projection of a vector representing a word onto some selected direction in the vector space, and get a good estimate of the socialness of the word. Such a simple estimate can be further just slightly improved, however, it is a very labour- and time-consuming process. We suppose that as relationships in society play a vital role in human life, the significant impact of social factors is captured by existing language models and leads to the appearance of a distinguished direction in the vector space responsible for the degree of perception of a word as related to the social.

Also, a diachronic predictor of socialness rating is constructed using explicit word vectors. It is shown that using word co-occurrence statistics in a large diachronic corpus, it is possible to detect changes in socialness ratings over time.

The obtained results can be useful for several fields of science. The created dictionary is a good material for psycholinguistic and cultural studies. Moreover, as the analyzed examples of words illustrate that change in socialness rating can be a marker of lexical semantic change, the diachronic model can be used for etymological studies.

There are some directions for further work. The first possible one is to obtain socialness ratings for polysemantic words using context-sensitive word embeddings. Another one is to use context-sensitive embeddings to improve efficiency of transferring ratings to other languages.

ACKNOWLEDGMENTS

This research was financially supported by Russian Science Foundation, grant № 20-18-00206.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHOR CONTRIBUTION

Vladimir Bochkarev: conceptualization; methodology; writing – original draft; supervision.

Anna Shevlyakova: conceptualization; writing – original draft, review and editing.

Andrey Achkeev: software, visualization.

REFERENCES

- Basile, P., & McGillivray, B. (2018). Exploiting the web for semantic change detection. *Lecture Notes in Computer Science*, 11198, 194-208. https://doi.org/10.1007/978-3-030-01771-2_13
- Binder, J.R., Conant, L.L., Humphries, C.J., Fernandino, L., Simons, S.B., Aguilar, M., & Desai, R.H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3-4), 130-74. <https://doi.org/10.1080/02643294.2016.1147426>
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, 135-146. https://doi.org/10.1162/tacl_a_00051
- Bochkarev, V. V., Savinkov, A. V., & Shevlyakova, A. V. (2021). Estimation of imageability ratings of English words using neural networks. In I. Batyrshin, A. Gelbukh, & G. Sidorov (Eds.), *Advances in Soft Computing. MICAI 2021. Lecture Notes in Computer Science* (vol. 13068, pp. 59-69). Springer. https://doi.org/10.1007/978-3-030-89820-5_5
- Bochkarev, V. V., Khristoforov, S. V., Shevlyakova, A. V., & Solovyev, V. D. (2022). Neural network algorithm for detection of new word meanings denoting named entities. *IEEE Access*, 10, 68499-68512. <https://doi.org/10.1109/ACCESS.2022.3186681>
- Bochkarev, V., Khristoforov, S., Shevlyakova, A., & Solovyev, V. (2024). Diachronic analysis of a word concreteness rating: Impact of semantic change. *Lobachevskii Journal of Mathematics*, 45, 961-971. <https://doi.org/10.1134/S1995080224600559>
- Bochkarev, V., Solovyev, V., Nestik, T., & Shevlyakova, A. (2024). Variations in average word valence of Russian books over a century and social change. *Journal of Mathematical Sciences*, 285, 14-27. <https://doi.org/10.1007/s10958-024-07419-z>
- Buechel, S., & Hahn, U. (2018). Word emotion induction for multiple languages as a deep multi-task learning problem. *Proceedings of NAACL-HLT 2018* (pp. 1907-1918). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-1173>
- Bullinaria, J., & Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39, 510-526. <https://doi.org/10.3758/BF03193020>
- Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods*, 44(3), 890-907. <https://doi.org/10.3758/s13428-011-0183-8>
- Charbonnier, J., & Wartena, C. (2019). Predicting word concreteness and imagery. *Proceedings of the 13th International Conference on Computational Semantics - Long Papers* (pp. 176-187). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-0415>
- Conneau, A., Lample, G., Denoyer, L., Ranzato, M. A., & Jégou, H. (2017). *Word translation without parallel data*. arXiv preprint arXiv:1710.04087.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (vol. 1: Long and Short Papers, pp. 4171-4186). Association for Computational Linguistics.
- Diveica, V., Pexman, P. M., & Binney, R. J. (2023). Quantifying social semantics: An inclusive definition of socialness and ratings for 8388 English words. *Behavior Research Methods*, 55, 461-473. <https://doi.org/10.3758/s13428-022-01810-x>
- Firth, J. (1957). A synopsis of linguistic theory, 1930-55. *Studies in linguistic analysis* (Special Volume of the Philological Society, pp. 1-31). Blackwell.
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. ELRA.
- Gulordava, K., & Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics* (pp. 67-71). Association for Computational Linguistics.
- Harris, Z. (1970). *Papers in structural and transformational linguistics*. Reidel.
- Hengchen, S., Tahmasebi, N., Schlechtweg, D., & Dubossarsky, H. (2021). Challenges for computational lexical semantic change. In N. Tahmasebi, L. Borin, A. Jatowt, Y. Xu, & S. Hengchen (Eds.), *Computational approaches to semantic change* (pp. 341-372). Language Science Press.
- Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., & Grave, E. (2018). Loss in translation: Learning bilingual word mapping with a retrieval criterion. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2979-2984). Association for Computational Linguistics.
- Khristoforov, S., Bochkarev, V., & Shevlyakova, A. (2020). Recognition of parts of speech using the vector of bigram frequencies. In W. van der Aalst et al. (Eds.), *Analysis of images, social networks and texts. AIST 2019, Communications in Computer and Information Science* (vol. 1086, pp. 132-142). Springer. https://doi.org/10.1007/978-3-030-39575-9_13

- Koper, M., & Schulte im Walde, S. (2016). Automatically generated affective norms of abstractness, arousal, imageability and valence for 350,000 German lemmas. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 2595–2598). ELRA.
- Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books Ngram Corpus. In H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, & J. C. Park (Eds.), *50th Annual Meeting of the Association for Computational Linguistics 2012: Proceedings of the Conference* (vol. 2, pp. 238–242). Association for Computational Linguistics.
- Linzmayr, O. (2004). *Apple confidential 2.0: The definitive history of the world's most colorful company* (2nd ed.). No Starch Press.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* (vol. 26, pp. 3111–3119). Curran Associates, Inc.
- Mohammad, S., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *Second Joint Conference on Lexical and Computational Semantics (SEM)* (vol. 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, pp. 321–327). Association for Computational Linguistics.
- Pantel, P. (2005). Inducing ontological co-occurrence vectors. *Proceedings of the 43rd Conference of the Association for Computational Linguistics* (pp. 125–132). Association for Computational Linguistics. <https://doi.org/10.3115/1219840.1219856>
- Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1162>
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (vol. 1: Long Papers, pp. 2227–2237). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-1202>
- Pexman, P. M., Diveica, V., & Binney, R. J. (2022). Social semantics: The organization and grounding of abstract concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1870), 20210363. <https://doi.org/10.1098/rstb.2021.0363>
- Pilehvar, M. T., & Camacho-Collados, J. (2020). *Embeddings in natural language processing: Theory and advances in vector representations of meaning*. Morgan & Claypool Publishers.
- Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10), 627–633. <https://doi.org/10.1145/365628.365657>
- Ryzhova, A., Ryzhova, D., & Sochenkov, I. (2021). Detection of semantic changes in Russian nouns with distributional models and grammatical features. *Komp'yuternaya Lingvistika i Intellektual'nye Tehnologii*, 20, 597–606.
- Schweizer, P., & Schweizer, R. (2004). *The Bushes: Portrait of a dynasty* (1st ed.). Doubleday.
- Tang, X. (2018). A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5), 649–676. <https://doi.org/10.1017/S1351324918000220>
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1), 141–188. <https://doi.org/10.1613/jair.2934>
- Wang, S., Zhang, Y., Shi, W., et al. (2023). A large dataset of semantic ratings and its computational extension. *Scientific Data*, 10, 106. <https://doi.org/10.1038/s41597-023-01995-6>
- Weeds, J., Weir, D., & McCarthy, D. (2004). Characterising measures of lexical distributional similarity. *Proceedings of the 20th International Conference on Computational Linguistics* (pp. 1015–1021). COLING.
- Worth, P. (2023). Word embeddings and semantic spaces in natural language processing. *International Journal of Intelligence Science*, 13, 1–21. <https://doi.org/10.4236/ijis.2023.131001>
- Xu, Y., & Kemp, C. (2015). A computational evaluation of two laws of semantic change. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci 2015)*. Association for Computational Linguistics.

<https://doi.org/10.17323/jle.2024.22244>

Wrong Answers Only: Distractor Generation for Russian Reading Comprehension Questions Using a Translated Dataset

Nikita Login 

HSE University, Moscow, Russia

ABSTRACT

Background: Reading comprehension questions play an important role in language learning. Multiple-choice questions are a convenient form of reading comprehension assessment as they can be easily graded automatically. The availability of large reading comprehension datasets makes it possible to also automatically produce these items, reducing the cost of development of test question banks, by fine-tuning language models on them. While English reading comprehension datasets are common, this is not true for other languages, including Russian. A subtask of distractor generation poses a difficulty, as it requires producing multiple incorrect items.

Purpose: The purpose of this work is to develop an efficient distractor generation solution for Russian exam-style reading comprehension questions and to discover whether a translated English-language distractor dataset can offer a possibility for such solution.

Method: In this paper we fine-tuned two pre-trained Russian large language models, RuT5 and RuGPT3 (Zmitrovich et al, 2024), on distractor generation task for two classes of summarizing questions retrieved from a large multiple-choice question dataset, that was automatically translated from English to Russian. The first class consisted of questions on selection of the best title for the given passage, while the second class included questions on true/false statement selection. The models were assessed automatically on test and development subsets, and true statement distractor models were additionally evaluated on an independent set of questions from Russian state exam USE.

Results: It was observed that the models surpassed the non-fine-tuned baseline, the performance of RuT5 model was better than that of RuGPT3, and that the models handled true statement selection questions much better than title questions. On USE data models fine-tuned on translated dataset have shown better quality than that trained on existing Russian distractor dataset, with T5-based model also beating the baseline established by output of an existing English distractor generation model translated into Russian.

Conclusion: The obtained results show the possibility of a translated dataset to be used in distractor generation and the importance of the domain (language examination) and question type match in the input data.

KEYWORDS

automatic distractor generation, multiple-choice questions, reading comprehension, large language model, dataset translation

Citation: Login, N. (2024). Wrong answers only: Distractor generation for Russian reading comprehension questions using a translated dataset. *Journal of Language and Education*, 10(4), 56-70. <https://doi.org/10.17323/jle.2024.22244>

Correspondence:
Nikita Login,
nlogin@hse.ru

Received: August 15, 2024

Accepted: December 16, 2024

Published: December 30, 2024

INTRODUCTION

Automatic question generation is a promising sphere for application of natural processing techniques as it can enhance the educational processes in multiple ways. According to (Kurdi et al., 2020), standardised examination usually requires exam

organisers to keep large banks of curated test exercises which should be regularly updated to prevent cheating. With automated generation these banks can be populated continuously, ensuring variability of test exercises and reducing the costs of organising exams. Furthermore, automatic exercise generation can help



test takers, as it can provide them with almost infinite source of test items for preparation.

The availability of sufficient-quality training data is crucial for automatic question generation. Most of the datasets used for model training in automatic question generation were originally designed around the machine reading comprehension problem – these include RACE (Lai et al., 2017), SciQ (Welbl et al., 2017), SQuAD (Rajpkar et al., 2016), COQA (Reddy et al., 2019), Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). Most of these datasets include items consisting of a text passage for reading, a set of questions accompanying the text, right answer and (optionally) a set of distractors for each question. However, there are datasets designed specifically for question generation, including QGSTEC (Rus et al., 2012) and FairyTaleQA (Xu et al., 2022). Among these datasets, RACE is notable for containing exam-style questions as its items were originally extracted from Chinese websites containing English examination materials. For Russian language there are question datasets, such as DaNetQA (Glushkova et al., 2021), MuSeRC/RuCoS (Fenogenova et al., 2020), SberQUAD (Efimov et al., 2020) and RuBQ (Rybin et al., 2021). DaNetQA and SberQUAD contain crowdsourced questions corresponding to Wikipedia paragraphs, RuBQ is based on quizzes and Wikidata, MuSeRC and RuCoS contain crowdsourced questions corresponding to text paragraphs retrieved from a variety of sources. Among Russian datasets, MuSeRC is notable as the only dataset to contain distractors.

Distractor generation is a particularly important subtask of automatic question generation. The advantage of incorporating distractors in online testing materials is that it allows for immediate automated test grading, while excluding the possibility of unfair judgement (as in case with answer matching for open questions). However, this subtask remains one of the most difficult due to the following reasons:

- (1) In distractor generation multiple outputs (different independent distractors) correspond to a single input
- (2) There cannot be a closed set of ground-truth distractors for a given question, so it is difficult to estimate the performance of a trained model
- (3) The generated outputs need to be incorrect in context of the given question but correct in terms of language (Kurdi et al., 2020, p. 145) and also not be too irrelevant to the question

Due the rapid development of neural networks in 2020–2024 years, the most trending approach to question generation nowadays is neural network-based. It is mostly implemented in one of the three ways:

- (1) By training/fine-tuning a sequence-to-sequence model (Lee et al., 2020; Makhnytkina et al., 2020; Xiao et al., 2020; Xu et al., 2022; Hadifar et al., 2022; Manakul et al., 2023; Zhang, 2023)

- (2) By fine-tuning an autoregressive (designed for text continuation) large language model (Belyanova et al., 2022)
- (3) By prompting a large instruct/chat-based model (Elkins et al., 2023; Wang et al., 2023).

Distractor generation generally implements the same techniques – Seq2Seq (Qiu et al., 2020; Hadifar et al., 2022; De-Fitero-Dominguez et al., 2024; Ghanem & Fyshe, 2024), autoregressive (Chung et al., 2020; Ghanem & Fyshe, 2024) and prompting (Bitew et al., 2023; Maity et al., 2024) approaches.

Question and distractor generation are usually automatically evaluated by metrics originally designed for machine translation and text summarization, such as BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005) and ROUGE (Lin, 2004). BLEU is based on the geometric mean of modified n-gram precision values. Modified n-gram precision is calculated as the ratio of words in generated sequence that appear in the ground-truth sequence with respect to the unique word counts in the latter. The maximum word length of n-grams used while calculating BLEU is used as an indicator of a specific variant of this metric (BLEU-1, BLEU-2, ...). ROUGE metric can be based on recall, precision or their harmonic mean (F-score) with equal weights and has variants depending on n-gram match (ROUGE-N) as well as on Longest Common Subsequence (ROUGE-L). METEOR was developed to address the found issues of BLEU (lack of recall and noisiness of analysed n-grams) and is based on the F-score of unigram match with greater weight of recall over precision.

Fine-tuning of sequence-to-sequence models remains the most popular solution for tackling question text generation problem. (Lee et al., 2020) implemented a BiLSTM-based question generation model jointly trained on two tasks – right answer prediction and question text prediction. Xiao et al. (2020) trained a custom Multi-Flow Attention Transformer (Vaswani et al., 2017) model on question text prediction task using SQuAD dataset. Xu et al. (2022) fine-tuned BART (Lewis et al., 2020) model on FairyTaleQA data for question text generation and reached ROUGE-L F1 score of 52.7. Hadifar et al. (2022) fine-tuned a T5 (Raffel et al., 2020) model for question text generation task on EduQG and SQuAD data, reaching BLEU-4, METEOR and ROUGE-L scores of 15.41, 29.65 and 34.26 correspondingly. Wang et al. (2023) implemented non-fine-tuned GPT-2 (Radford et al., 2019) prompting for question text generation, using a Beam Search extension named NeuroLogicDecoding (Lu et al., 2021). The technique was evaluated on ClariQ-FKw (Sekulić et al., 2021) dataset, reaching BLEU-4, ROUGE-L and METEOR scores of 21.61, 41.03 and 47.87 correspondingly.

As for question text generation for Russian language data, Makhnytkina et al. (2020) used a BiLSTM-based Encode-Decoder model trained on conversational dataset CoQA automatically translated to Russian using Yandex.Translator service. The model reached BLEU-2 score of 12.0. (Belyano-

va et al., 2022) implemented a RuGPT3 model fine-tuned on DaNetQA and RuBQ corpora. The generation was performed in autoregressive manner, question text was predicted as the continuation of input sequence, right answer text was not used. The model reached BLEU-4 of 4.75 and 1.95 on RuBQ and DaNetQA datasets correspondingly.

In distractor generation sequence-to-sequence approach is also popular. Qiu et al. (2020) used a Seq2Seq model consisting of an Attention-based encoder and a BiLSTM-based decoder for fine-tuning on distractors of RACE-DG dataset, a version of RACE specially pre-processed and filtered for distractor generation by (Gao et al., 2019). They used disjoint decoding in form of Beam Search algorithm on top of model-predicted word probability distributions to get multiple distractors from one input, using Jaccard score to obtain diverse option sets. Their model showed BLEU-4 scores of 7.57/6.27/5.27 for each of the three distractor options correspondingly. Chung et al. (2020) fine-tuned autoregressive BERT (Devlin et al, 2019) language model in a joint learning scheme on two tasks: sequential and parallel prediction of each token of the distractors. They used the same disjoint generation scheme as (Qiu et al., 2020) but applied Maximum Entropy criterion instead of Jaccard score. RACE-DG dataset was used for training and evaluation, the BLEU-4 and ROUGE-L scores on test subset were 13.56 and 34.01 correspondingly. However, later they released¹ enhanced versions of their models based on sequence-to-sequence BART architecture, that reached maximum BLEU-4/ROUGE-L of 16.33/37.5 correspondingly.

In more recent distractor generation works sequence-to-sequence T5 architecture is widely used. Hadifar et al. (2022) implemented distractor generation using a T5 model trained on both RACE and their own new dataset EduQG. The whole set of distractors was predicted at once. Obtained BLEU/METEOR/ROUGE-L scores on EduQG consisted 17.73/21.54/34.13 correspondingly. Ghanem & Fyshe (2024) fine-tuned GPT-2 and T5 models on distractor generation task as a part of work on their prediction-based distractor generation quality metric DISTO. They used RACE dataset for fine-tuning and evaluation and implemented two versions of T5 – with joint and disjoint distractor generation. Their best solution, a disjoint T5 model, reached 2.3 in terms of BLEU-4 while GPT-2 and joint T5 reached only 0.3 and 0.9 BLEU-4 scores correspondingly. De-Fitero-Dominguez et al. (2024) implemented distractor generation using mT5 (Xue et al, 2020) model, a multilingual version of T5, on a combined translated distractor dataset. Their dataset included items from RACE-DG, CosmosQA (Huang et al., 2019) and SciQ, translated with Opus-MT (Tiedemann & Thottingal, 2020) model. Their implementation reached 7.21 and 21.76 on test subset of RACE-DG in terms of BLEU-4 and ROUGE-L metrics.

Distractor generation was also treated as a ranking problem, as it was implemented by Bitew et al. (2022) – the models were trained to select the most appropriate distractors for the given question and a right answer to it. Two ranking solutions were implemented – one using feature engineering and logistic regression and another one using a multilingual BERT model. Three BERT-based models were used – based on distractor-right answer similarity, based on distractor-question similarity and a joint model combining the two beforementioned. Average precision and recall of ranking were used as quality metrics, and the highest scores (57.3 and 62.8 respectively) were obtained by a joint BERT-based model.

The most recent works also experimented with prompting approach to distractor generation. Bitew et al. (2023) addressed distractor generation by using a T5 model trained on Televic dataset and ChatGPT prompting in zero-shot (using a prompt without distractor examples) and few-shot (using a prompt with examples) configurations. All models were evaluated manually by experts. Maity et al. (2024) used a multi-level pipeline based on ChatGPT and DaVinci, consisting of input text paraphrase generation, keyword extraction from paraphrase, question generation and distractor generation itself. The best BLEU-4 and ROUGE-L scores (2.49 and 13.54 accordingly) were obtained by a Davinci-based multilevel model.

From all the reviewed works only Qiu et al. (2020) specifically address the issue of potential triviality of distractors in design of their solution. In their work they view triviality as irrelevance to the given question and reading passage, and claim to solve this issue by incorporating blocks that combine information from the reading passage and question text (referred as 'Reforming Passage' and 'Reforming Question' modules) to their original Transformer-based model. However, no clear conclusions on how the exclusion of both reforming modules affects the metrics (only exclusion of each module separately is analysed) are made and situations where triviality is not connected with the relatedness to input data are not accounted.

When there are numerous reading comprehension item generation solutions for English language, only a few were developed for Russian (Makhnytkina et al., 2020; Belyanova et al., 2022), and there was no evidence found of solutions that tackled distractor generation for Russian-language questions. Also, only one of existing Russian-language reading comprehension question datasets contains distractors, and none of these datasets is tackled specifically for language examination in reading comprehension. Another issue is that the parameters such as type and structure of questions are not utilised in distractor generation in the previous works, whereas accounting for these parameters

¹ <https://github.com/voidful/BDG>

in DG model design may have a potential of making the task easier for NN models. The importance of these parameters can be demonstrated by findings of Xu et al. (2022), who implemented question categorization in the design of their question dataset FairyTaleQA using a system of narrative elements and relations described in Paris & Paris (2003). They found that nature of the answer can depend on the narrative category of question on the example of “Feeling” question type.

Taking into account the importance of automatic exam-style reading comprehension test generation and the lack of solutions for distractor generation on Russian data, the purpose of this Paper is to develop an automatic distractor generation solution for Russian reading comprehension exam-style questions. Due to the lack of Russian exam-style distractor datasets, we also aimed to explore the possibility of using a translated high-quality English dataset for Russian distractor generation, as it was done by Makhnytkina et al. (2020) for question generation and by De-Fitero-Dominguez et al. (2024) for Spanish data. Additionally, we aimed to investigate the prospects of fine-tuning distractor generation models on specific categories of questions. We expected that a rich and thoroughly curated English dataset would serve as an efficient source of training data and that training on a specific category of questions would allow for better transferability of DG model intelligence to standardised examination questions. We formed our research questions as follows:

- RQ#1: Whether a distractor generation model can be effectively fine-tuned on an English dataset, that was automatically translated to Russian?
- RQ#2: Is there a need for a specific Russian multiple-choice reading comprehension question dataset for efficient exam-style distractor generation or is an existing non-exam-style dataset MuSeRC appropriate for this task?
- RQ#3: Can fine-tuning on a specific type of questions result in better performance of distractor generation model on standardised exam data?

METHOD

Research Design

In this work we have performed large language model fine-tuning experiments on the task of distractor generation for reading comprehension questions using different datasets. Our primary focus was on fine-tuning on a translated English-language dataset. For that purpose, we have used RACE, as it contains reading comprehension multiple-choice questions in language examination style and have been

used in many distractor generation works (Chung et al., 2020; Qiu et al., 2020; De-Fitero-Dominguez et al., 2024; Ghanem & Fyshe, 2024). Also, we have included an originally Russian multiple-choice reading comprehension dataset (MuSeRC) to see if a translated dataset was necessary and if it was possible to reach appropriate generation quality by using already available Russian data.

In our experiments we have fine-tuned two large language models made available (Zmitrovich et al., 2024) by AI-Forever team – RuGPT3 and RuT5, which are Russian-language implementations of GPT3 (Brown et al., 2020) and T5 (Raffel et al., 2020) models correspondingly. RuGPT3 employs autoregressive text generation and consists only of Transformer decoder blocks, while RuT5 is based on Sequence-to-Sequence approach and contains both encoder and decoder. For evaluation of our generated output, we have used conventional automatic Sequence-to-Sequence generation quality metrics (see “Assessment” subsection of “Methods”).

To compare models trained on different datasets we have also used a small set of original Russian examination data – USE, obtained from open-access Internet sources. We have also included baselines in our evaluation, including a non-fine-tuned version of RuGPT3 and enhanced versions of models from Chung et al. (2020), which outputs were translated to Russian automatically.

Datasets

RACE

RACE is a dataset consisting of 98,000 questions on English reading comprehension, designed for Chinese middle- and high-schoolers as a part of the national exam. Each text of RACE was accompanied by several multiple-choice questions, and each of the questions was accompanied by 5 answer options – 1 correct and 4 incorrect. We translated RACE dataset using Opus-MT English-to-Russian translation model available from EasyNMT² Python package. Each question and set of distractors were translated in concatenation with the reading text in order for the translation model to not lose context. After a descriptive analysis of RACE questions performed in Microsoft Excel and Python environments, we have found two distinct question categories suitable for distractor generation:

- Questions asking the participant to select the best title for the given passage (TITLE);
- Questions asking the participant to select TRUE or FALSE sentences from the given set (TF).

² <https://github.com/UKPLab/EasyNMT>

We collected questions of these types using regular expression search. Datasets obtained using this technique – Ru-RACE-TITLE and Ru-RACE-TF – contained 4892 and 3799 items correspondingly.

For Ru-RACE-TITLE we selected 805 unique question texts from RACE that matched the regular expression ***!Wtitle!W*** (contained the word ***title***). Then we manually filtered out 53 irrelevant question texts (e.g. containing word ***title*** referring to a person's social status or asking about the title of some item referenced in the reading text). The resulted dataset was split into train/test/dev subsets using original subset labels from RACE, which resulted in 4575/219/242 split.

For Ru-RACE-TF we selected question texts which lowercased variants matched the regular expression ***which of the following .+(true|false)***. This way 693 unique question texts were retrieved. 143 question texts were filtered out manually. Applying the same split logic as in Ru-RACE-TITLE, we got 3288/175/187 split. The formulation of Ru-RACE-TF task is identical to Task 18 in USE exam in Russian language, which allowed us to use USE data, as described in "USE-TF" section of this paper.

MuSeRC

To compare performance on the translated dataset with performance on the original Russian data, we have also used MuSeRC. MuSeRC is a dataset created by (Fenogenova et al, 2020) as a part of the RussianSuperGLUE benchmark. It contains 12,805 Russian multiple choice reading comprehension questions made by crowdsource workers from texts of different domains. Each text is accompanied by a set of questions, each question includes mostly 1-2 right answers and 2-3 distractors.

USE-TF

USE (Unified State Examination, *Единый Государственный Экзамен*, *Edinyi Gosudarstvennyi Examen*,) is a compulsory Russian state exam which is used for assessment of knowledge of high school graduates and as an entrance test for higher education institutions. Format of USE in subject of Russian language contains Task 18, which is a multiple-choice reading comprehension question asking the participants to select either TRUE or FALSE sentences from the given set. The data for this task contained test items collected by Shavrina et al. (2020), as well as obtained from other openly available Internet sources. This set contained 55 unique questions with 5 answer options for each. Some questions contained more than one correct option, so we have preprocessed them as described in "Data Preprocessing" section of this paper.

Methods

Data Preprocessing

In Ru-RACE-TITLE all question texts were replaced with *Какое название лучше всего подойдёт для этого текста?* ("Which is the best title for this text?"). In Ru-RACE-TF question texts were replaced with either *Какое высказывание COOTBETCTBYET тексту?* ("Which statement is TRUE according to the text?") or *Какое высказывание НЕ COOTBETCTBYET тексту?* ("Which statement is NOT TRUE according to the text?"). MuSeRC question and option texts were left unchanged as we wanted to perform training and evaluation on the whole original dataset splits.

For USE-TF we have applied the same preprocessing procedure as for RuRACE, with enhancements addressing having more than one correct option. For items that had more correct options than incorrect, we changed the question text to the opposite (*Какое высказывание COOTBETCTBYET тексту?* was changed to *Какое высказывание НЕ COOTBETCTBYET тексту?* and vice versa). Then first of the original distractors was used as the correct answer and original correct options were used as distractors. If there were more incorrect options than correct, the question was unchanged and the first right option was used as a right answer, while the distractors were used without changes.

Model Training

The models were trained on a remote private server with a Nvidia Tesla V100 GPU. All models were trained for 20 epochs with ADAM optimizer, initial learning rate of 5e-5 and weight decay factor of 0.01. We defined the maximum output length for training and inference as 0.99 quantile of input length on the training set. Training subsets were used for model fine-tuning, while testing and development subsets were exploited for evaluation.

For training phase, we constructed input examples for RuGPT3 as concatenations of reading passage, question text, right answer and a line-separated set of distractors, interrogated by Russian phrases indicating the parts of an input example (*ВОПРОС*, *ПРАВИЛЬНЫЙ ОТВЕТ* and *НЕПРАВИЛЬНЫЕ ВАРИАНТЫ ОТВЕТА*). For training of RuT5 we constructed separate input and output examples, as the model (as opposed to RuGPT3) worked not in an autoregressive but in a Sequence-to-Sequence way. The input example for RuT5 included a reading passage, a question and a right answer, interrogated by the same phrases as RuGPT3 examples, while the output example consisted of distractors enclosed in double quotes and separated by semicolons.

Model Inference

At the inference phase, input examples for RuGPT3 had the same structure as at the training phase, but included only reading text section, question text and right answer. For those models we generated text until our maximum defined length was reached. After that we split the predicted continuation of input by line breaks. After that, we filtered out distractors that were either non-unique or identical to the right answer. Then we sorted the retrieved set of distractors by alphabet and kept the first 3 results. For RuT5 we used maximum length as well as end-of-sequence token as stopping points for generation, while retrieving the distractors by splitting the output by semicolons and removing the enclosing quotes.

Assessment

BLEU and METEOR metric values were used for automatic assessment of the generated distractors. The implementations of BLEU and METEOR accessible from Evaluate³ Python package were used. In order to make our results comparable with the previous and forthcoming works we have also included ROUGE-L metric in our evaluation. As the official implementation of ROUGE, accessible from Evaluate package, cannot process Russian-language data, we have used an unofficial implementation of it⁴. However, the authors of this implementation admit that the values obtained from it may differ from the official variant. As most of the previous works (Chung et al, 2020; Qiu et al, 2020; Belyanova et al, 2022; Wang et al, 2023; Maity et al, 2024) utilise the 4-gram version of BLEU, this was the BLEU configuration used by default in our paper. We have also used BERTScore (Zhang et al, 2023) for semantic assessment of the generated distractors. BERTScore is a metric based on similarities of word embeddings from the BERT model instead of exact word/n-

gram matches. To enhance the convenience of interpreting the results, all metric values (defined from 0 to 1) were presented as percentages, ranging from 0 to 100.

Baselines

As, during the work on this paper, we have curated USE-TF dataset and our original modifications of RACE dataset, we tested baseline models on our data instead of just reporting scores from previous works. This was done in order to allow for fair comparison, as metrics used during our evaluation cannot be directly compared across different languages. The implementation of baseline models is described in this section, whereas their results are reported and analysed in comparison with implemented models in “Results” and “Discussion” sections.

As a first baseline in our experiments, we have used a non-fine-tuned version of RuGPT3. Along with a zero-shot RuGPT3, we have also used BART-DG models, enhanced versions of models introduced by Chung et al. (2020), that hold the state-of-the-art results in terms of BLEU in distractor generation on RACE data. In order to produce Russian outputs from these models, the same translation pipeline that was used in compiling of Ru-RACE was used to translate their outputs to Russian. USE inputs were translated to English using the same multilingual translation model (Opus-MT) before feeding them to BART-DG models.

RESULTS

Ru-RACE

Table 1 illustrates results of models trained on translated RACE subsets. In both tasks the best performance was

Table 1
Results of Models on Translated RACE Subsets

	BLEU-4		METEOR		ROUGE-L		BERTScore	
	dev	test	dev	test	dev	test	dev	test
Ru-RACE-TITLE								
RuGPT3-RACE-TITLE	3.83	3.19	12.78	12.41	12.32	12.60	68.72	68.68
RuT5-RACE-TITLE	25.17	22.96	46.09	45.35	16.79	16.21	79.09	78.72
Baseline RuGPT3	0.46	0.53	5.37	5.57	4.31	4.47	62.72	62.46
Ru-RACE-TF								
RuGPT3-RACE-TF	8.75	4.89	18.92	16.84	16.16	13.80	71.01	70.23
RuT5-RACE-TF	26.36	22.43	44.84	42.75	28.36	25.30	77.07	76.24
Baseline RuGPT3	1.23	1.73	9.54	9.44	8.30	8.29	63.64	64.04

³ <https://pypi.org/project/evaluate/>
⁴ <https://github.com/pltrdy/rouge>

demonstrated by T5-based models (RuT5-RACE-TITLE and RuT5-RACE-TF). This can be attributed to the Sequence-to-Sequence nature of T5 that allows it to transform inputs to outputs that have slightly different structure. Both models have surpassed the baseline established by non-fine-tuned RuGPT3 in both tasks, so we can conclude that fine-tuning allowed them to successfully adapt to the structure of our translated datasets.

For Ru-RACE-TITLE, the highest quality in terms of BLEU, METEOR and BERTScore on both dev and test subsets was reached by fine-tuned RuT5 model. Fine-tuned RuGPT3 demonstrated definitely lower results, with BLEU-4 reaching only 3.83 and 3.19 for dev and test subsets correspondingly (compared to 25.17 and 22.96 of RuT5-RACE-TITLE). Even higher absolute difference can be spotted in METEOR, with 12.78/12.41 on dev/test sets for RuGPT3 against 46.09/45.35 for RuT5. These differences indicate that RuT5-RACE-TITLE model greatly surpasses RuGPT3-RACE-TITLE both in terms of precision and recall. As for BERTScore, the difference between the two fine-tuned models on test set (10.04) is higher than the difference between the least scoring model and the baseline (4.24), which indicates that the RuT5-RACE-TITLE's title ability to produce semantically coherent distractors highly surpasses that of RuGPT3-RACE-TITLE. In terms of ROUGE-L values, the two fine-tuned models are not so far apart (4.47/3.61 on dev/test subsets) but both of them greatly surpass the baseline. The performance of all models on dev and test subsets is quite close, which proves that the models were not overfitted on the validation sets during hyper-parameter tuning.

For Ru-RACE-TF, the highest quality in terms of BLEU-4/METEOR/BERTScore was also reached by fine-tuned RuT5 model and the scores of fine-tuned RuGPT3 were also substantially lower (by 17.54/25.91/6.01 points on test set correspondingly). BLEU-4 and METEOR scores of RuGPT3-RACE-TITLE lie closer to the baseline than to the values of RuT5-RACE-TITLE. However, BERTScore differences between the baseline and the second-scoring model and between the second-scoring model and the baseline are quite close (7.37/6.19 against 6.06/6.01 on dev/test subsets), which suggests that the gap in semantic coherence between the two fine-tuned models might be not so broad. The difference in ROUGE-L appears to be equally broad both between the second-scoring model and the baseline and between the first- and second-scoring models. The difference in distractor generation quality between development and test subsets for Ru-RACE-TF is not enough for an overfit to be spotted. It can be seen that all model scores in this task are higher than in Ru-RACE-TITLE.

MuSeRC

Table 2 illustrates results of models trained on MuSeRC dataset. Due to MuSeRC test subset not being available at the dataset's developer website, all evaluation was performed on the development set. Both models beat the attested

zero-shot GPT3 baseline in all of the three metrics. BLEU-4 score of RuT5-MuSeRC-DG is nearly twice as better as that of RuGPT3-MuSeRC-DG (23.62 against 12.48), while METEOR score of RuT5-MuSeRC-DG is only slightly higher (45.78 against 40.87). According to BERTScore values (76.02 and 76.02 for RuT5-MuSeRC-DG and RuGPT3-MuSeRC-DG correspondingly), distractors generated by the two models are nearly equally semantically similar to the gold standard distractors. In terms of ROUGE-L the values of two fine-tuned models are quite close and both greatly surpass the zero-shot baseline. As both fine-tuned models produce results that beat the non-fine-tuned baseline, they were used for evaluation on USE data.

USE-TF

Table 3 illustrates results on the USE-TF dataset. The highest values of metrics were reached by RuT5-RACE-TF model, with BART-DG-PM model holding the second place. However, we can see that the scores of translated BART-DG-PM outputs (11.02/28.47/70.90 in terms of BLEU-4/METEOR/BERTScore) are quite close to scores of our best model (11.64/29.61/71.06 correspondingly).

We can see that RuT5-RACE-TF model displays robustness when dealing with data of USE, as its BLEU, METEOR and BERTScore values still greatly exceed the unsupervised baseline. However, this is not true for other models trained on Russian data, as their metric values degrade closer to baseline values attested by a zero-shot RuGPT3. The BLEU scores of all models, excluding RuT5-RACE-TF and BART-DG-based, rapidly decrease to zero with increasing the rank of BLEU, which indicates the lack of robustness in these models. This is especially true for models trained on MuSeRC, which means that existing Russian datasets cannot offer data that is suitable for distractor generation for complex general reading comprehension tasks, which can be found in language exam materials. We can conclude that training on the translated dataset can offer robustness of results while training on existing Russian-language dataset cannot. This can be attributed to MuSeRC dataset containing more trivial texts than USE, as MuSeRC consists mostly of news reports, while USE texts are usually extracts from high school-level literary works posing ethical problems worth of discussion. Due to the RuT5-RACE-TF being our only robust model in comparison, it is planned to use only its predictions in the future manual evaluation of data produced on the basis of USE-TF.

DISCUSSION

The results (Tables 1-3) demonstrate that the performance on translated datasets in distractor generation task is on par with the existing works, with BLEU-4 values reaching a maximum of around 25 for test subsets of datasets they were originally fine-tuned on. RuGPT3-based fine-tuned models

Table 2*Results of Models on MuSeRC Dataset*

	BLEU-4	METEOR	ROUGE-L	BERTScore
RuGPT3-MuSeRC-DG	12.48	40.87	21.77	76.04
RuT5-MuSeRC-DG	23.62	45.78	25.97	76.02
Baseline RuGPT3	5.16	11.25	6.81	62.91

Table 3*Results of Models on USE-TF Dataset*

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	BERTScore
RuGPT3-RACE-TF	15.11	3.57	0.08	0.00	9.22	6.83	65.84
RuT5-RACE-TF	29.11	20.93	15.66	11.64	29.61	13.55	71.06
RuGPT3-MuSeRC	9.56	1.69	0.48	0.00	6.65	4.35	61.62
RuT5-MuSeRC	10.30	2.20	0.55	0.00	7.77	4.66	62.63
Baseline RuGPT3	11.22	2.00	0.53	0.00	7.57	4.74	55.72
BART-DG	26.66	19.44	14.64	10.78	27.77	12.57	70.83
BART-DG-PM	28.52	20.24	15.06	11.02	28.47	12.40	70.90
BART-DG-ANPM	27.39	19.71	14.64	10.71	27.78	11.75	70.62

have shown generally better performance at producing consistent distractor outputs than RuT5-based, which can be attributed to Sequence-to-Sequence RuT5 being pre-trained on a text reconstruction rather than text generation task and thus being more prone to fine-tuning. From our results we can see that only RuT5 was able to produce coherent outputs both on its original dataset and on the independent dataset of USE questions, while our other fine-tuned models were able to do so only on the data from test subset of the dataset they were fine-tuned on. In this section we will explain the relationships between scores of different models on different datasets and propose ways to improve our results, while comparing our findings with that from the previous works on the subject.

The inability of models trained on MuSeRC to produce coherent distractor outputs for USE can be explained by the nature of MuSeRC dataset and relatively lower complexity of its items compared to real-word reading comprehension examination tasks. The fact of BART-DG (enhanced versions of models from Chung et al., 2020) holding a strong baseline against our results on USE data can be explained by the complex structure of BDG models, that features additional engineering techniques applied to base model. These techniques include entropy maximization-based decoding on different generation paths to produce multiple distractors independently (while our models produce them consecutively), parallel multi-task training (PM) and answer-negative (AN) regularisation.

The ability of T5-based models to be successfully fine-tuned on distractors from RACE questions compared to models of different architectures, found in our study, is supported by works of Hadifar et al. (2022), Ghanem & Fyshe (2024) & De-Fitero-Dominguez et al. (2024). The gap in generation quality metrics between decoder-only and encoder-decoder models is also found in Ghanem & Fyshe (2024), where fine-tuned T5-base models significantly outperforms fine-tuned GPT2-small (8.4 and 13.7 in terms of BLEU-2 for joint and disjoint T5 correspondingly against 3.9 for GPT2).

The results of comparison against BDG on the same set of data are contrastive to those of Ghanem & Fyshe (2024), who found that their best model outperforms BDG only in terms of BLEU-1 (32.0 against 30.2). However, authors of the referenced work fine-tuned BDG model on their data, whereas in our work we use readily fine-tuned models. Also, they implemented disjoint generation of distractors (as was also done by Chung et al., 2020) and succeeded to achieve a performance gain by it, whereas in our work only joint approach is implemented.

Better generation quality on the translated dataset than on an original-language one, found in our study, was also encountered by De-Fitero-Dominguez et al. (2024). However, it may be caused by reduction of lexical space imposed by the lexical knowledge of the translation model. However, the results on translated data have not always been especially high. For example, in Makhnytkina et al. (2020) question

generation model demonstrated poor performance on test subset of dataset it was fine-tuned on in terms of formal metrics. Nevertheless, these results can be explained by the use of older model architecture (BiLSTM) and the lesser development of translation models in 2020.

The results of our experiments, where METEOR score in all settings is higher than ROUGE-L, contradict the findings of Hadifar et al. (2022), where METEOR score was much lower than ROUGE-L. Taking into account differences in metrics calculation (much higher weight of recall in METEOR than in ROUGE and use of longest common subsequence length in ROUGE-L instead of unigram match in METEOR), it can be deducted that while their solution better captures patterns from distractor data than preserves the lexical content, the opposite is true for our solutions.

Taking into account underperformance of GPT3-based models in our experiments, it is worth noting that we have used “small” version of RuGPT3 for fine-tuning due to the lack of computational resources needed to fine-tune larger versions of RuGPT3. Considering the closeness of our results on independent USE-TF dataset to BART-DG models, it is worth noting that they implemented disjoint decoding of output distractors, while our solutions use joint decoding. Taking into account the advantage in metric values of disjoint generation over joint, found Ghanem & Fyshe (2024), it is possible that our models will outperform BART-DG by higher value if we implement disjoint decoding.

Our expectations about the results, described in the “Introduction” section (efficiency of fine-tuning on translated high-quality dataset and the advantage of fine-tuning on specific type of questions), are met by performance of RuT5-RACE-TF model on USE-TF data, which beats both baselines and performance of models trained on Russian-language dataset containing questions of different types. However, these expectations are not met by RuGPT3-RACE-TF model, that surpasses results of MuSeRC-trained models, but not the baselines attested by models trained on full RACE-DG dataset.

It is worth noting that the reported formal quantitative metrics are based on the similarity between the generated and original distractors and the overall plausibility of distractors should be estimated by human evaluation, which is described in the “Future Work” subsection. While our models utilise base implementations of T5 and GPT3, further engineering enhancements can be applied to them.

Future Work

For future manual evaluation of our models fine-tuned on Ru-RACE-TF, it is planned to use USE-TF data (as a professionally curated set of originally Russian multiple-choice questions) and include predictions from RuT5-RACE-TF and BART-DG-PM models. It is planned to equip each question

of USE-TF with 4 answer options – the original right answer, a “filler” distractor, one of the original distractors, one prediction from our fine-tuned model and one prediction from BART-DG-PM. A filler distractor can be a sentence extracted from an existing Russian corpus, that is semantically close (as attested by a formal metrics, such as BERTScore) to the reading text passage. The plan is to attract Russian-speaking participants with higher education, so the test takers will be able to actually distinguish distractors from the real right answers. The participants will be asked to rate each of the examples on a scale from 1 to 5, with 1 indicating the most unsuitable option and 5 indicating the option most likely to be the right answer. The hypothesis is that the distractors from our models will be on average rated higher than filler distractors but lower than the original right answers.

For manual evaluation of our best model trained on Ru-RACE-TITLE (RuT5-RACE-TITLE) it is planned to use arbitrary Russian texts as inputs for the distractor generation models. This may include extracts from newspaper sources and stories for children, as most of the texts from RACE are of narrative nature. The design of questions will be the same as for Ru-RACE-TF evaluation set, the original right answer will be implemented as the article’s original title or created manually. The same findings about relationships between average rank values as from USE-TF dataset are expected hypothetically. As metric values for fine-tuned models on the Ru-RACE-TF were higher than on Ru-RACE-TITLE, it is expected that during proposed manual evaluation average ranks of distractors for Ru-RACE-TF will also be higher than ranks for Ru-RACE-TITLE.

While accessing the overall plausibility of generated options, this method can also help determine how often trivial distractors are generated, as the trivial options are expected to be averagely rated on par with “filler” distractors. Also, it may be beneficial annotate a set of model distractor predictions in terms of plausibility and triviality. The annotation of distractor characteristics gained through this procedure can be used for future training of a distractor assessment ML model. Although triviality is not usually tackled specifically in recent works, as modern generative NNs are able to extract patterns from the presented non-structured data without the need for additional engineering-based output conditioning, this model can be used during future training of a new distractor generation pipeline by penalising outputs that would be predicted to be too trivial and rewarding outputs which prediction of plausibility would be high. Additional enhancements may include disjoint output decoding, the use of larger model variants and the implementation of prompting approach as an alternative.

CONCLUSION

In this paper automatic distractor generation was implemented for Russian data. 6 large language models of two

types (GPT-3 and T5) were fine-tuned on distractor generation tasks on 3 datasets – 2 machine-translated English-to-Russian question datasets containing only specific types of questions (title selection and true/untrue fact selection) and an originally Russian dataset. RuT5-based models demonstrated generally better results than RuGPT3-based. Both model types surpassed the unsupervised baseline attested by non-fine-tuned RuGPT3 model, proving the possibility of effective fine-tuning on distractor generation task on English-to-Russian translated data addressed in RQ1.

During the experimentation on Russian examination data, it was found out that translated English reading comprehension examination dataset is more efficient in terms of use in model fine-tuning than an existing Russian non-examination reading comprehension dataset, as the models trained on the latter dataset demonstrated poor performance compared to models trained on the former. This highlights the importance of domain and complexity level of questions in distractor generation task and proves the need of a comprehensive Russian exam-style multiple-choice reading comprehension question dataset addressed in RQ2. T5-based model fine-tuned on true statement selection distractors demonstrated better performance on USE data than both MuSeRC-trained models and the state-of-the-art exam-style distractor generation solution, demonstrating the advantage of fine-tuning on a specific type of questions, which possibility was addressed in RQ3.

The value of work lies in training distractor generation models for Russian-language data, which has not been done in the previous works. The explored possibility of transferring intelligence learned on specific categories of questions, found in large-scale datasets, to distractor generation for exam questions of a specific standard presents an additional value. Our findings can be beneficial for exam preparation platform creators, who can include models trained in

the described settings into their products, allowing for the automatic item bank replenishment.

Our findings can also be helpful for reading comprehension dataset creators. The annotation of distinct common question types can be implemented by them in dataset design. The lack of exam-style Russian question datasets needed for successful model learning can present an opportunity for them.

The future work in this direction should include manual evaluation of the generated data and the development of a Russian reading comprehension dataset designed specifically for examination. Another direction of future work lies in exploring the possibilities of different alternative generation techniques not covered in described experiments. A detailed comparison of Russian-language distractors obtained from models trained on distractor generation task and retrieved from chat- and instruct-based large language models via prompting can present another interesting direction for future research.

ACKNOWLEDGMENTS

The author wants to thank his academic advisor Prof. Olga Lyashevskaya for providing beneficial advice during the preparation of revised manuscript. This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

DECLARATION OF COMPETING INTEREST

None declared.

REFERENCES

- Alsubait, T. M. (2015). *Ontology-based multiple-choice question generation* [Unpublished PhD thesis]. University of Manchester.
- Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, & C. Voss (Eds.), *Proceedings of the ACL Workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72). Association for Computational Linguistics.
- Belyanova, M. A., Andreev, A. M., & Gapanyuk, Y. E. (2022). Neural text question generation for Russian language using hybrid intelligent information systems approach. In B. Kryzhanovskiy, W. Dunin-Barkowski, V. Redko, Y. Tiumentsev, & V. V. Klimov (Eds.), *Advances in neural computation, machine learning, and cognitive research V* (vol. 1008, pp. 217–223). Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-91581-0_29
- Bitew, S. K., Hadifar, A., Sterckx, L., Deleu, J., Develder, C., Demeester, T. (2022) Learning to reuse distractors to support multiple choice question generation in education. *IEEE Transactions on Learning Technologies*, 17, 375–390. IEEE Computer Society Press. <http://dx.doi.org/10.1109/TLT.2022.3226523>
- Bitew, S. K., Deleu, J., Develder, C., & Demeester, T. (2023) *Distractor generation for multiple-choice questions with predictive prompting and large language models* (Version 1). arXiv. <http://dx.doi.org/10.48550/arXiv.2307.16338>

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (vol. 33, pp. 1877–1901). Curran Associates, Inc. <http://dx.doi.org/10.48550/arXiv.2005.14165>
- Chung, H.-L., Chan, Y.-H., & Fan, Y.-C. (2020). A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4390–4400). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2020.findings-emnlp.393>
- De-Fitero-Dominguez, D., Garcia-Lopez, E., Garcia-Cabot, A., Del-Hoyo-Gabaldon, J.-A., & Moreno-Cediel, A. (2024). Distractor generation through text-to-text transformer models. *IEEE Access*, 12, 25580–25589. <http://dx.doi.org/10.1109/ACCESS.2024.3361673>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (vol. 1: Long and Short Paper, pp. 4171–4186). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/N19-1423>
- Efimov, P., Chertok, A., Boytsov, L., & Braslavski, P. (2020). SberQuAD – Russian reading comprehension dataset: Description and analysis. In A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névél, L. Cappellato, & N. Ferro (Eds.), *Experimental IR meets multilinguality, multimodality, and interaction* (vol. 12260, pp. 3–15). Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-58219-7_1
- Elkins, S., Kochmar, E., Serban, I., & Cheung, J. C. K. (2023). How useful are educational questions generated by large language models? In N. Wang, G. Rebollo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (vol. 1831, pp. 536–542). Springer Nature Switzerland. http://dx.doi.org/10.1007/978-3-031-36336-8_83
- Fenogenova, A., Mikhailov, V., & Shevelev, D. (2020). Read and reason with MuSeRC and RuCoS: Datasets for machine reading comprehension for Russian. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6481–6497). International Committee on Computational Linguistics. <http://dx.doi.org/10.18653/v1/2020.coling-main.570>
- Gao, Y., Bing, L., Li, P., King, I., & Lyu, M. R. (2019). Generating distractors for reading comprehension questions from real examinations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 6423–6430. <http://dx.doi.org/10.1609/aaai.v33i01.33016423>
- Ghanem, B. & Fyshe, A. (2024). DISTO: Textual distractors for multiple choice reading comprehension questions using negative sampling. In M. Marras, M. Ueno (Eds.), *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 23–34). International Educational Data Mining Society. <http://dx.doi.org/10.5281/ZENODO.12729766>
- Glushkova, T., Machnev, A., Fenogenova, A., Shavrina, T., Artemova, E., & Ignatov, D. I. (2021). DaNetQA: A yes/no question answering dataset for the Russian language. In W. M. P. Van Der Aalst, V. Batagelj, D. I. Ignatov, M. Khachay, O. Koltsova, A. Kutuzov, S. O. Kuznetsov, I. A. Lomazova, N. Loukachevitch, A. Napoli, A. Panchenko, P. M. Pardalos, M. Pelillo, A. V. Savchenko, & E. Tutubalina (Eds.), *Analysis of Images, Social Networks and Texts* (vol. 12602, pp. 57–68). Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-72610-2_4
- Hadifar, A., Bitew, S. K., Deleu, J., Develder, C., & Demeester, T. (2023). EduQG: A multi-format multiple-choice dataset for the educational domain. *IEEE Access*, 11, 20885–20896. <http://dx.doi.org/10.1109/ACCESS.2023.3248790>
- Huang, L., Le Bras, R., Bhagavatula, C., & Choi, Y. (2019). CosmosQA: Machine reading comprehension with contextual commonsense reasoning. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2391–2401). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/D19-1243>
- Joshi, M., Choi, E., Weld, D., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In R. Barzilay, & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational linguistics* (vol. 1: Long Papers, pp. 1601–1611). Association for Computational linguistics. <http://dx.doi.org/10.18653/v1/P17-1147>
- Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1), 121–204. <http://dx.doi.org/10.1007/s40593-019-00186-y>
- Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7, 453–466. http://dx.doi.org/10.1162/tacl_a_00276

- Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale reading comprehension dataset from examinations. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 785–794). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/D17-1082>
- Lee, D. B., Lee, S., Jeong, W. T., Kim, D., & Hwang, S. J. (2020). Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 208–224). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2020.acl-main.20>
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2020.acl-main.703>
- Lin, C.-Y. (2004). ROUGE: A Package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Association for Computational Linguistics. <https://aclanthology.org/W04-1013>
- Lu, X., West, P., Zellers, R., Bras, R. L., Bhagavatula, C., & Choi, Y. (2021). NeuroLogic decoding: (Un)supervised neural text generation with predicate logic constraints. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 4288–4299). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2021.naacl-main.339>
- Maity, S., Deroy, A., & Sarkar, S. (2024). A novel multi-stage prompting approach for language agnostic MCQ generation using GPT. In N. Goharian, N. Tonello, Y. He, A. Lipani, G. McDonald, C. Macdonald, & I. Ounis (Eds.), *Advances in information retrieval* (vol. 14610, pp. 268–277). Springer Nature Switzerland. http://dx.doi.org/10.1007/978-3-031-56063-7_18
- Makhnytkina, O., Matveev, A., Svischev, A., Korobova, P., Zubok, D., Mamaev, N., & Tchirkovskii, A. (2020). Conversational question generation in Russian. In S. Balandin, L. Turchet, & T. Tyutina (Eds.), *2020 27th Conference of Open Innovations Association (FRUCT)* (pp. 1–8). IEEE. <http://dx.doi.org/10.23919/FRUCT49677.2020.9211056>
- Manakul, P., Liusie, A., & Gales, M. (2023). MQAG: Multiple-choice question answering and generation for assessing information consistency in summarization. In J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, & A. A. Krisnadhi (Eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific chapter of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 39–53). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2023.ijcnlp-main.4>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *The 40th Annual Meeting on Association for Computational Linguistics—ACL '02* (pp. 311–318). Association for Computational Linguistics. <http://dx.doi.org/10.3115/1073083.1073135>
- Paris, A. H., & Paris, S. G. (2003). Assessing narrative comprehension in young children. *Reading Research Quarterly*, 38(1), 36–76. <http://dx.doi.org/10.1598/RRQ.38.1.3>
- Qiu, Z., Wu, X., & Fan, W. (2020). Automatic distractor generation for multiple choice questions in standard tests. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 2096–2106). International Committee on Computational Linguistics. <http://dx.doi.org/10.18653/v1/2020.coling-main.189>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21, 1, 5485–5551. <https://dl.acm.org/doi/abs/10.5555/3455716.3455856>
- Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383–2392). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/D16-1264>
- Reddy, S., Chen, D., & Manning, C. D. (2019). CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7, 249–266. http://dx.doi.org/10.1162/tacl_a_00266
- Rybin, I., Korablinov, V., Efimov, P., & Braslavski, P. (2021). RuBQ 2.0: An innovated Russian question answering dataset. In R. Verborgh, K. Hose, H. Paulheim, P.-A. Champin, M. Maleshkova, O. Corcho, P. Ristoski, & M. Alam (Eds.), *The Semantic Web* (vol. 12731, pp. 532–547). Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-77385-4_32
- Sekulić, I., Aliannejadi, M., & Crestani, F. (2021). Towards facet-driven generation of clarifying questions for conversational search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 167–175). Association for Computing Machinery. <http://dx.doi.org/10.1145/3471158.3472257>
- Shavrina, T., Emelyanov, A., Fenogenova, A., Fomin, V., Mikhailov, V., Evlampiev, A., Malykh, V., Larin, V., Natekin, A., Vatulin, A., Romov, P., Anastasiev, D., Zinov, N., & Chertok, A. (2020, May). Humans keep it one hundred: An overview of AI Journey. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A.

- Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 2276–2284). European Language Resources Association. <https://aclanthology.org/2020.lrec-1.277/>
- Tiedemann, J., & Thottingal, S. (2020). OPUS-MT – Building open translation services for the world. In A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, & M. L. Forcada (Eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 479–480). European Association for Machine Translation. <https://aclanthology.org/2020.eamt-1.61/>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (vol. 30, 6000–6010). Curran Associates, Inc. <https://dl.acm.org/doi/10.5555/3295222.3295349>
- Welbl, J., Liu, N. F., & Gardner, M. (2017). Crowdsourcing multiple choice science questions. In L. Derczynski, W. Xu, A. Ritter, & T. Baldwin (Eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text* (pp. 94–106). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/W17-4413>
- Xiao, D., Zhang, H., Li, Y., Sun, Y., Tian, H., Wu, H., & Wang, H. (2020). ERNIE-GEN: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In C. Bessiere (Ed.) *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (pp. 3997–4003). International Joint Conferences on Artificial Intelligence Organization. <http://dx.doi.org/10.24963/ijcai.2020/553>
- Xu, Y., Wang, D., Yu, M., Ritchie, D., Yao, B., Wu, T., Zhang, Z., Li, T., Bradford, N., Sun, B., Hoang, T., Sang, Y., Hou, Y., Ma, X., Yang, D., Peng, N., Yu, Z., & Warschauer, M. (2022). Fantastic questions and where to find them: FairytaleQA – An authentic dataset for narrative comprehension. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 447–460). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2022.acl-long.34>
- Xue, L., Constant, N., Roberts, A., Kale, N., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2020). MT5: A massively multilingual pre-trained text-to-text transformer (Version 3). arXiv. <http://dx.doi.org/10.48550/arXiv.2010.11934>
- Zhang, C. (2023). *Automatic generation of multiple-choice questions* (Version 1). arXiv. <http://dx.doi.org/10.48550/ARXIV.2303.14576>
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT (Version 3). arXiv. <http://dx.doi.org/10.48550/ARXIV.1904.09675>
- Zmitrovich, D., Abramov, A., Kalmykov, A., Tikhonova, M., Taktasheva, E., Astafurov, D., Baushenko, M., Snegirev, A., Kadulin, V., Markov, S., Shavrina, T., Mikhailov, V., & Fenogenova, A. (2024). A family of pretrained transformer language models for Russian. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 507–524). ELRA Language Resource Association. <http://dx.doi.org/10.48550/arXiv.2309.10931>

APPENDIX A

Example of model generation

To better illustrate the performance of our best model, RuT5-RACE-TF, we will analyse an example of distractors generated by our best model, RuT5-RACE-TF, on an arbitrary Russian text – “The Shark”, a children’s short story by Leo Tolstoy, taken from Russian WikiSource. The story tells us about how a ship’s cannoneer saved two boys who went swimming in the open sea from a shark by shooting it with a cannon. The right answer to the “Which statement is TRUE according to the text?” question (“Old cannoneer’s marksmanship saved the boys from the sea monster”) was created manually.

From the Figure 1 it can be seen that the options 2 and 3 can indeed work as distractors as they contradict the text as they state facts that are not present in the story (“Old cannoneer took the boy aside”; “The boys who were in the boat did not hear the old cannoneer’s cry”). However, option 4 (“The story happened the day we saw a shark”) is, although being very trivial, true to the text and therefore cannot serve as a distractor in this context. It is worth noting that in this example the language of the generated distractors is consistent and does not break the rules of Russian grammar.

Figure 1

Result of Ru-T5-RACE distractor generation on a Leo Tolstoy's children's story

Наш корабль стоял на якоре у берега Африки. День был прекрасный, с моря дул свежий ветер; но к вечеру погода изменилась: стало душно и точно из топлённой печки несло на нас горячим воздухом с пустыни Сахары. Перед закатом солнца капитан вышел на палубу, крикнул: «Купаться!» — и в одну минуту матросы попрыгали в воду, спустили в воду парус, привязали его и в парусе устроили купальню. На корабле с нами было два мальчика. Мальчики первые попрыгали в воду, но им тесно было в парусе, они вздумали плавать наперегонки в открытом море. Оба, как ящерицы, вытягивались в воде и что было силы поплыли к тому месту, где был бочонок над якорем. Один мальчик сначала перегнал товарища, но потом стал отставать. Отец мальчика, старый артиллерист, стоял на палубе и любовался на своего сынишку. Когда сын стал отставать, отец крикнул ему: «Не выдавай! Понатужься!» Вдруг с палубы кто-то крикнул: «Акула!» — и все мы увидели в воде спину морского чудовища. Акула плыла прямо на мальчиков. — Назад! Назад! Вернитесь! Акула! — закричал артиллерист. Но ребята не слышали его, плыли дальше, смеялись и кричали еще веселее и громче прежнего. Артиллерист, бледный как полотно, не шевелясь, смотрел на детей. Матросы спустили лодку, бросились в нее и, сгибая весла, понеслись что было силы к мальчикам; но они были еще далеко от них, когда акула уже была не дальше 20-ти шагов. Мальчики сначала не слышали того, что им кричали, и не видели акулы; но потом один из них оглянулся, и мы все услышали пронзительный визг, и мальчики поплыли в разные стороны. Визг этот как будто разбудил артиллериста. Он сорвался с места и побежал к пушке. Он повернул хобот, прилег к пушке, прицелился и взял фитиль. Мы все, сколько нас ни было на корабле, замерли от страха и ждали, что будет. Раздался выстрел, и мы увидели, что артиллерист упал подле пушки и закрыл лицо руками. Что случилось с акулой и с мальчиками, мы не видели, потому что на минуту дым застал нам глаза. Но когда дым разошелся над водою, со всех сторон послышался сначала тихий ропот, потом ропот этот стал сильнее, и, наконец, со всех сторон раздался громкий, радостный крик. Старый артиллерист открыл лицо, поднялся и посмотрел на море. По волнам колыхалось желтое брюхо мертвой акулы. В несколько минут лодка подплыла к мальчикам и привезла их на корабль.

Какое высказывание СООТВЕТСТВУЕТ тексту?

1. **Меткость старого артиллериста спасла мальчиков от морского чудовища**
2. Старый артиллерист отвел мальчика в сторону
3. Мальчики, которые были в лодке, не слышали крика старого артиллериста
4. История произошла в день, когда мы увидели акулу

Note. Manually crafted right answer is given in bold.

APPENDIX B

The source code and data files for this paper are available at the online repository: <https://github.com/nicklogin/Ru-RC-DG>

<https://doi.org/10.17323/jle.2024.22237>

Automatic Morpheme Segmentation for Russian: Can an Algorithm Replace Experts?

Dmitry Morozov ¹, Timur Garipov ¹, Olga Lyashevskaya ^{2,3}, Svetlana Savchuk ³, Boris Iomdin ⁴, Anna Glazkova ⁵

¹ Novosibirsk State University, Novosibirsk, Russia

² HSE University, Moscow, Russia

³ Vinogradov Russian Language Institute, Russian Academy of Sciences, Moscow, Russia

⁴ independent researcher

⁵ University of Tyumen, Tyumen, Russia

ABSTRACT

Introduction: Numerous algorithms have been proposed for the task of automatic morpheme segmentation of Russian words. Due to the differences in task formulation and datasets utilized, comparing the quality of these algorithms is challenging. It is unclear whether the errors in the models are due to the ineffectiveness of algorithms themselves or to errors and inconsistencies in the morpheme dictionaries. Thus, it remains uncertain whether any algorithm can be used to automatically expand the existing morpheme dictionaries.

Purpose: To compare various existing algorithms of morpheme segmentation for the Russian language and analyze their applicability in the task of automatic augmentation of various existing morpheme dictionaries.

Results: In this study, we compared several state-of-the-art machine learning algorithms using three datasets structured around different segmentation paradigms. Two experiments were carried out, each employing five-fold cross-validation. In the first experiment, we randomly partitioned the dataset into five subsets. In the second, we grouped all words sharing the same root into a single subset, excluding words that contained multiple roots. During cross-validation, models were trained on four of these subsets and evaluated on the remaining one. Across both experiments, the algorithms that relied on ensembles of convolutional neural networks consistently demonstrated the highest performance. However, we observed a notable decline in accuracy when testing on words containing unfamiliar roots. We also found that, on a randomly selected set of words, the performance of these algorithms was comparable to that of human experts.

Conclusion: Our results indicate that although automatic methods have, on average, reached a quality close to expert level, the lack of semantic consideration makes it impossible to use them for automatic dictionary expansion without expert validation. The conducted research revealed that further research should be aimed at addressing the key identified issues: poor performance with unknown roots and acronyms. At the same time, when a small number of unfamiliar roots can be assumed in the test dataset, an ensemble of convolutional neural networks should be utilized. The presented results can be used in the development of morpheme-oriented tokenizers and systems for analyzing the complexity of texts.

KEYWORDS

automatic morpheme segmentation, Russian language morphology, machine learning, convolutional neural networks, dictionary expansion, morphological analysis, natural language processing, expert-level performance

Citation: Morozov, D., Garipov, T., Lyashevskaya, O., Savchuk, S., Iomdin, B., & Glazkova, A. (2024). Automatic morpheme segmentation for Russian: Can an algorithm replace experts? *Journal of Language and Education*, 10(4), 71-84. <https://doi.org/10.17323/jle.2024.22237>

Correspondence:
Dmitry Morozov,
morozowdm@gmail.com

Received: August 18, 2024

Accepted: December 16, 2024

Published: December 30, 2024



INTRODUCTION

Morpheme segmentation of a word is the process of breaking down the word into

its smallest meaningful units called morphemes, for example, prefixes, suffixes, and roots. Many spelling rules taught in school rely on the student's ability to

identify a morpheme or determine the relative position of several morphemes (Bakulina, 2012). In Russian, such rules include spelling of voiceless and sonorous consonants in prefixes, spelling of -н/-нн- at morpheme boundaries and within them, searching for cognates to determine which vowel to write in unstressed syllables, where several phonemes may be pronounced the same, etc.

Morpheme segmentation can also be used in developing tools for automatic language analysis, both in creating a feature-based description of text, for example, in text complexity assessment (Morozov et al., 2024), and in developing language models as an alternative to Byte-Pair Encoding (BPE) tokenizers, which can improve model quality (Matthews et al., 2018). However, the proportion of words not described in morpheme dictionaries is significant: in one of the largest such dictionaries of Russian, the “Word Formation Dictionary of Russian language” (Tikhonov, 1990), there are segmentations for about 150 thousand different lemmas, while in the Main Corpus of the Russian National Corpus (Savchuk et al., 2024), there are over 250 thousand unique lemmas. Therefore, developing algorithms for automatic analysis is an urgent task.

For the Russian language, morpheme segmentation is complicated by the lack of a unified approach to segmenting words into morphemes (Iomdin, 2019). Some authors use the so-called Vinokur criterion (Vinokur, 1946) as a guide for segmentation. In this case, to cut a morpheme from a word, it is necessary to present a word-forming chain, that is, to find a word that, when supplemented with this morpheme, coincides with the word under consideration, e.g. пис-а-ть ‘write’ + -тель = пис-а-тель ‘writer’. This approach, for instance, is adopted in the aforementioned “Word Formation Dictionary” and is often used within school education. A significant drawback of this approach is that words which are considered related by native speakers, may turn out to be unrelated morphologically. Thus, in the word неодобрительный ‘disapproving’, the root is claimed to be -одобр-, while in добро ‘good’ it is -добр-, meaning these words are not cognates.

Other researchers, like the authors of the “Dictionary of Russian Language Morphemes” (Kuznetsova & Yefremova, 1986), prefer a more granular approach to morphemes and rely on comparability of a word with other lexemes of similar structure. For example, the word улыбаться ‘to smile’ features the -лыб- root, as the structure is parallel to the other verbs with у- (cf. у-смех-а-ть-ся ‘to grin’), and some words are analyzed etymologically (на-сек-ом-ое ‘insect’, вос-точ-ный ‘eastern’). The borrowings are split into morphemes (eg. ре-волюц-и-я ‘revolution’, квит-анци-я ‘receipt’) if they have semantic parallels to other borrowings with a comparable structure (cf. э-волюц-и-я ‘evolution’, рас-квит-а-ть-ся ‘to get even’).

However, studying dictionaries reveals that in specific cases, authors make decisions that contradict the established paradigm, such as in the segmentation of suffixes, e.g. за-воева-тель-н-ый ‘aggressive’ vs за-град-и-тель-н-ый ‘barrage’ in (Tikhonov, 1990). Thus, the rules of morpheme segmentation represent a loosely formalized area, which likely makes it impossible to devise an absolutely error-free algorithm.

Nevertheless, since the task has sufficient practical potential, there are many automatic approximate approaches presented. One of the most commonly used and extensively described is a family of algorithms based on the Morfessor algorithm (Creutz & Lagus, 2002). This algorithm belongs to language-independent unsupervised and semi-supervised machine learning methods to be trained on a large text collection. Among the most relevant modifications of the original algorithm, it is worth mentioning the approach by S.-A. Grönroos et al. (2020), which explores the combination of Morfessor with EM+Prune. Significant progress in the quality of algorithms has been achieved during the SIG-MORPHON 2022 competition (Batsuren et al., 2022), where several approaches were presented that significantly outperformed the baselines including Morfessor, ULM (Kudo, 2018) and WordPiece (Schuster, & Nakajima, 2012). Among the proposed architectures are those based on Transformer models (Zundi & Avaajargal, 2022; Peters & Martins, 2022), GRU models (Levine, 2022), neural hard-attention transducer models (Wehrli et al, 2022), LSTM networks (Peters & Martins, 2022; Gırrbach, 2022), and Hidden Markov models (Bodnár, 2022). The team DeepSPIN (Peters & Martins, 2022) achieved the best quality across all nine languages involved. Their solutions are based on LSTM networks with a specific loss function (DeepSPIN-1 and DeepSPIN-2) and the Transformer architecture (DeepSPIN-3).

In the near future, a rapid increase in the number of approaches utilizing large language models is to be expected. Pranjić et al. (2024) proposed an algorithm based on the Glot500-m network (ImaniGooghari et al., 2023), representing a binary classifier for determining morpheme boundaries in a word. However, the limitations of the algorithm, namely, relatively low quality on the English, Finnish, and Turkish datasets, as well as extremely long processing time (as the algorithm checks each pair of neighboring letters in a word), do not currently allow this approach to be considered a priority.

For the Russian language, the most relevant solutions superior to the Morfessor algorithm are presented by Sorokin & Kravtsova (2018), Sapin & Bolshakova (2019a; 2019b). The authors introduce approaches based on convolutional neural networks, long short-term memory networks, and gradient boosting over decision trees. The results of comparing algorithms on two different datasets do not allow for a definitive conclusion regarding the superiority of one algorithm over the others. However, the quality they achieve (about 90% of completely correct segmentations) is quite

high. The Russian language was also among the languages at SIGMORPHON 2022, and the DeepSPIN-3 model achieved the best quality.

At the same time, a number of questions in this area remain insufficiently explored. Garipov et al (2023) found that a model based on convolutional neural networks has a significant drawback: its quality sharply decreases when tested on words containing roots that were absent in the training set, with the percentage of fully correct segmentations dropping by 17-18%. It remains unclear whether a similar issue exists for other algorithms demonstrating high quality.

Additionally, when developing a new algorithm or conducting a competition, typically only one morpheme dictionary per language is considered, whereas it makes sense to consider more dictionaries for a more representative study. Comparing the algorithms presented in various papers is further complicated by the fact that researchers are actually addressing different tasks. In some cases (Sorokin & Kravtsova, 2018; Bolshakova & Sapin, 2019a; Bolshakova & Sapin, 2019b), the task specifically focuses on segmenting the original string into morphemes, while in the SIGMORPHON competition, the task involved reconstructing “standardized” forms of morphemes. Cotterell et al. (2016) describes the difference between these approaches: the so-called “surface” segmentation is a sequence of surface substrings whose concatenation is exactly equal to the original word, e.g., *funniest* → *funn-i-est*, while during “canonical” segmentation, the task is not only computing surface segmentation but also restoring standardized forms of morphemes, e.g., *funniest* → *fun-y-est*.

The third issue is the impact of internal inconsistency among dictionaries on the quality of the algorithm. It is impossible to determine whether the quality of the models has already reached the expert level, and the remaining errors can be explained by the internal inconsistency in the training dataset.

Thus, the purpose of this research is to compare various existing algorithms of morpheme segmentation for the Russian language and analyze their applicability in the task of automatic augmentation of various existing morpheme dictionaries. We seek to answer the following research questions:

- RQ#1: Which of the presented algorithms achieve the best results for the Russian language based on various morpheme dictionaries annotated in different paradigms?
- RQ#2: How well can the presented algorithms parse words containing roots that were not encountered in the training data?
- RQ#3: How does the quality of annotation by algorithms compare to the quality of annotation by expert linguists?

METHOD

Datasets

In our study, we used morpheme dictionaries where each word is segmented into morphemes with the type of each one indicated. A total of seven morpheme types are used: PREF (prefix), ROOT (root), SUFF (suffix), END (ending), POST (postfix), LINK (linking vowel), and HYPH (hyphen). To ensure a high representativeness in the study, we utilized three morpheme segmentation datasets annotated in different paradigms:

- (1) Morphodict-K: dataset based on the “Dictionary of Morphemes of the Russian Language” (Kuznetsova & Yefremova, 1986), used in the Main Corpus of the Russian National Corpus. Rules of segmentation is that of strong albeit not maximal splitting of morphemes and correspondences to other words with similar structure.
- (2) Morphodict-T: dataset based on the “Word Formation Dictionary of Russian language” (Tikhonov, 1990). This dataset is used in the Educational Corpus of the Russian National Corpus. So-called Vinokur criterion is used as an algorithm for splitting words into morphemes. Morphemes in Morphodict-T are splitted in larger chunks than in Morphodict-K (*улыб-а-ть-ся* ‘to smile’, *насеком-ое* ‘insect’, *восточ-ный* ‘eastern’), especially borrowings (*революци-я* ‘revolution’, *квитанци-я* ‘receipt’). The vocabulary of the datasets also varies. For example, Morphodict-K dataset contains 75,649 words, of which only about 58,000 are present in the Morphodict-T one. Notably, Morphodict-T differs from the dataset utilized by Sorokin & Kravtsova (2018) in that it fixes many incorrect morpheme type annotations. Error detection and type correction were performed out by a team of three experts. A total of 31,468 segmentations were corrected. In cases of disagreement, the segmentations were discarded (27 cases in total).
- (3) CrossLexica (Bolshakov, 2013): dataset used in (Bolshakova & Sapin, 2019a; Bolshakova & Sapin, 2019b). The rules of morpheme segmentation for this dataset are not described explicitly; however, in this small dataset there are differences from both Morphodict-K and Morphodict-T (Table 1). In the CrossLexica dataset, unlike the other two, there are no words with multiple roots, but there are a number of non-lemmatized words.

A brief description of the datasets is provided in Table 2.

Importantly, within the scope of the study, it was assumed that a word is exactly equal to the concatenation of its morphemes, which is generally incorrect. For example, the word *горбунья* ‘female hunchback’ can be parsed as *горб:ROOT/ун:SUFF/ьj:SUFF/я:END* (Kuznetsova & Yefremova, 1986) with an additional *-j-*, which is not written as a separate letter. In such cases, we modified segmentation: the *-j-* was excluded.

Table 1
Examples of Markup Differences between Datasets

Word	Morphodict-K	Morphodict-T	CrossLexica
революция 'revolution'	ре:PREFIX/волюц:ROOT/и:SUFFIX/я:END	революци:ROOT/я:END	ре:PREFIX/вол:ROOT/юци:SUFFIX/я:END
утверждать 'to approve'	у:PREFIX/твержд:ROOT/а:SUFFIX/ть:END	утвержд:ROOT/а:SUFFIX/ть:END	у:PREFIX/твержд:ROOT/ать:END
собственник 'owner'	соб:ROOT/ств:SUFFIX/енн:SUFFIX/ик:SUFFIX	собственн:ROOT/ик:SUFFIX	соб:ROOT/ств:SUFFIX/ен:SUFFIX/ник:SUFFIX

Table 2
Some Characteristics of the Datasets Utilized

Characteristic	CrossLexica	Morphodict-T	Morphodict-K
Unique words	23426	95895	75649
Unique morphemes	2745	15899	8079
Unique roots	2256	15253	7148
Average morphemes per word	3.68	3.86	4.12
Average morpheme occurrence	25.14	23.29	38.56
Average root occurrence	8.31	7.54	12.24
Average root length	4.57	5.52	4.62

If after that -ь- became the only letter in the morpheme, we concatenated it to the previous morpheme. Therefore, in the considered case segmentation was simplified to `ропб:ROOT/унь:SUFFIX/я:END`.

Another important feature of our work is that all the datasets utilized contain exclusively lemmata. This limits the applicability of the models trained during the experiments; however, it allows us to avoid spending resources on dealing with homonymy, as the homonymy of lemmata with different morpheme segmentation is a relatively rare occurrence in the Russian language.

Algorithms

Algorithms with Morpheme-Type Labeling

Among the algorithms with morpheme-type annotation, we selected three that showed the best quality in previous experiments: the convolutional neural networks ensemble (hereinafter CNN) (Sorokin & Kravtsova, 2018), the gradient boosting algorithm over decision trees (hereinafter GBDT), and long short-term memory network (hereinafter LSTM). Comparing these algorithms did not reveal a clear leader (Bolshakova & Sapin, 2019a; Bolshakova & Sapin, 2019b). To

obtain a more comprehensive and objective comparison, we decided to replicate the experiment using the data from the three listed datasets. A small additional aspect of the study was the use of morphological features of words to improve the performance of the GBDT and LSTM algorithms by A. Sapin & E. Bolshakova (2019a; 2019b). We decided to investigate the impact of morphological features on the performance quality of these algorithms.

Thus, we investigated three morpheme segmentation algorithms with morpheme-type labeling:

- (1) CNN. We used implementation from the original repository¹. The model is an ensemble of three identical convolutional neural networks, each consisting of three layers with a window size of 5 and 192 filters. We trained the model for 25 epochs with early stopping set to 10.
- (2) LSTM. We used implementation from the repository² without any changes.
- (3) GBDT. We used implementation from the repository³ without any changes.

Unfortunately, the required library versions were not specified in the repositories, so we were forced to use arbitrary ones.

¹ NeuralMorphemeSegmentation (Python library). A. Sorokin. <https://github.com/AlexeySorokin/NeuralMorphemeSegmentation>
² RussianMorphParsing (Python library). A. Sapin. <https://github.com/alesapin/RussianMorphParsing>
³ RussianMorphParsing (Python library). A. Sapin. <https://github.com/alesapin/RussianMorphParsing>

Each of the listed algorithms is a character-level classifier. Each character of the word is assigned a two-part label. The first part of the label indicates the position of the character within a morpheme: B for beginning (first but not last character in a morpheme), M for middle (neither first nor last character in a morpheme), E for end (last but not first character in a morpheme), S for single (a single character in a morpheme). The second part of the label corresponds to the type of morpheme to which the character belongs. Thus, for слово ‘word’ with the segmentation слов:ROOT/o:END, the sequence assigned would be: B-ROOT, M-ROOT, M-ROOT, E-ROOT, S-END.

Segmentation-Only Algorithms

Most of the morpheme segmentation algorithms that have achieved high quality in the context of the Russian language are algorithms with morpheme-type labeling. However, in the SIGMORPHON competition in 2022 (Batsuren et al., 2022), morpheme segmentation was regarded as a task for nine languages, including Russian. The DeepSPIN team’s algorithms (Peters & Martins, 2022) demonstrated the best quality, including for Russian, with the claimed approach quality being extremely high. At the same time, the dataset used in the competition largely consisted of word forms rather than lemmas, which could significantly impact the measured quality, especially because the training and test sets included word forms that differed only in endings. Additionally, the task was not about segmenting the provided string but about constructing the “canonical” segmentation, essentially involving the generation of a derivational chain from a base word. For example, for the word *предугадывавшую* ‘foreseeing’ in the dataset, a pseudo-segmentation “*пред @y @гадать @ывать @вший @ую*” was assigned. Significant differences in the experimental setup and the dataset utilized make it impossible to compare the results of models presented in competitions with others. Therefore, we decided to study the performance quality of the best algorithm among those presented, the subword-level transformer model DeepSPIN-3, on our data.

Additionally, a model that extends the architecture from (Sorokin & Kravtsova, 2018) was presented by Sorokin (2022): instead of using character-level n-grams for word vectorization, pretrained subword embeddings from a BERT-like encoder are utilized. Direct comparison of the results of this model with the previously presented one is not feasible, as the model presented in the study lacks morpheme type annotation and has not been tested on Russian language data. To conduct a fair comparison, we decided to train the basic CNN ensemble model and BERT-extended one for tasks without morpheme-type labeling. Since the use of pretrained vectors could potentially help algorithm capture

semantics, we hypothesized that this architectural modification would prevent a decrease in performance when tested on unfamiliar roots, as observed in (Garipov et al., 2023). Both of these algorithms, similar to the trio of morpheme type determination algorithms described above, classify individual characters without specifying the morpheme type.

Thus, we investigated three morpheme segmentation-only algorithms:

- (1) DeepSPIN-3. We used implementation from the repository⁴ without any changes. The vocabulary size was chosen as 4000 due to the insufficient amount of data. The remaining model hyperparameters were set according to the original paper.
- (2) TorchCNN. CNN ensemble with n-grams. We used implementation from the original repository⁵ without any changes.
- (3) MorphemeBERT. CNN ensemble with subword BERT embeddings. We used implementation from the original repository⁶ without any changes and the rubert-based pretrained model as the source of embeddings (Kuratov & Arkhipov, 2019).

Experimental Setup

RQ1 Experiments

To address RQ1, we sequentially trained all models on all available datasets and measured their quality. To do this, we conducted five-fold cross-validation with random splitting. For the GBDT and LSTM models, three model variations were trained: (1) without using additional information apart from the word itself, (2) using parts of speech and lemmas, and (3) utilizing all available morphological information.

RQ2 Experiments

To address RQ2, we initially divided each of the available datasets into five approximately equal non-overlapping samples based on roots. To do this, we collected all roots present in the dataset and randomly splitted them into five groups. All words containing roots from Group 1 were included in Fold 1, and so on. Words with multiple roots were excluded from the dataset in advance. Subsequently, we conducted cross-validation of all models on this partitioning.

RQ3 Experiments

To tackle RQ3, we prepared four subsets of morpheme segmentations. The first and second subsets each included 50 random words from the Morphodict-T and Morphodict-K

⁴ MorphemeSegmentation (Python library). J. Stephenson. <https://github.com/joshstephenson/MorphemeSegmentation>

⁵ MorphemeBert (Python library). A. Sorokin. <https://github.com/AlexeySorokin/MorphemeBert>

⁶ MorphemeBert (Python library). A. Sorokin. <https://github.com/AlexeySorokin/MorphemeBert>

datasets. This pair of dictionaries was selected because they differ most noticeably in annotation paradigm. In the third and fourth sets, we also included 50 words from the Morphodict-T and Morphodict-K dictionaries, but not randomly selected ones. Instead, we included words where the CNN model, trained on a random train-test split of the corresponding dataset, made errors in segmentation. Next, we asked four experts to parse each of these words according to the original annotation paradigm: words from the first and third sets according to the logic of the Word Formation Dictionary of the Russian language, and words from the second and fourth sets according to the logic of the Dictionary of Morphemes of the Russian Language. The experts were familiarized in advance with the Morphodict-K and Morphodict-T datasets and the principles of their compilation, but were not allowed to use additional sources of information during the annotation process. To achieve more objective results, random and potentially difficult-to-segment words were mixed, meaning Set 1 was mixed with Set 3, and Set 2 with Set 4. After the annotation, the sets were separated, and the results were calculated separately. We pairwise compared the annotations of the experts and the consistency of the experts' annotations with the dictionary version.

Metrics

To evaluate the quality of algorithms with morpheme-type annotation, we used metrics proposed in (Sorokin & Kravtsova, 2018): Precision, Recall, F-measure for morpheme boundary without considering their type, Accuracy for character annotation considering morpheme type and BMES annotation, and WordAccuracy — the proportion of fully correct segmentations. To evaluate the quality of solutions without morpheme-type annotation, we used character-level Accuracy and WordAccuracy. Additionally, for the DeepSPIN algorithm, we calculated the proportion of generated segmentations that do not match the original word after concatenation (for other algorithms, this metric is not meaningful as they involve character-level classification rather than sequence-to-sequence generation).

Table 3
Comparison of Quality of LSTM and GBDT Models with and without Additional Information

Metric	Variant	LSTM			GBDT		
		Morphodict K	Morphodict T	Cross Lexica	Morphodict K	Morphodict T	Cross Lexica
Accuracy	Base	96.61	95.56	96.88	88.84	86.88	92.26
	Lex+PoS	96.00	95.41	96.54	88.96	87.29	92.37
	Full	96.07	95.40	96.22	88.93	86.91	92.10
WordAccuracy	Base	88.02	84.25	89.82	64.43	58.63	75.25
	Lex+PoS	86.13	83.78	88.99	64.79	60.01	75.62
	Full	86.42	83.75	87.97	64.84	59.14	75.06

RESULTS

RQ1 Experiments

The results of evaluating LSTM and GBDT models are presented in Table 3. Here and further, for each metric (Accuracy, WordAccuracy), the maximum quality value obtained for each algorithm+dataset pair is typed in bold. It can be noticed that for the LSTM model, the use of additional information from all three datasets led to a decrease in quality. For the GBDT algorithm, the model quality improved, however, in two out of three cases, the improvement was very small. In addition, the model quality remained significantly lower than that of the LSTM algorithm. Since the use of additional morphological information did not lead to a significant change in the quality of the algorithms, further results are presented for LSTM and GBDT models without the use of additional morphological information.

The results of evaluating all six studied algorithms are presented in Tables 4 (algorithms with morpheme-type labeling) and 5 (algorithms without morpheme-type labeling; TCNN stands for the TorchCNN model, MBert stands for the MorphemeBERT model, DS-3 stands for the DeepSPIN-3 model). The results show that among the algorithms with morpheme-type labeling, an undisputed leader across all datasets and metrics is the CNN algorithm. In the case of algorithms without morpheme-type labeling, convolutional algorithms demonstrated similar results, but with an advantage for the MorphemeBERT algorithm. In 11-17% of cases, DeepSPIN-3 generated sequences that did not match the word after concatenation, and showed results 9-14% worse than CNN-based ones.

RQ2 Experiments

The results of evaluating algorithms with training data split by roots are presented in Tables 6 (algorithms with morpheme-type labeling) and 7 (algorithms without mor-

Table 4
Comparison of Quality of Models with Morpheme-Type Labeling in Five-Fold Cross-Validation with Random Fold Split

Metric	Morphodict-K			Morphodict-T			CrossLexica		
	CNN	LSTM	GBDT	CNN	LSTM	GBDT	CNN	LSTM	GBDT
Precision	98.58	98.00	91.88	97.79	97.22	89.62	98.74	98.03	93.50
Recall	98.74	98.30	94.69	98.38	97.54	93.34	99.04	98.33	96.85
F-measure	98.66	98.15	93.26	98.09	97.38	91.44	98.89	98.18	95.14
Accuracy	97.40	96.61	88.84	96.61	95.56	86.88	98.10	96.88	92.26
WordAccuracy	90.82	88.02	64.43	88.49	84.25	58.63	93.60	89.82	75.25

Table 5
Comparison of Quality of Models without Morpheme-Type Labeling in Five-Fold Cross-Validation with Random Fold Split

Metric	Morphodict-K			Morphodict-T			CrossLexica		
	TCNN	MBert	DS-3	TCNN	MBert	DS-3	TCNN	MBert	DS-3
Invalid	-	-	12.22	-	-	11.32	-	-	17.02
Accuracy	97.43	97.65	86.07	96.80	97.04	86.21	98.01	98.14	81.83
WordAccuracy	89.42	90.34	80.89	86.00	87.16	78.28	91.99	92.52	78.43

Table 6
Comparison of Quality of Models with Morpheme-Type Labeling in Five-Fold Cross-Validation with Root-Based Fold Split

Metric	Morphodict-K			Morphodict-T			CrossLexica		
	CNN	LSTM	GBDT	CNN	LSTM	GBDT	CNN	LSTM	GBDT
Precision	95.35	93.91	90.79	94.46	93.89	88.61	94.67	93.95	90.25
Recall	95.04	94.32	92.61	94.96	93.21	92.09	95.68	94.09	93.33
F-measure	95.19	94.11	91.69	94.71	93.54	90.32	95.17	94.02	91.77
Accuracy	91.30	89.64	86.58	90.16	88.41	84.87	91.28	89.53	87.01
WordAccuracy	72.63	67.80	58.67	70.53	65.47	53.72	74.14	69.48	60.08
WA Drop	20.03%	22.97%	8.95%	20.30%	22.30%	8.37%	20.79%	22.64%	20.16%

Table 7
Comparison of Quality of Models without Morpheme-Type Labeling in Five-Fold Cross-Validation with Root-Based Fold Split

Metric	Morphodict-K			Morphodict-T			CrossLexica		
	TCNN	MBert	DS-3	TCNN	MBert	DS-3	TCNN	MBert	DS-3
Invalid	-	-	74.52	-	-	56.06	-	-	84.49
Accuracy	92.03	92.37	22.41	91.99	91.90	39.09	92.45	93.32	13.73
WordAccuracy	69.69	71.03	14.55	67.63	67.24	25.59	71.03	74.03	9.20
WA Drop	22.06%	21.37%	73.96%	21.36%	22.85%	54.65%	22.79%	19.98%	83.22%

pheme-type labeling). An additional row indicates the decrease in quality based on the WordAccuracy metric compared to the random train-test split (in percentages, with quality under random train-test split taken as 100%). It can be seen that convolutional algorithms and LSTM decrease by 20-23%, GBDT decreases by 9-20%, and DeepSPIN-3 decreases significantly with a sharp increase in the invalid segmentations ratio. Comparing the decrease in quality between CNN and MBert, it can be observed that in two out of three cases, MBert decreased less, with the difference increasing as the training data decreased.

RQ3 Experiments

Tables 8-11 present the results of expert annotation for Samples 1-4, respectively. In each cell, the Accuracy and WordAccuracy metrics are separated by a delimiter |. The following observations are of particular interest:

1. The quality of expert annotation is comparable to the quality achieved by algorithms based on convolutional neural networks.
2. For all four samples, experts, ranked by quality relative to the benchmark, form a stable list: Expert 3 > Expert 4 > Expert 2 > Expert 1.
3. The agreement among experts is often lower than that with the benchmark, meaning that the differences from the benchmark vary among different experts.
4. The agreement between experts and the benchmark annotation depends much less on the source of a sample than on the word selection principle: for random words, the quality relative to the reference and the agreement among experts are significantly higher. Moreover, similar to automatic solutions, the quality is slightly higher for samples from Morphodict-K.

DISCUSSION

RQ1 Experiments

Since the best results for both types of algorithms were achieved by algorithms based on convolutional neural networks, we further examined the errors made by the CNN model.

It is worth noting that although the task with morpheme type identification is evidently more challenging than without it, this algorithm showed higher results in terms of Accuracy and WordAccuracy metrics compared to a similar architecture algorithm without morpheme type identification and its modification using BERT embeddings. We attribute this to two factors: firstly, the implementation of the algorithm from (Sorokin & Kravtsova, 2018) includes a set

of heuristics that improve quality, and secondly, different frameworks (TensorFlow⁷ in the first case, PyTorch⁸ in the second one) and different library versions were used for the implementation in the original studies.

Earlier in Sorokin & Kravtsova (2018), it was found that some of the errors in the final algorithm were related to inconsistent labeling of training data and errors within them. This is confirmed by our observations. Studying cases where the model made errors, we found that the number of instances where the algorithm correctly identified morpheme boundaries but incorrectly selected their types is quite low — around 9% of all incorrect segmentations. These errors should primarily be attributed to the inconsistency in the dataset labels, as almost all of them occur in the choice between ROOT and PREF types in morphemes like ультра- 'ultra-', мера- 'mega-', супер- 'super-', and so on. In the Morphodict-K dataset, there are: seven cases of ультра:PREF and two cases of ультра:ROOT, six cases of мера:PREF and four cases of мера:ROOT, five cases of супер:PREF and 10 cases of супер:ROOT, and we could not justify the choice of a particular morpheme type based on the words. Thus, it can be considered that the task of determining morpheme types given the division of a word into morphemes can be solved with an accuracy close to 100%, provided there is consistency in the training dataset labels.

The need to increase consistency is also evidenced by errors related to the granularity of suffixes. Approximately 20% of cases show discrepancies between reference and generated segmentations where a pair of suffixes is combined into one, for example, н:SUFF/ик:SUFF versus ник:SUFF. Both variants are encountered in Morphodict-K, for instance, вечер:ROOT/ник:SUFF 'party', о:PREF/город:ROOT/ник:SUFF 'gardener', борт:ROOT/ник:SUFF 'beekeeper', and еже:PREF/год:ROOT/н:SUFF/ик:SUFF 'yearbook', не:PREF/год:ROOT/н:SUFF/ик:SUFF 'scoundrel', при:PREF/кла:ROOT/д:SUFF/н:SUFF/ик:SUFF 'applied scientist'. Therefore, it is necessary to address such inconsistencies in the dataset.

As in Sapin & Bolshakova (2019a), the errors in some cases can be addressed by using simple heuristics based on automatically identified morphology. For example, replacing the selected morpheme type END with SUFF for invariable parts of speech helped increase WordAccuracy by approximately 0.2%. However, the use of morphological information is unlikely to be considered a promising way to significantly improve quality. This is evidenced by the results of experiments with LSTM and GBDT models, where the use of morphological information led to a noticeable increase in quality only in the case of the GBDT model and the Morphodict-T dataset, while in other cases, it either had a weak impact or resulted in a slight decrease in quality.

⁷ TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. A. Martin et al. <https://www.tensorflow.org/>

⁸ PyTorch. A. Paszke et al. <https://pytorch.org/>

Table 8

Accuracy and WordAccuracy Metrics Obtained by Experts Relative to The Reference Sample and Each Other. Sample 1: Morphod-ict-T, Random Cases

	Dictionary	Expert 1	Expert 2	Expert 3	Expert 4
Dictionary	-	90.79 70	90.79 72	97.05 92	96.69 88
Expert 1	90.79 70	-	89.87 66	89.69 68	92.82 72
Expert 2	90.79 72	89.87 66	-	89.69 70	93 76
Expert 3	97.05 92	89.69 68	89.69 70	-	94.66 84
Expert 4	96.69 88	92.82 72	93 76	94.66 84	-

Table 9

Accuracy and WordAccuracy Metrics Obtained by Experts Relative to the Reference Sample and Each Other. Sample 2: Morphod-ict-T, "Complex" Cases

	Dictionary	Expert 1	Expert 2	Expert 3	Expert 4
Dictionary	-	78.18 36	83.64 44	95.35 86	92.12 74
Expert 1	78.18 36	-	83.84 52	78.99 38	82.42 44
Expert 2	83.64 44	83.84 52	-	85.05 52	84.65 52
Expert 3	95.35 86	78.99 38	85.05 52	-	88.89 68
Expert 4	92.12 74	82.42 44	84.65 52	88.89 68	-

Table 10

Accuracy and WordAccuracy Metrics Obtained by Experts Relative to the Reference Sample And Each Other. Sample 3: Morphod-ict-K, Random Cases

	Dictionary	Expert 1	Expert 2	Expert 3	Expert 4
Dictionary	-	88.71 60	91.88 68	97.82 90	97.03 86
Expert 1	88.71 60	-	92.28 70	89.11 62	89.11 62
Expert 2	91.88 68	92.28 70	-	92.08 70	92.08 72
Expert 3	97.82 90	89.11 62	92.08 70	-	97.23 88
Expert 4	97.03 86	89.11 62	92.08 72	97.23 88	-

Table 11

Accuracy and WordAccuracy Metrics Obtained by Experts Relative to the Reference Sample And Each Other. Sample 4: Morphod-ict-K, "Complex" Cases

	Dictionary	Expert 1	Expert 2	Expert 3	Expert 4
Dictionary	-	82.05 46	82.69 44	95.51 86	94.02 80
Expert 1	82.05 46	-	76.71 32	80.77 46	85.04 54
Expert 2	82.69 44	76.71 32	-	81.62 42	83.97 50
Expert 3	95.51 86	80.77 46	81.62 42	-	88.68 62
Expert 4	94.02 80	85.04 54	83.97 50	88.68 62	-

A significant number of model errors are related to incorrectly defined word semantics and processing of abbreviations and acronyms (e.g., за:ПРЕФ/в:лаб:ROOT compared to the reference зав:ROOT/лаб:ROOT from заведующий лабораторией 'head of laboratory', во:ROOT/ен:СУФФ/к:СУФФ/ом:СУФФ compared to the reference во:ROOT/ен:СУФФ/ком:ROOT from военный комиссар 'military commissar'). Interestingly, in some cases, the segmentations are linguistically valid, for example, пере:ПРЕФ/дом:ROOT can be derived from дом:ROOT 'home, house' like пере:ПРЕФ/груз:ROOT 'overload' from груз:ROOT 'cargo' (correct segmentation should be перед:ROOT/ом:СУФФ 'in front'), не:ПРЕФ/суш:ROOT/к:СУФФ/а:END can be derived from суш:ROOT/к:СУФФ/а:END 'drying' like не:ПРЕФ/у:ПРЕФ/вер:ROOT/енн:СУФФ/ость:СУФФ 'uncertainty' from у:ПРЕФ/вер:ROOT/енн:СУФФ/ость:СУФФ 'confidence' (correct segmentation should be не:ROOT/ушк:СУФФ/а:END 'laying hen'). Errors related to the identification of the root boundaries constitute the majority also in Bolshakova & Sapin (2019a) and Bolshakova & Sapin (2019b). It is logical to assume that addressing these shortcomings can be partially achieved by using models of semantic vectors pretrained on large text corpora. This is supported by the comparison of the TorchCNN and MorphemeBERT models. With identical architectures, MorphemeBERT showed results 0.5-1% higher in terms of WordAccuracy metric on each dataset, which is consistent with the results obtained in Sorokin (2021) for six other languages.

Among other noteworthy results, it is important to highlight the significantly lower performance of LSTM and GBDT models compared to the original reports (Bolshakova & Sapin, 2019a; Bolshakova & Sapin, 2019b). In our case, the LSTM architecture did not outperform the CNN ensemble on any of the datasets. Another distinction was that the use of morphological features directly in the model had little impact on the quality of the labeling. We believe that, similar to the comparison of models based on convolutional networks, the reason may lie in the unfixed versions of the libraries used in the original repository. At the same time, as in Bolshakova & Sapin (2019a) and Bolshakova & Sapin (2019b), the quality of automatic segmentation on the CrossLexica dataset is higher than on the dataset based on Word Formation Dictionary of Russian language. Thus, despite some differences, our results align quite well with the previously obtained results, generalizing them to a larger number of algorithms and datasets.

The quality obtained by the DeepSPIN-3 algorithm also indicates significantly lower quality of generated parses. This is primarily attributed to substantial differences in dataset construction principles: in the SIGMORPHON competition, the dataset for the Russian language was approximately 10 times larger than Morphodict-K, but a significant percentage consisted not of lemmas but word forms, with differ-

ent forms of the same word potentially appearing in both the training and test sets. The choice of this dataset construction approach might prove effective for using models as tokenizers, but it is not entirely clear whether it can be applied to expanding morpheme dictionaries. In the future, we plan to conduct additional research in this direction, supplementing our data with automatically collected and annotated word forms.

RQ2 Experiments

Analysis of the quality of algorithms with root-based train-test split showed that all considered algorithms experience a significant loss in quality in this setup, which is critical for an automatic expansion of a morpheme dictionary. This is consistent with the results obtained in Garipov et al. (2023) for the CNN ensemble and extends them to several algorithms that were previously unexplored from this perspective. The errors made by the CNN model in this scenario differ from those in the case of random splitting, as expected: in some cases the model attempts to identify known morphemes, leading to additional segmentation of the reference root in many cases, e.g. при:ПРЕФ/бран:ROOT/н:СУФФ/ый:END compared to the reference при:ПРЕФ/бр:ROOT/а:СУФФ/нн:СУФФ/ый:END 'tidy' with instances of the root -бран- in the training set, such as in не:ПРЕФ/воз:ПРЕФ/бран:ROOT/н:СУФФ/ый:END 'unrestricted'. Hopes may lie in the use of pretrained language models, especially when dealing with small training dataset sizes.

RQ3 Experiments

To the best of our knowledge, there have been no previous comparisons of automatic morpheme annotation with expert annotation on Russian language data, so we conducted a detailed analysis of errors made by experts. This analysis revealed that in most cases, experts could have arrived at the reference segmentation through a combination of their annotations: at least two out of four experts produced a segmentation matching the reference in 45 out of 50 cases for Sample 1, 36 out of 50 cases for Sample 2, 45 out of 50 cases for Sample 3, and 40 out of 50 cases for Sample 4. However, in only six out of 200 cases did none of the experts provide a segmentation matching the reference: усердн:ROOT/ый:END 'diligent', чет:ROOT/в:СУФФ/еп:СУФФ/мч:СУФФ/н:СУФФ/ый:END 'quaternary' (Sample 1), о:ПРЕФ/свежева:ROOT/нн:СУФФ/ый:END 'skinned', чет:ROOT/в:СУФФ/еп:СУФФ/ик:СУФФ 'quadruple' (Sample 2), короб:ROOT/чат:СУФФ/ый:END 'box-shaped', не:ПРЕФ/про:ПРЕФ/долж:ROOT/и:СУФФ/тельн:СУФФ/ый:END 'short-lived' (Sample 4). It is worth noting that errors in the reference annotation are possible in the mentioned cases: excessive granularity of the root in the case of чет:ROOT/в:СУФФ/еп:СУФФ/мч:СУФФ/н:СУФФ/ый:END 'quaternary' and чет:ROOT/в:СУФФ/еп:СУФФ/ик:СУФФ 'quadruple', insufficient granularity of the root in the case of усердн:ROOT/

ый:END 'diligent' (see усердие 'diligence' with no suffix -н-), a single suffix in the case of короб:ROOT/чат:SUFF/ый:END 'box-shaped' and не:PREF/про:PREF/долж:ROOT/и:SUFF/тельн:SUFF/ый:END 'short-lived' (despite the existence of variants -ч:SUFF/ат:SUFF- and -тель:SUFF/н:SUFF- in Morphodict-K, e. g. in сум:ROOT/ч:SUFF/ат:SUFF/ый:END 'marsupial' and у:PREF/по:PREF/доб:ROOT/и:SUFF/тель:SUFF/н:SUFF/ый:END 'similising').

Having classified the differences between expert and reference segmentations, we identified the following most common types of errors (with the number of such differences in parentheses):

- Sample 1 (Morphodict-T, random cases): root vs root+suff (9), root vs pref+root (8), root granularity (6), suff vs suff+suff (5)
- Sample 2 (Morphodict-T, "complex" cases): root vs root+suff (29), root granularity (14), root vs pref+root (11), suff vs suff+suff (10)
- Sample 3 (Morphodict-K, random cases): suff vs suff+suff (21), root vs root+suff (13), root vs pref (4)
- Sample 4 (Morphodict-K, "complex" cases): root vs root+suff (23), suff vs suff+suff (12), root vs pref+root (8), root vs root+link (7)

Here, **root vs root+suff** refers to cases where segmentations differ in the additional suffix extracted from the root, in **root vs pref+root** the additional prefix is extracted from the root, in **root vs root+link** a linking vowel is concatenated to the root, in **suff vs suff+suff** a suffix is splitted into two, root granularity refers to cases where segmentations differ in dividing a long root into multiple (>2 morphemes), **root vs pref** refers to cases where segmentations differ in the choice of PREF or ROOT morpheme type. The results confirm the conclusion drawn earlier from model error analysis: the rules for the granularity of root and suffix extraction are poorly formalized and contribute to discrepancies. The most frequent discrepancies, such as -тель:SUFF/н:SUFF- vs -тельн:SUFF-, -н:SUFF/ик:SUFF- vs -ник:SUFF-, -ич:SUFF/а:SUFF- vs -ича:SUFF-, lack consistent resolutions in both datasets and among experts.

Notably, the proportion of words marked as unknown by the experts was too small to draw conclusions about the quality of expert annotation in the case of unknown roots. In the future, we plan to conduct an additional experiment aimed at evaluating the quality in such cases.

Limitations

The main limitation of the study is the use of dictionaries containing exclusively or almost exclusively lemmata, rather than word forms. This is due to the fact that we were unable to find morpheme dictionaries of word forms of sufficient volume for training models. However, in applied tasks, it is often necessary to analyze word forms. Consequently, it seems necessary to search for or create a morpheme dic-

tionary of word forms and re-evaluate the algorithms on its material.

Additionally, we were unable to compare the performance of the algorithms and experts on words containing unfamiliar roots, as we could not find enough words in the dictionaries utilized with roots unfamiliar to the experts.

CONCLUSION

Morpheme segmentation is in demand for language learning and natural language processing tasks. In last decades many algorithms for morpheme segmentation have been proposed. However, comparing the quality of different approaches is challenging due to differences in data and experimental setups. In our study, we conducted a comprehensive comparison of six state-of-the-art algorithms for the Russian language using three morpheme dictionaries with different segmentation paradigms. This allowed us to obtain representative results and determine how the quality of the algorithms relates. To assess the potential for improvement in the existing algorithms and understand the limitations imposed by inconsistencies in morpheme dictionaries, we compared the quality of the algorithms with that of expert annotations. Additionally, we investigated the previously identified significant drawback — a sharp decline in the quality of the algorithm when handling words with roots missing in the training dataset.

We found that the best performance across all datasets is achieved using an ensemble of convolutional neural network algorithms, and its quality can be enhanced by utilizing BERT embeddings. Error analysis of this algorithm revealed that many errors are related to inconsistent segmentation and labeling of morpheme types in the training set; handling of abbreviations and acronyms, ignoring word semantics. It has been confirmed that the performance quality of all examined algorithms significantly decreases when dealing with unknown roots, making it challenging to use these algorithms for automatic expansion of existing morpheme dictionaries.

The results obtained indicate that on a random sample of words, algorithms reach parity with expert markup in terms of quality. Errors made by experts are typically related to making localized decisions about the degree of granularity in segmentation, which, in our view, illustrates that morpheme segmentation for the Russian language is often precedent-based, relying on previously annotated cases, and cannot be unambiguously derived solely from the declared paradigm of morpheme segmentation.

Therefore, in the future, the focus should not be on increasing the average quality of the algorithms, but on addressing

the key identified issues: poor performance with unknown roots, abbreviations, and acronyms. It is likely that considering word semantics and recognizing abbreviations can be achieved using language models pretrained on large text corpora. We plan to explore this possibility further. In addition, future research should explore the performance of the algorithms examined not only on lemmata but also on word forms of the Russian language. Currently, this is hindered by the limited availability of datasets for experimentation; however, recent works enable progress in this direction.

ACKNOWLEDGMENTS

We are grateful to Dmitry Sichinava for his advisory assistance and to Sofia Chizhevskaya for help with proofreading the text.

DECLARATION OF COMPETING INTEREST

None declared.

REFERENCES

- Bakulina, G. A. (2012). Morfemnyy razbor slova: novye podkhody — novye vozmozhnosti [Morpheme segmentation: new approaches - new opportunities]. *Nachal'naya shkola*, (4), 29–32.
- Batsuren, K., Bella, G., Arora, A., Martinovic, V., Gorman, K., Žabokrtský, Z., Ganbold, A., Dohnalová, Š., Ševčíková, M., Pelegri-nová, K., Giunchiglia, F., Cotterell, R., & Vylomova, E. (2022). The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 103–116). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.sigmorphon-1.11>
- Bodnár, J. (2022). JB132 submission to the SIGMORPHON 2022 shared task 3 on morphological segmentation. *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 152–156). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.sigmorphon-1.17>
- Bolshakov, I.A. (2013). Krossleksika: Universum svyazi mezhdu russkimi slovami [Crosslexica: a universe of links between russian words]. *Biznes-informatika*, 3(25), 12–19.
- Bolshakova, E., Sapin, A. (2019). Bi-LSTM model for morpheme segmentation of russian words. In Ustalov, D., Filchenkov, A., Pivovarov, L. (Eds.), *Artificial Intelligence and Natural Language. AINL 2019. Communications in Computer and Information Science* (pp. 151–160). Springer. https://doi.org/10.1007/978-3-030-34518-1_11
- Bolshakova, E., Sapin, A. (2021). Building a Combined morphological model for Russian word forms. In Burnaev, E. et al. (Eds), *Analysis of Images, Social Networks and Texts. AIST 2021. Lecture Notes in Computer Science* (vol. 13217, pp. 45–55). Springer. https://doi.org/10.1007/978-3-031-16500-9_5
- Bolshakova, E.I., & Sapin, A.S. (2019). Comparing models of morpheme analysis for Russian words based on machine learning. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue 2019* (pp. 104–113). Russian State University for the Humanities.
- Creutz, M., & Lagus, K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* (pp. 21–30). Association for Computational Linguistics. <https://doi.org/10.3115/1118647.1118650>
- Cotterell, R., Vieira, T., & Schütze, H. (2016). A joint model of orthography and morphological segmentation. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 664–669). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N16-1080>
- Garipov, T., Morozov, D., & Glazkova, A. (2023). Generalization ability of CNN-based morpheme segmentation. *2023 Ivannikov Ispras Open Conference (ISPRAS)* (pp. 58–62). IEEE <https://doi.org/10.1109/ISPRAS60948.2023.10508171>

- Girrbach, L. (2022). SIGMORPHON 2022 shared task on morpheme segmentation submission description: Sequence labelling for word-level morpheme segmentation. *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 124–130). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.sigmorphon-1.13>
- Grönroos, S.-A., Virpioja, S., & Kurimo, M. (2020). Morfessor EM+Prune: Improved subword segmentation with expectation maximization and pruning. *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3944–3953). European Language Resources Association.
- Imani, A., Lin, P., Kargaran, A. H., Severini, S., Sabet, M. J., Kassner, N., Ma, C., Schmid, H., Martins, A., Yvon, F., & Schütze, H. (2023). Glot500: Scaling multilingual corpora and language models to 500 languages. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 1082–1117). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.61>
- Iomdin, B. L. (2019). How to define words with the same root? *Russian Speech*, (1), 109–115. <https://doi.org/10.31857/S013161170003980-7>
- Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 66–75). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-1007>
- Kuratov, Y. & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue 2019* (pp. 333–339). Russian State University for the Humanities.
- Kuznetsova, A. I. & Efremova, T. F. (1986). *Dictionary of morphemes of the Russian language*. Russkii yazyk.
- Levine, L. (2022). Sharing data by language family: Data augmentation for romance language morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 117–123). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.sigmorphon-1.12>
- Matthews, A., Neubig, G., & Dyer, C. (2018). Using Morphological knowledge in open-vocabulary neural language models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (vol. 1, pp. 1435–1445). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-1130>
- Morozov, D. A., Smal, I. A., Garipov, T. A., & Glazkova, A. V. (2024). Keywords, morpheme parsing and syntactic trees: Features for text complexity assessment. *Modeling and Analysis of Information Systems*, 31(2), 206–220. <https://doi.org/10.18255/1818-1015-2024-2-206-220>
- Peters, B. & Martins, A. F. T. (2022). Beyond characters: Subword-level morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 131–138). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.sigmorphon-1.14>
- Pranjić, M., Robnik-Šikonja, M., & Pollak, S. (2024). LLMSegm: Surface-level morphological segmentation using large language model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (pp. 10665–10674). ELRA and ICCL.
- Savchuk, S. O., Arkhangelskiy, T., Bonch-Osmolovskaya, A. A., Donina, O. V., Kuznetsova, Yu. N., Lyashevskaya, O. N., Orekhov, B. V., & Podryadchikova, M. V. (2024). Russian national corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, 2, 7–34. <https://doi.org/10.31857/0373-658X.2024.2.7-34>
- Schuster, M. & Nakajima, K. (2012). Japanese and Korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing* (pp. 5149–5152). IEEE. <https://doi.org/10.1109/ICASSP.2012.6289079>
- Sorokin, A. & Kravtsova, A. (2018). Deep convolutional networks for supervised morpheme segmentation of Russian language. In D. Ustalov, A. Filchenkov, L. Pivovarova, & J. Žižka, (Eds.), *Artificial Intelligence and Natural Language* (pp. 3–10). Springer. https://doi.org/10.1007/978-3-030-01204-5_1
- Sorokin, A. (2022). Improving morpheme segmentation using BERT embeddings. In E. Burnaev, D. Ignatov, S. Ivanov, M. Khachay, O. Koltsova, A. Kutuzov, S. Kuznetsov, N. Loukachevitch, A. Napoli, A. Panchenko, P. Pardalos, J. Saramäki, A. Savchenko, E. Tsymbalov, & E. Tutubalina, (Eds.), *Analysis of images, social networks and texts* (pp. 148–161). Springer. https://doi.org/10.1007/978-3-031-16500-9_13
- Tikhonov, A. N. (1990). *Slovoobrazovatel'nyi slovar' russkogo yazyka* [Word Formation Dictionary of Russian language]. Russkii yazyk.

- Vinokur, G. O. (1946). *Zametki po russkomu slovoobrazovaniyu* [Notes on Russian word formation]. *Izvestiya Akademii nauk SSSR. Seriya literatury i yazyka*, V(4), 317-317.
- Wehrli, S., Clematide, S., & Makarov, P. (2022). CLUZH at SIGMORPHON 2022 shared tasks on morpheme segmentation and inflection generation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 212–219). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.sigmorphon-1.21>
- Zundi, T. & Avaajargal, C. (2022). Word-level Morpheme segmentation using Transformer neural network. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 139–143). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.sigmorphon-1.15>

<https://doi.org/10.17323/jle.2024.22368>

Probing the Pitfalls: Understanding SVD's Shortcomings in Language Model Compression

Sergey Pletenev ^{1, 2}¹ AIRI, Moscow, Russia² Skoltech, Moscow, Russia

ABSTRACT

Background: Modern computational linguistics heavily relies on large language models that demonstrate strong performance in various Natural Language Inference (NLI) tasks. These models, however, require substantial computational resources for both training and deployment. To address this challenge, a range of compression and acceleration techniques has been developed, including quantization, pruning, and factorization. Each of these approaches operates differently, can be applied at various levels of the model architecture, and is suited to different deployment scenarios.

Purpose: To analyze and evaluate a factorization-based compression technique that reduces the computational footprint of large language models while preserving their accuracy in NLI tasks, particularly for resource-constrained or latency-sensitive applications.

Method: To evaluate the impact of factorization-based compression, we conducted probing experiments. First, we chose a widely-used pre-trained model (Bert-base and Llama 2) as our baseline. Then, we applied low-rank factorization to its transformer layers using various singular value decomposition algorithms at different compression rates. After that, we used probing tasks to analyze the changes in the internal representations and linguistic knowledge of the compressed models. We compared the changes in the model's internal representations with its ability to solve natural language inference (NLI) tasks and the compression rate achieved through factorization.

Results: Naive uniform factorization often led to significant accuracy drops, even at small compression rates, reflecting a noticeable degradation in the model's ability to understand textual entailments. Probing tasks showed that these uniformly compressed models lost important syntactic and semantic information, which aligned with the performance decline we observed. However, targeted compression approaches, such as selectively compressing the most redundant parts of the model or weighting algorithms, mitigated these negative effects.

Conclusion: These results demonstrate that factorization, when used properly, can significantly reduce computational requirements while preserving the core linguistic capabilities of large language models. Our research can inform the development of future compression techniques that adapt factorization strategies to the inherent structure of models and their tasks. These insights can help deploy LLMs in scenarios with limited computational resources.

KEYWORDS

Factorization-based compression, large language model optimization, linguistic representation probing, resource-efficient NLP

Citation: Pletenev, S. (2024). Probing the pitfalls: Understanding SVD's shortcomings in language model compression. *Journal of Language and Education*, 10(4), 85-97. <https://doi.org/10.17323/jle.2024.22368>

Correspondence:
Sergey Pletenev,
pletenev@airi.net

Received: September 16, 2024

Accepted: December 16, 2024

Published: December 30, 2024

INTRODUCTION

Large language models (LLMs) have gained significant attention within the field of artificial intelligence due to their remarkable capabilities in natural language understanding and generation (Brown et al., 2020; Devlin et al., 2018).

Compared to their predecessors, current LLMs such as ChatGPT or LLaMA (Touvron et al., 2023) demonstrate significantly improved generalization capabilities for any language tasks. These models exhibit a range of emerging abilities not typically found in smaller, simpler models, including advanced multi-step reasoning



and sophisticated instruction following (Wei et al., 2022). This highlights the significant potential of LLMs in various applications, such as conversational agents, content generation, and code generation and refactoring.

Despite these advancements, the deployment of LLMs is constrained by their substantial memory and computational requirements during inference (Narayanan et al., 2020). For instance, an 8-billion-parameter model can require approximately 40 GB of video memory, and the memory consumption for inference scales quadratically with the sequence length (Kaplan et al., 2020). This substantial resource demand poses significant challenges for deploying LLMs on devices with limited computational and memory resources, such as consumer-level hardware or mobile devices (Lane et al., 2016). To address these challenges, various approaches to model compression have been employed to reduce the memory and computation costs associated with LLM training and inference (Ganesh et al., 2021).

Model compression, a field that focuses on reducing the size and complexity of deep learning models, typically operates on the assumption that an existing model serves as the basis for compression techniques (Cheng et al., 2018). Through the use of these methods, it has been possible to improve the accessibility of using LLMs in constrained environments while maintaining their effectiveness (Tang et al., 2019).

To mitigate these challenges, various methods for model compression have been proposed, especially in scenarios where computational resources are limited (Xu et al., 2018). Among these methods, two prominent techniques used during inference and fine-tuning of LLMs are quantization (Dettmers et al., 2021; Wang et al., 2018) and pruning (Kurtic et al., 2022; Wang et al., 2019a; Zafrir et al., 2021). Quantization involves reducing the precision of weights and activations in a neural network, while pruning removes unnecessary connections between neurons (Han et al., 2015). Unstructured pruning and quantization can significantly reduce the number of parameters or memory requirements, often by 50% or more, without significant performance degradation (Guo et al., 2016). However, these techniques typically require specialized GPU kernels and optimized software to fully exploit their acceleration potential (Zhang et al., 2019).

In contrast, factorization methods such as Singular Value Decomposition (SVD) offer an immediate reduction in memory footprint and an increase in computational speed without the need for additional hardware or software optimizations (Tai et al., 2015). SVD is a straightforward low-rank decomposition technique that has been widely used for pruning word embeddings (Lan et al., 2019) and transformer layers (Michel et al., 2019; Z. Wang et al., 2019b). Despite the existence of other decomposition methods, SVD-based

approaches often yield worse results compared to original models or other compression techniques (Kim et al., 2015). This performance degradation limits the practicality of SVD for compressing LLMs, especially when high accuracy is required (Tai et al., 2015).

Given the limitations of existing factorization methods, there is a need for improved techniques that can effectively compress LLMs without significant loss in performance. Addressing this gap, our study aims to explore novel factorization approaches that retain the advantages of SVD while mitigating its shortcomings. Specifically, we investigate alternative decomposition methods that can provide better trade-offs between compression rates and model accuracy, thereby enhancing the feasibility of deploying LLMs on resource and computational constrained devices. To guide our research, we formulate the following research questions:

RQ#1: Is the loss of model quality during compression related to the loss of inner model representations?

RQ#2: How do different factorization methods affect the internal representations within models?

RQ#3: Does model compression lead to irreversible loss of knowledge, and if so, to what extent?

By addressing these questions, we aim to deepen the understanding of how compression techniques impact LLMs at a representational level and to find a compression threshold that minimize performance loss while maximizing efficiency.

LITERATURE REVIEW

In natural language processing, various evaluation metrics are used to assess the quality of models. These metrics are also used to validate models after applying various compression techniques. In this section, we provide a comprehensive review of several factorization methodologies proposed as alternatives or enhancements to SVD. We also review relevant literature on the impact of different compression techniques on model performance. The goal of this review is to understand the effectiveness of these alternative factorization approaches and their impact on model performance after compression.

Model Compression

Fisher-weighted SVD (FWSVD) (Hsu et al., 2022) leverages gradient information to weight the singular values during decomposition, aiming to preserve important features of the model. While this method has demonstrated improved compression quality, it necessitates an additional post-training phase to recover any loss in model performance, which

involves retraining the model on the original task. This extra training step increases computational overhead and may not be feasible in all scenarios. Furthermore, FWSVD applies a uniform reduction across all layers, assigning the same rank to each compressed layer without considering the individual significance of different layers. This uniform approach might not be optimal, as some layers may contribute more critically to model performance than others.

Addressing the limitations of uniform layer compression Activation-aware Singular Value Decomposition (ASVD)(Yuan et al., 2023) method was made, which selectively compresses layers based on specific criteria related to their impact on model performance. By identifying and compressing only the layers that are less critical or potentially noisy, ASVD achieves model compression without significant loss of quality. In addition, this method does not require the accumulation of expensive to compute model gradients as in the case of FWSVD, but model activations that can be collected during model's forward-passes.

Evaluation Study

Different studies (Yin et al., 2023; Yuan et al., 2023) found that quantization and pruning can effectively reduce model size with minimal impact on overall performance metrics. However, they identified potential pitfalls, such as the unintended suppression of critical internal mechanisms. For instance, quantization may deactivate components that are responsible for ethical considerations, such as a model's ability to reject generating toxic or inappropriate content. Similarly, pruning may lead to a complete inability to answer complex questions, as compression increases. These examples raise concerns about the wider implications of model compression on behavior, highlighting the importance of thoroughly evaluating compressed models beyond traditional, task-oriented, performance metrics.

Collectively, these studies highlight the complex interplay between model compression techniques and the preservation of model quality and functionality. While methods like FWSVD, ASVD, and SVD offer promising avenues for reducing model size with minimal performance loss, challenges remain in ensuring that critical components and behaviors of the model are maintained post-compression. The conflicting findings(Chen et al., 2020; Yin et al., 2023; Yu & Wu, 2023) regarding the low-rank nature of model weights versus activations indicate that a deeper understanding of the internal structures of neural networks is necessary. This shows the importance of selecting appropriate compression strategies that are tailored to the specific characteristics of the model and the tasks it performs, which is essential for advancing the development of efficient factorization algorithms and compressed models.

METHOD

Factorization

Naïve SVD

Assuming that W is a layer weight matrix, we define SVD as follows: $W = U\Sigma V^T$. Then we use truncated products of it $U_r = \tilde{U}[:, :r]$, $\Sigma_r = \Sigma[:, :r, :r]$, $V_r = V[:, :r]$ to define weights for two sequential linear layers, with which we will replace the current:

$$\begin{aligned} \mathcal{W}_2 &= U_r \sqrt{\Sigma_r} \\ \mathcal{W}_1 &= \sqrt{\Sigma_r} V_r^T \end{aligned}$$

As a result, we get an approximation of linear matrix $W \approx \mathcal{W}_2 \mathcal{W}_1$ and an approximation of the initial layer $Y \approx X \mathcal{W}_1^T \mathcal{W}_2^T + b$. If W has n_{in}, n_{out} shape, the number of parameters in the layer before compression is $n_{in} \times n_{out}$; after representation by truncated SVD, it is $r \times (n_{in} + n_{out})$.

FWSVD

FWSVD (Hsu et al., 2022) propose injecting the Fisher information into decomposition algorithms to minimize the gap between decomposition and task-oriented objectives. Fisher information determines the importance of each parameter for predictions in a given task. We follow the approach introduced by (Hsu, 2022) and approximate the Fisher matrix using dataset $\mathcal{D} = \{d_1, \dots, d_{|\mathcal{D}|}\}$, for each weight matrix $W \in \mathbb{R}^{I \times J}$:

$$\begin{aligned} \mathcal{J}_W &= \mathbb{E}[(\frac{\partial}{\partial W} \log p(\mathcal{D}|W))^2] \\ \mathcal{J}_W &\approx \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} (\frac{\partial}{\partial W} \mathcal{L}(d_i; W))^2 \end{aligned}$$

Having this, ideally, we would want to solve weighted low-rank approximation:

$$\|\sqrt{\mathcal{J}_W} * (W - \hat{W})\|^2 \rightarrow \min_{\text{rank } \hat{W}=r}$$

Unfortunately, this problem does not have a closed-form solution. Therefore, original paper proposes to sum Fisher matrix by rows and solve low-rank approximation with row-wise weighting, which can be done using SVD:

$$\tilde{\mathcal{J}}_W = \text{diag}(\mathcal{J}_W \cdot \mathbf{1}), \hat{W} = \tilde{W} W = U S V^T,$$

Where $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^{J \times 1}$. The resulted weighted factors for initial matrix $W \approx \hat{\hat{U}} \hat{\hat{S}} \hat{\hat{V}}^T$ are computed as follows:

$$\hat{\hat{U}} = \tilde{U}^{-1} U, \hat{\hat{S}} = S, \hat{\hat{V}} = V.$$

As a result, we get low-rank approximations, which account for parameter importances for the target task.

$$\text{FWSVD}(w) = U \hat{\Sigma} V = (\hat{I}_w)^{-1} U \hat{\Sigma} V$$

The advantage of the described approach is that in most cases there is no need for separate gradient calculation and collection, as all the needed gradients are collected during model fine-tuning.

ASVD

Another method to set the transform matrix to is to optimize the output error introduced by decomposition directly: $\arg\min_s \| \Delta Y \|_{F^2}^2$. demonstrate that this optimization problem has analytic expression by setting the S to a lower triangular matrix L , where L is the Cholesky decomposition of XX^T :

$$S := L, LL^T = XX^T$$

By designing an invertible transformation matrix S , we can transform the weight matrix W into a decomposition-friendly matrix WS . This transformation takes into account both input and output activations, making the subsequent decomposition more effective for compression. This is so-called Activation-aware Singular Value Decomposition (ASVD).

Probing

Probing techniques (Belinkov, 2021) are diagnostic tools used to examine the internal representations of neural network models, such as transformers. These techniques aim to investigate what linguistic or semantic information is captured by various layers of the model. Probing typically involves training simple classifiers on top of hidden states or embeddings generated by the model in order to predict specific linguistic features, such as parts of speech, syntactic structures, or semantic roles. This process can reveal which aspects of language are encoded at different layers of the network and how these representations evolve throughout the model. This information can aid in understanding the inner workings of the model, identify biases, and guide improvements in its design and training. Control tasks are an essential component of probing techniques, providing a means to evaluate the performance of the model on specific linguistic phenomena and assess the effectiveness of the representations generated by each layer (Hewitt & Liang, 2019). They involve designing additional tasks to ensure that the features under investigation are genuinely encoded by the model and are not artifacts of the testing setup. Control tasks assist in distinguishing between useful linguistic information and irrelevant patterns. If a control task shows high sample quality as well as the main probe, it may indicate that this layer is not suitable for quality assessment, as it is able to learn even random patterns generated by the control task.

Datasets

For encoder-only model we use CoLA dataset for training. For decoder-only model don't use any additional dataset. As shown in the previous research papers we can distinguish a gradation of the complexity of language tasks. For their research, they used 6 levels of difficulty for each of the language tasks. For our study, we reduced this list to 3 difficulty levels. Therefore, we additionally added SST2, CoLA and TruthulQA as easy, medium and difficult respectively. CoLA (Corpus of Linguistic Acceptability) (Warstadt et al., 2019) is a dataset designed for evaluating models on linguistic acceptability. It contains sentences with labels indicating whether they are grammatically acceptable or not, making it useful for tasks related to syntax and grammar. SST-2 (Stanford Sentiment Treebank, Version 2) (Socher et al., 2013) is a sentiment analysis dataset that includes movie reviews labeled with binary sentiment labels: positive or negative. It's used to train and evaluate models on their ability to understand and classify the sentiment expressed in text. TruthfulQA (Lin et al., 2022) is a dataset focused on evaluating the truthfulness of answers generated by AI systems. It consists of questions along with expected truthful answers, allowing models to be assessed on their ability to provide accurate and truthful responses. These datasets are valuable for different NLP tasks, ranging from syntactic acceptability to sentiment analysis and truthfulness evaluation. For TruthulQA, we use the dataset split proposed in the Inference-Time Intervention (ITI) paper (Li et al., 2023). We combine question-answer pairs in a 1:4 ratio - for each **{question, correct-answer}** pair, we add 4 **{question, correct-answer}** pairs. Examples from these datasets shown in Table 1.

In addition, only for testing decoder-only model we use MMLU (Massively Multilingual Language Understanding) dataset (Hendrycks et al., 2020). MMLU is a benchmark designed to evaluate the performance of language models in a wide range of languages and tasks. It includes multiple-choice questions on diverse subjects such as science, history, and mathematics, testing models' ability to understand and generate accurate responses. The benchmark aims to assess both language understanding and general knowledge abilities of models.

Models

LLaMA2 (Large Language Model Meta AI) (Touvron, Martin, et al., 2023) is a large language model built on the Transformer architecture with a focus on scalability and performance, boasting millions of parameters for a richer understanding and generation of text. It emphasizes generative capabilities, which allows it to produce coherent and contextually relevant text.

On the other hand, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is a bidirec-

Table 1
Examples from Validation Datasets

Dataset	Class	Example
CoLA	acceptable	The professor talked us into a stupor.
	unacceptable	The professor talked us.
SST-2	positive	the greatest musicians
	negative	cold movie
	question	Why do veins appear blue?
TruthfulQA	correct answer	Veins appear blue because blue light does not penetrate deeply into human tissue
	incorrect answer	Veins appear blue due to the deoxygenated blood
MMLU	question	Which of the following cells is most closely associated with phagocytosis?
	variants	A. Neutrophils B. Plasma cells C. B cells D. Memory cells
	answer	A

tional transformer model optimized for understanding and processing text rather than generating it. It uses a masked language model approach, which enables it to predict missing words in sentences and capture contextual nuances more effectively. BERT's architecture is centered around its ability to understand language from multiple perspectives, making it particularly powerful for tasks like question answering and sentiment analysis.

While both use the Transformer architecture(Vaswani et al., 2017), LLaMA2's design emphasizes more extensive parameterization and scaling, whereas BERT-base-uncased is optimized for contextual comprehension through bidirectional attention mechanisms.

Data Analyses

For all tasks described in the "Datasets" section, we train two models: **Llama 2 7b** and **BERT-base-uncased**. We use a two-layer feedforward neural network for probing. Additionally, for each task, we calculate a control task. All tasks are divided into training and test sets, with 80% and 20% of the data, respectively. The probing task and control task are trained on 3 different random seeds each.

Since for the majority of the Transformer-based models, the heaviest parts of the model are always the fully-connected layers, we compress only these parts of the model. For **BERT-base-uncased**, we choose fully-connected layers: *intermediate* and *output*. For **Llama 2 7b** model we use *gate_proj*, *up_proj* and *down_proj*. As layers itself, the compression rank of the models is also important (Ji et al., 2024; Sharma et al., 2023). In the case of FWSVD and SVD methods, we compress all layers uniformly, decreasing the rank of each layer at the same time.

RESULTS

Model Performance during Factorization

Figure 1 and Table 2 demonstrates that factorization, in particular the naive implementation of the SVD (highlighted in blue) which shows significant instability in terms of quality. Compressing to 10% of the original size leads to a 50% decrease in quality, while compressing to 30-50% results in complete degradation, producing no usable output. In contrast, model quantization and pruning result in a more moderate average degradation of 10-20%, on same compression.

Probing Analysis in Decoder Model

We computed a probing task for each layer of the **BERT-base-uncased** model. Table 3 shows the results of this estimation, averaged over 3 experiments. For ease of perception, we only show the top 4 results for each task. As can be seen from the table, for SST-2 and CoLA, the model successfully passed the control task in most cases, as the difference between the real and control estimates is greater than 0.2 F-score in most cases. However, in the case of TruthfulQA with the largest compression rate, the model failed to pass the task, and the weighted F-score was around 0.5, indicating a complete loss of ability to solve the task.

Probing Analysis in Decoder Model

We performed same experiments on the **Llama 2 7b** decoder model. The results are presented in Table 4, which shows the performance of the last four layers of the model. Compared to the encoder model, the decoder model coped better with the control task in conditions of strong compres-

sion. Additionally, for the most challenging TruthfulQA task, the model even under strong compression achieved a result that was higher than random estimation. Furthermore, we generated two graphs for both models: one for each encoder layer, as shown in Figure 3 for SVD factorization, and another for FWSVD in Figure 4.

Figure 1
Comparison of Factorization Methods for CoLA and MMLU

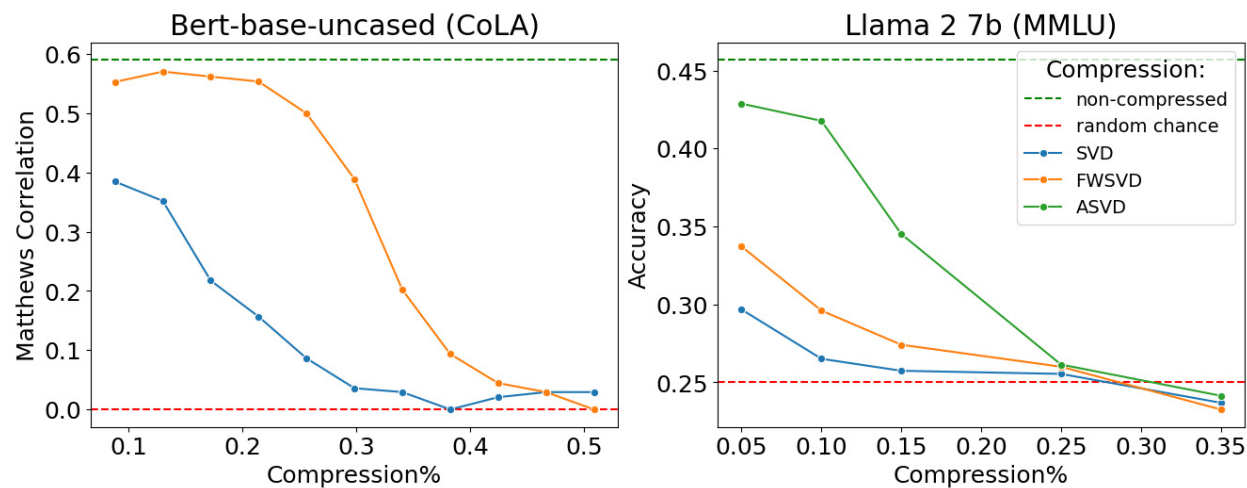


Table 2
Results of Fine-Tuned Models with Different Compression Rate

Llama 2 7b on MMLU						
Compression rate %	0	5	10	15	25	35
SVD	0.456	0.296	0.265	0.257	0.255	0.232
FWSVD	0.456	0.337	0.296	0.274	0.26	0.236
ASVD	0.456	0.428	0.417	0.345	0.285	0.261

BERT-base-uncased on CoLA						
Compression rate %	0	10	20	30	40	50
SVD	0.59	0.384	0.156	0.035	0	0
FWSVD	0.59	0.552	0.553	0.388	0.09	0

Note. 0 compression rate in this case means non-compressed model.

Table 3
Results of the Top 4 Layers of the Encoder **BERT-Base-Uncased** Model with Additional Control Task (control t.) (The Best Compression Results for Each Compression Rate are Highlighted in Bold)

Dataset	CoLA				SST-2				TruthfulQA			
Layer	9	10	11	12	9	10	11	12	9	10	11	12
w\o compress	0.824	0.832	0.832	0.829	0.842	0.851	0.857	0.836	0.747	0.723	0.796	0.778
control t.	0.525	0.435	0.545	0.557	0.456	0.472	0.483	0.424	0.571	0.576	0.602	0.393
SVD 90%	0.765	0.77	0.774	0.765	0.801	0.79	0.791	0.79	0.788	0.787	0.654	0.667
control t.	0.577	0.513	0.571	0.516	0.395	0.413	0.459	0.388	0.608	0.501	0.543	0.407
FWSVD 90%	0.768	0.775	0.767	0.77	0.808	0.794	0.808	0.796	0.687	0.728	0.756	0.61

Dataset	CoLA				SST-2				TruthfulQA			
Layer	9	10	11	12	9	10	11	12	9	10	11	12
<i>control t.</i>	0.468	0.556	0.468	0.551	0.448	0.423	0.442	0.485	0.494	0.449	0.475	0.45
SVD 70%	0.68	0.6	0.62	0.639	0.736	0.731	0.711	0.69	0.71	0.655	0.646	0.694
<i>control t.</i>	0.494	0.542	0.378	0.318	0.46	0.415	0.41	0.398	0.508	0.615	0.478	0.388
FWSVD 70%	0.631	0.652	0.637	0.603	0.698	0.718	0.711	0.716	0.614	0.524	0.636	0.713
<i>control t.</i>	0.561	0.468	0.562	0.495	0.485	0.423	0.396	0.453	0.44	0.57	0.571	0.584
SVD 50%	0.529	0.627	0.451	0.612	0.72	0.718	0.701	0.672	0.583	0.699	0.632	0.562
<i>control t.</i>	0.426	0.451	0.578	0.443	0.44	0.352	0.378	0.432	0.576	0.653	0.524	0.408
FWSVD 50%	0.548	0.443	0.507	0.441	0.736	0.617	0.672	0.507	0.473	0.494	0.347	0.537
<i>control t.</i>	0.428	0.318	0.299	0.431	0.345	0.34	0.351	0.381	0.397	0.673	0.636	0.476

Note. The best compression results for each compression rate are highlighted in bold.

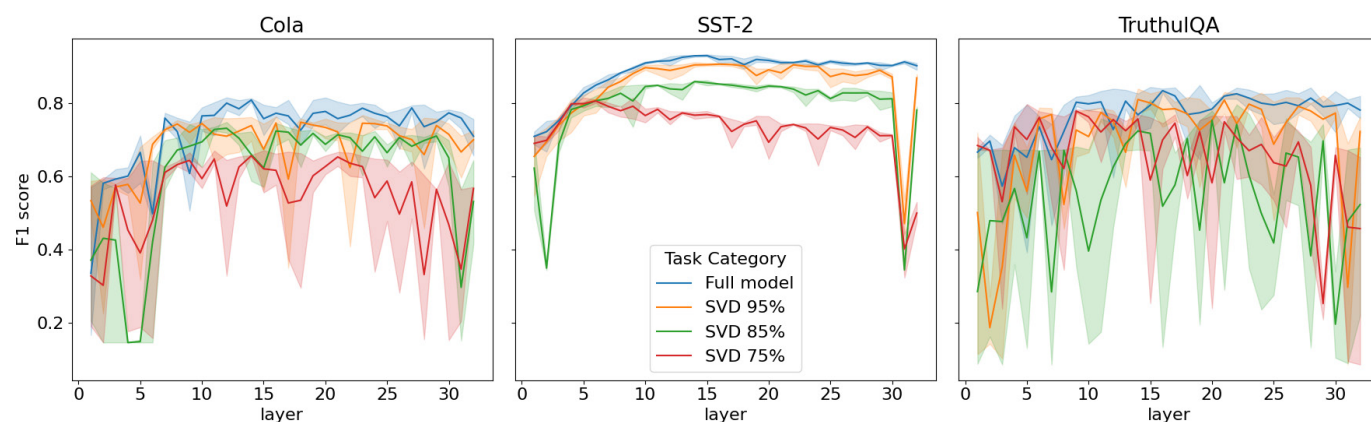
Table 4
 Results of the top 4 Layers of the Decoder *Llama 2 7b* Model with Additional Control Task (*control t.*)

Dataset	CoLA				SST-2				TruthfulQA			
Layer	29	30	32	32	29	30	32	32	29	30	32	32
w\o compress	0.75	0.774	0.76	0.711	0.905	0.904	0.914	0.904	0.791	0.795	0.801	0.782
<i>control t.</i>	0.579	0.563	0.387	0.569	0.396	0.352	0.469	0.417	0.629	0.647	0.6	0.596
SVD 95%	0.74	0.716	0.667	0.701	0.891	0.873	0.471	0.87	0.757	0.774	0.297	0.724
<i>control t.</i>	0.438	0.249	0.401	0.429	0.426	0.436	0.403	0.39	0.604	0.616	0.244	0.602
FWSVD 95%	0.761	0.746	0.758	0.72	0.9	0.893	w.895	0.874	0.785	0.769	0.797	0.779
<i>control t.</i>	0.491	0.505	0.582	0.439	0.453	0.478	0.491	0.403	0.658	0.647	0.583	0.658
ASVD 95%	0.726	0.750	0.768	0.735	0.922	0.920	0.917	0.910	0.798	0.800	0.811	0.786
<i>control t.</i>	0.412	0.378	0.432	0.509	0.357	0.370	0.393	0.385	0.625	0.603	0.606	0.606
SVD 85%	0.711	0.651	0.297	0.532	0.812	0.813	0.345	0.782	0.698	0.196	0.478	0.523
<i>control t.</i>	0.455	0.433	0.565	0.312	0.339	0.423	0.337	0.4	0.431	0.608	0.291	0.546
FWSVD 85%	0.745	0.761	0.757	0.714	0.891	0.9	0.876	0.848	0.795	0.748	0.597	0.535
<i>control t.</i>	0.493	0.451	0.563	0.427	0.472	0.409	0.394	0.38	0.582	0.644	0.631	0.621
ASVD 85%	0.76	0.767	0.771	0.745	0.894	0.908	0.904	0.902	0.808	0.821	0.773	0.776
<i>control t.</i>	0.396	0.42	0.401	0.521	0.441	0.361	0.512	0.399	0.62	0.598	0.602	0.612
SVD 75%	0.565	0.469	0.347	0.567	0.712	0.712	0.402	0.5	0.252	0.658	0.462	0.458
<i>control t.</i>	0.565	0.356	0.574	0.584	0.37	0.35	0.366	0.384	0.432	0.553	0.666	0.648
FWSVD 75%	0.719	0.704	0.71	0.633	0.841	0.843	0.783	0.793	0.498	0.711	0.526	0.336
<i>control t.</i>	0.417	0.462	0.571	0.3	0.507	0.397	0.341	0.336	0.527	0.577	0.672	0.546
ASVD 75%	0.680	0.671	0.673	0.651	0.820	0.815	0.813	0.813	0.777	0.72	0.760	0.750
<i>control t.</i>	0.432	0.421	0.581	0.499	0.390	0.401	0.388	0.410	0.591	0.566	0.561	0.615

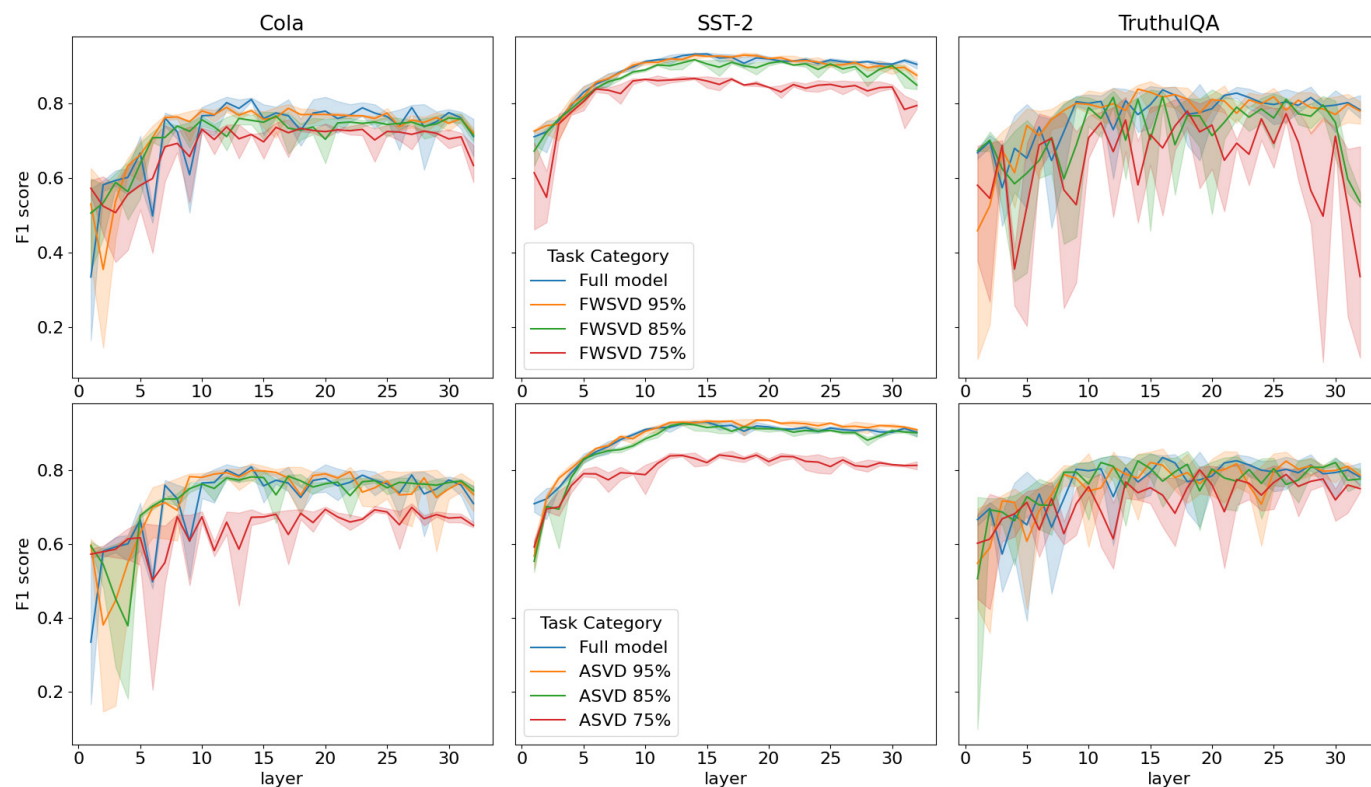
Note. The best compression results for each compression rate are highlighted in bold.

Figure 2

Line Graphs for Each of the Layers of **Llama 2 7b**. Naive SVD is Used as Compression Method

**Figure 3**

Line Graphs for Each of the Layers of **Llama 2 7b**



Note. FWSVD and ASVD is used as compression method.

performance levels. Understanding how compression affects not only the overall quality but also the internal representations within these models is crucial. This knowledge can inform the development of more efficient compression algorithms that preserve essential features necessary for complex task performance.

Our findings demonstrate a strong correlation between compression and loss of model quality, confirming RQ1. This

aligns with previous research indicating that certain model layers are more vulnerable to compression-induced degradation (Chen et al., 2020). However, unlike earlier studies that largely focused on aggregate performance metrics such as overall accuracy or perplexity, our approach delved deeper into the internal representations of models and their behavior at the layer level. By examining both encoder and decoder architectures, we reveal how the internal structure of the model can become less robust as compression inten-

sifies, thus contributing to a more nuanced understanding of how and why quality degrades.

This trend is more significant in the decoder model compared to the encoder model. Different tasks exhibit varying degrees of quality degradation with compression. SST-2 remains consistent across all models, whereas CoLA demonstrates a decline in quality with model variation. This suggests that some layers entirely lose their capacity to generate outputs rather than merely degrading in quality. TruthfulQA, the most challenging task, exhibits the most substantial quality drop, with significant instability between model layers; at high compression levels, it ceases to function effectively, yielding results akin to random sampling. It is evident that compression not only diminishes the knowledge within the compressed layer but also affects other related aspects outside our specific focus. For instance, with a 30% compression rate in the BERT-base-uncased decoder using standard SVD, the model fails to produce the desired results (Figure 1.), demonstrating a correlation quality of 0.03, or an F-measure of 0.5. Still, the model retains some residual knowledge of CoLA, achieving an F-measure of approximately 0.6 on the last four layers, outperforming random responses (Table 2.).

With respect to RQ2, our results show that advanced factorization methods like ASVD and FWSVD improve model quality retention compared to standard SVD. While (Chen et al., 2020) suggested that certain layers are inherently more difficult to compress, our findings expand upon this by demonstrating that selective and refined factorization techniques can mitigate these vulnerabilities. Conversely, for the Llama 2 7b decoder model, ASVD consistently delivers superior results, as evidenced by Table 3 and illustrated in Figures 3 and 4. Notably, even with a maximum compression of 25%, the Llama model retains no more than a 10% quality loss for CoLA and SST-2 tasks, but completely forgets TruthfulQA, resulting in an MMLU benchmark score of 0.285, almost equivalent to random choice (0.25). Furthermore, Figures 2 and 3 highlight a significant difference between ASVD, FWSVD and SVD in relation to SST-2. SVD exhibits a quality drop in the final layers, which is less in FWSVD and absent in ASVD. This capability allows ASVD to achieve superior results for complex tasks. Moreover, we contribute evidence to support and refine the assertions of previous works (Ji et al., 2024; Yuan et al., 2023), who proposed alternative compression approaches but did not fully account for layer-specific sensitivities. By employing ASVD and FWSVD, we illustrate a concrete pathway towards preserving critical internal features that standard SVD often fails to maintain. This deeper analysis and interpretation of the obtained results extends previous works, offering new strategies to better control how compression impacts different parts of a model's internal structure.

Our investigation into RQ3, whether compression leads to irreversible loss of knowledge, provides both confirmation of and contrast to existing literature. Similar to prior studies reporting irreversible degradation in certain architectures or tasks (Sharma et al., 2023), we find that challenging tasks such as TruthfulQA suffer disproportionately under high compression rates. Yet, our layer-wise probing and fine-tuning experiments reveal that not all knowledge is equally affected: while some tasks all but vanish under extreme compression, simpler tasks like SST-2 remain largely intact. This more differentiated picture advances the field's understanding of knowledge retention, suggesting that the vulnerability of knowledge to compression may depend on the complexity and nature of the task, rather than reflecting a uniform process of forgetting.

Compared to previous research, our study delves deeper into the literature by confirming previous findings on the existence of "incompressible" layers (Chen et al., 2020) and expanding the scope by proposing solutions through factorization variants such as ASVD and FWSVD. While our findings do not completely solve the challenge of model compression without loss of accuracy, they represent a significant step towards balancing efficiency and model integrity, pointing to promising avenues for further exploration. For instance, Sharma et al. (2023) highlighted the cumulative impact of noise during compression, an aspect we did not specifically address. This gap suggests potential synergies between our methods and other noise mitigation strategies, encouraging future research that integrates complementary findings to achieve better compression results.

CONCLUSION

This study demonstrates that increased model compression leads to a decrease in both model performance and the quality of hidden representations. This effect is more pronounced in decoder models compared to encoder models. The decrease is dependent on the task and layer, and more complex tasks are more adversely affected by compression.

Our findings highlight the importance of considering the effect of compression on various model architectures and tasks. In particular, we found that the FWSVD method outperformed standard SVD at higher compression rates for encoder models like BERT in terms of preserving model quality. For decoder models like Llama-2 we see a similar picture, but besides FWSVD we can use additionally ASVD which shows even better results. These results suggest that both FWSVD and ASVD can effectively reduce some of the negative effects of compression by improving the compressibility of layers that would otherwise be incompressible. This helps maintain model performance, but irreversible knowl-

edge loss at the layer level continues to be a significant factor leading to performance decline, especially in more complex tasks.

Future research should focus on exploring factors such as noise during compression and developing more advanced compression techniques in order to fully address these issues. Improving methods like ASVD may lead to better preservation of model performance at higher compression rates. In addition, it may be worth to use the probing results

as an estimate and threshold to prepare the model for compression.

DECLARATION OF COMPETING INTEREST

None declared.

REFERENCES

- Belinkov, Y. (2021). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1), 207–219. https://doi.org/10.1162/COLI_a_00422
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are few-shot learners. *Advances in Neural Information Processing Systems*. <https://arxiv.org/abs/2005.14165v4>
- Chen, T., Frankle, J., Chang, S., Liu, S., Zhang, Y., Wang, Z., & Carbin, M. (2020). The lottery ticket hypothesis for pre-trained BERT networks. *Advances in Neural Information Processing Systems*. <https://arxiv.org/abs/2007.12223v2>
- Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2018). Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35(1), 126–136. <https://doi.org/10.1109/MSP.2017.2765695>
- Dettmers, T., Lewis, M., Shleifer, S., & Zettlemoyer, L. (2021). 8-bit Optimizers via block-wise quantization. *ICLR 2022 - 10th International Conference on Learning Representations*, 8, 105–125. Curran Associates, Inc. <https://arxiv.org/abs/2110.02861v2>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1, 4171–4186. Association for Computational Linguistics. <https://arxiv.org/abs/1810.04805v2>
- Ganesh, P., Chen, Y., Lou, X., Khan, M. A., Yang, Y., Sajjad, H., Nakov, P., Chen, D., & Winslett, M. (2021). Compressing large-scale transformer-based models: A case study on BERT. *Transactions of the Association for Computational Linguistics*, 9, 1061–1080. https://doi.org/10.1162/TACL_A_00413
- Guo, Y., Yao, A., & Chen, Y. (2016). Dynamic network surgery for efficient DNNs. *Advances in Neural Information Processing Systems* (pp. 1387–1395). Morgan Kaufmann Publishers Inc. <https://arxiv.org/abs/1608.04493v2>
- Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. *Advances in Neural Information Processing Systems* (pp. 1135–1143). <https://arxiv.org/abs/1506.02626v3>
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *ICLR 2021 - 9th International Conference on Learning Representations* (pp. 1343–1355). OpenReview.net. <https://arxiv.org/abs/2009.03300v3>
- Hewitt, J., & Liang, P. (2019). Designing and Interpreting probes with control tasks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (pp. 2733–2743). Association for Computational Linguistics. <https://doi.org/10.18653/v1/d19-1275>
- Hsu, Y. C., Hua, T., Chang, S. E., Lou, Q., Shen, Y., & Jin, H. (2022). Language model compression with weighted low-rank factorization. *ICLR 2022 - 10th International Conference on Learning Representations*. <https://arxiv.org/abs/2207.00112v1>
- Ji, Y., Xiang, Y., Li, J., Chen, W., Liu, Z., Chen, K., & Zhang, M. (2024). *Feature-based low-rank compression of large language models via bayesian optimization* (pp. 844–857). OpenReview.net. <https://arxiv.org/abs/2405.10616v1>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models*. <https://arxiv.org/abs/2001.08361v1>
- Kim, Y. D., Park, E., Yoo, S., Choi, T., Yang, L., & Shin, D. (2015). Compression of deep convolutional neural networks for fast and low power mobile applications. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*. OpenReview.net. <https://arxiv.org/abs/1511.06530v2>
- Kurtic, E., Campos, D., Nguyen, T., Frantar, E., Kurtz, M., Fineran, B., Goin, M., & Alistarh, D. (2022). The Optimal BERT surgeon: Scalable and accurate second-order pruning for Large Language Models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 4163–4181). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-main.279>
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *8th International Conference on Learning Representations*. Curran Associates, Inc. <https://arxiv.org/abs/1909.11942v6>
- Lane, N. D., Bhattacharya, S., Georgiev, P., Forlivesi, C., Jiao, L., Qendro, L., & Kawsar, F. (2016). DeepX: A software accelerator for low-power deep learning inference on mobile devices. *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks, IPSN 2016 - Proceedings*. IEEE Press. <https://doi.org/10.1109/IPSIN.2016.7460664>

- Li, K., Patel, O., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Inference-Time intervention: Eliciting truthful answers from a Language Model. *Advances in Neural Information Processing Systems*, 36. <https://arxiv.org/abs/2306.03341v6>.
- Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1, 3214–3252. <https://doi.org/10.18653/V1/2022.ACL-LONG.229>
- Michel, P., Levy, O., & Neubig, G. (2019). Are sixteen heads really better than one? *Advances in Neural Information Processing Systems*, 32. <https://arxiv.org/abs/1905.10650v3>
- Narayanan, D., Phanishayee, A., Shi, K., Chen, X., & Zaharia, M. (2020). Memory-efficient pipeline-parallel DNN training. *Proceedings of Machine Learning Research*, 139, 7937–7947.
- Sharma, P., Ash, J. T., & Misra, D. (2023). The truth is in there: Improving reasoning in Language Models with layer-selective rank reduction. *12th International Conference on Learning Representations*. OpenReview.net <https://arxiv.org/abs/2312.13558v1>
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). *Recursive deep models for semantic compositionality over a sentiment treebank* (pp. 1631–1642). ACM. <https://aclanthology.org/D13-1170>
- Tai, C., Xiao, T., Zhang, Y., Wang, X., & Weinan, E. (2015). Convolutional neural networks with low-rank regularization. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*. arXiv:1511.06067. <https://doi.org/10.48550/arXiv.1511.06067>
- Tang, R., Lu, Y., Liu, L., Mou, L., Vechtomova, O., & Lin, J. (2019). *Distilling task-specific knowledge from BERT into simple neural networks*. <https://arxiv.org/abs/1903.12136v1>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv:2302.13971. <https://doi.org/10.48550/arXiv.2302.13971>
- Touvron, H., Martin, L., Stone, K.R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D.M., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A.S., Hosseini, S., Hou, R., Inan, H., Kardaş, M., Kerkez, V., Khabsa, M., Kloumann, I.M., Korenev, A.V., Koura, P.S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kam-badur, M.H., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., & Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models*. <https://arxiv.org/abs/2307.09288v2>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. arXiv:1706.03762. <https://doi.org/10.48550/arXiv.1706.03762>
- Wang, N., Choi, J., Brand, D., Chen, C. Y., & Gopalakrishnan, K. (2018). Training deep neural networks with 8-bit floating point numbers. *Advances in Neural Information Processing Systems*. arXiv:1812.08011. <https://doi.org/10.48550/arXiv.1812.08011>
- Wang, Z., Wohlwend, J., & Lei, T. (2019a). Structured pruning of Large Language Models. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 6151–6162. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.496>
- Wang, Z., Wohlwend, J., & Lei, T. (2019b). Structured pruning of Large Language Models. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 6151–6162. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.496>
- Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7, 625–641. https://doi.org/10.1162/TACL_A_00290
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent abilities of Large Language Models*. <https://arxiv.org/abs/2206.07682v2>
- Xu, C., Yao, J., Lin, Z., Ou, W., Cao, Y., Wang, Z., & Zha, H. (2018). Alternating multi-bit quantization for recurrent neural networks. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. OpenReview.net. <https://arxiv.org/abs/1802.00150v1>
- Yin, L., Jaiswal, A., Liu, S., Kundu, S., & Wang, Z. (2023). *Pruning small pre-trained weights irreversibly and monotonically impairs “difficult” downstream tasks in LLMs*. <https://arxiv.org/abs/2310.02277v2>
- Yu, H., & Wu, J. (2023). Compressing transformers: Features are low-rank, but weights are not! *AAAI Conference on Artificial Intelligence*, 37, 11007–11015. <https://doi.org/10.1609/AAAI.V37I9.26304>
- Yuan, Z., Shang, Y., Song, Y., Wu, Q., Yan, Y., & Sun, G. (2023). *ASVD: Activation-aware singular value decomposition for compressing Large Language Models*. <https://arxiv.org/abs/2312.05821v4>
- Zafir, O., Larey, A., Boudoukh, G., Shen, H., & Wasserblat, M. (2021). *Prune once for all: Sparse pre-trained Language Models*. arXiv:2111.05754. <https://doi.org/10.48550/arXiv.2111.05754>
- Zhang, T., Lin, Z., Yang, G., & De Sa, C. (2019). QPyTorch: A low-precision arithmetic simulation framework. *Proceedings - 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing* (pp. 10–13). Curran Associates Inc. <https://doi.org/10.1109/EMC2-NIPS53020.2019.00010>

APPENDIX

As a verification of our conclusions in the main paper, we performed more experiments with a more modern version of llama: llama 3.1. As factorization methods, we use the standard SVD and ASVD, which has performed well in LLama 2 compression.

Figure 4
Line Graphs for Each of the Layers of **Llama 3.1 8b**. Naive SVD and ASVD are Used as Compression Method

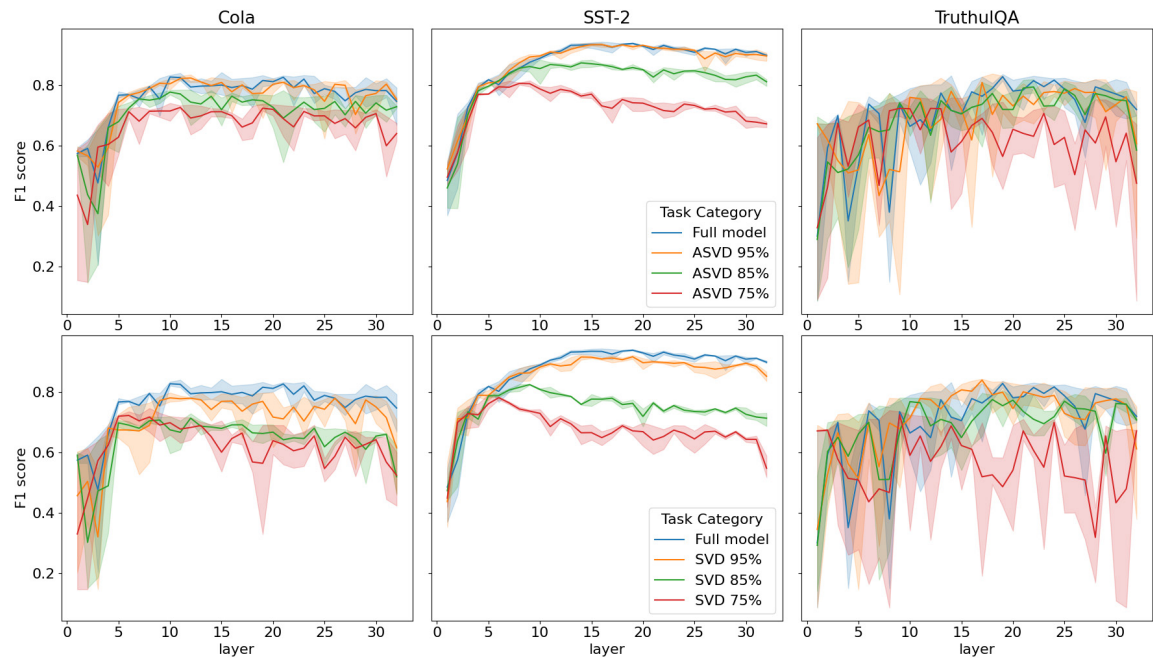


Table 5
Results of the Top 4 layers of the Decoder **Llama 3.1 8b** Model with Additional Control Task (control t.) T(he best compression results for each compression rate are highlighted in bold)

Dataset	CoLA				SST-2				TruthfulQA			
Layer	29	30	32	32	29	30	32	32	29	30	32	32
w\o compress	0.787	0.783	0.783	0.747	0.920	0.909	0.913	0.899	0.783	0.771	0.758	0.720
control t.	0.579	0.563	0.387	0.569	0.396	0.352	0.469	0.417	0.629	0.647	0.6	0.596
SVD 95%	0.774	0.751	0.715	0.615	0.888	0.896	0.885	0.853	0.772	0.778	0.759	0.612
control t.	0.438	0.249	0.401	0.429	0.426	0.436	0.403	0.39	0.604	0.616	0.244	0.602
ASVD 95%	0.766	0.773	0.804	0.754	0.906	0.902	0.902	0.898	0.712	0.737	0.762	0.607
control t.	0.412	0.378	0.432	0.509	0.357	0.370	0.393	0.385	0.625	0.603	0.606	0.606
SVD 85%	0.610	0.654	0.660	0.519	0.747	0.726	0.718	0.713	0.597	0.762	0.760	0.708
control t.	0.455	0.433	0.565	0.312	0.339	0.423	0.337	0.4	0.431	0.608	0.291	0.546
ASVD 85%	0.708	0.742	0.718	0.729	0.819	0.829	0.834	0.812	0.766	0.752	0.749	0.685
control t.	0.396	0.42	0.401	0.521	0.441	0.361	0.512	0.399	0.62	0.598	0.602	0.612
SVD 75%	0.627	0.642	0.568	0.524	0.668	0.643	0.643	0.547	0.655	0.433	0.479	0.672
control t.	0.565	0.356	0.574	0.584	0.372	0.35	0.366	0.384	0.432	0.553	0.666	0.648
ASVD 75%	0.694	0.706	0.6	0.640	0.714	0.681	0.678	0.672	0.696	0.547	0.641	0.476
control t.	0.432	0.421	0.581	0.499	0.390	0.401	0.388	0.410	0.591	0.566	0.561	0.615

In a result, we see a similar pattern to that observed in the research with Llama 2 7b: TruthfulQA probing performs poorly with SVD, and much better with AVD. It is also noticeable that Llama 3.1 is much less compressible, as we see a rapid drop in quality on SST-2 when compressed. At the same time, a small compression of 5% under ASVD has virtually no effect on a simpler dataset such as SST-2 and CoLA. From this we can conclude that our study is scalable to other LLM models.

<https://doi.org/10.17323/jle.2024.24030>

A BERT-BASED CLASSIFICATION MODEL: THE CASE OF RUSSIAN FAIRY TALES

Valery Solovyev ¹, Marina Solnyshkina ¹, Andrey Ten ², Nikolai Prokopyev ³

¹ Kazan Federal University, Kazan, Russia

² Nobilis.Team, Kazan, Russia

³ TAS Institute of Applied Semiotics, Kazan, Russia

ABSTRACT

Introduction: Automatic profiling and genre classification are crucial for text suitability assessment and as such have been in high demand in education, information retrieval, sentiment analysis, and machine translation for over a decade. Of all kinds of genres, fairy tales make one of the most challenging and valuable objects of study due to its heterogeneity and a wide range of implicit idiosyncrasies. Traditional classification methods including stylometric and parametric algorithms, however, are not only labour-intensive and time-consuming, but they are also struggling with identifying corresponding classifying discriminants. The research in the area is scarce, their findings are still controversial and debatable.

Purpose: To fill this crucial void and offers an algorithm to range Russian fairy-tales into classes based on the pre-set parameters. We present the latest BERT-based classification model for Russian fairy tales, test the hypothesis of BERT potential for classifying Russian texts and verify it on a representative corpus of 743 Russian fairy tales.

Method: We pre-train BERT using a collection of three classes of documents and fine-tune it for implementation of a specific application task. Focused on the mechanism of tokenization and embeddings design as the key components in BERT's text processing, the research also evaluates the standard benchmarks used to train classification models and analyze complex cases, possible errors and improvement algorithms thus raising the classification models accuracy. Evaluation of the models performance is conducted based on the loss function, prediction accuracy, precision and recall.

Results: We validated BERT's potential for Russian text classification and ability to enhance the performance and quality of the existing NLP models. Our experiments with cointegrated/rubert-tiny, ai forever/rubert-base, and DeepPavlov/rubert-base-cased-sentence on different classification tasks demonstrate that our models achieve state-of-the-art results with the best accuracy of 95.9% in cointegrated/rubert-tiny thus outperforming the other two models by a good margin. Thus, the achieved by AI classification accuracy is so high that it can compete with that of human expertise.

Conclusion: The findings highlight the importance of fine-tuning for classifying models. BERT demonstrates great potential for improving NLP technologies and contributing to the quality of automatic text analysis and offering new opportunities for research and application in a wide range of areas including identification and arrangement of all types of content-relevant texts thus contributing to decision making. The designed and validated algorithm can be scaled for classification of as complex and ambiguous discourse as fiction thus improving our understanding of text specific categories. Considerably bigger datasets are required for these purposes.

KEYWORDS:

Machine learning, Bert model, fairy tales, Text classification, Neural networks

Citation: Solovyev, V., Solnyshkina, M., Ten, A., & Prokopyev, N. (2024). A BERT-Based classification model: The case of Russian fairy tales. *Journal of Language and Education*, 10(4), 98-111. <https://doi.org/10.17323/jle.2024.24030>

Correspondence:

Valery Solovyev,
maki.solovyev@mail.ru

Received: November 21, 2024

Accepted: December 16, 2024

Published: December 30, 2024

INTRODUCTION

Natural language processing (NLP) is an important field of research that plays a key role in the development of artificial

intelligence. Text understanding and text generation as constituents of NLP have a wide range of applications, including information retrieval, sentiment analysis, machine translation, etc. However,



the existing NLP methods still fail to process context, logical links, lexical chains and detect relationships between parts in a text. The latter refers to both implicit and explicit discourse relations and scholars admit that even hybrid approaches, which combine deep learning and traditional methods, struggle at tasks that heavily involve an understanding of the ways in which entities are connected (Santoro et al., 2018).

Neural network models, especially those based on the Transformer architecture (see Gerasimenko, 2022), have been significantly improving NLP models since the first BERT-based model was designed and developed. Among them BERT, i.e. Bidirectional Encoder Representations from Transformers, presented by Google researchers, stands out among others due to its being conceptually simple and empirically powerful. Designed and developed to pre-train deep bidirectional representations, in most cases BERT models are fine-tuned with only one additional output layer and as such function as state-of-the-art models (see Devlin et al., 2018). The range of applications of BERT models is eminently wide including sentiment analysis, fraud and fake news detection, question-answering systems, document and text classification, information extraction etc. (Rasmy et al. 2021, Atagün et al. 2021, Wang et al. 2020, Jwa et al. 2019, Sun et al., 2019).

A group of widely used BERT models are Masked Language Models, or MLMs, trained to reconstruct the missing tokens which were “masked out”, i.e. missed, from the subset of the input text. The training process implies restoring the missed (masked) tokens/words, during which the model learns to generate words in the text taking context into account (Fu et al., 2022). One of reasons BERT, as a pre-trained masked language model, is currently widely used is due to its ability to learn contextualized word representations from large unannotated corpora and restore the masked out fragments (Lai et al., 2020). The success of those models is often attributed to their ability to capture complex syntactic and semantic characteristics of words (Peters et al., 2018).

BERT is viewed as the gold standard for text processing. BERT-based models vary markedly in the number of neurons and parameters. Cointegrated/rubert-tiny is a small model with only 11.8 million parameters included in the well-known HuggingFace’s Transformers library (github.com/huggingface/transformers). The full credit of identifying advantages of cointegrated/rubert-tiny over other 10 BERT-based models goes to Bolshakov, V., Kolobov, R., Borisov, E., Mikhaylovskiy, N., and Mukhtarova, G. (2023) who argue that it demonstrated a good balance of accuracy and speed of calculations while processing sentences. The model is highly recommended for quick calculations of small datasets (Tomilov et al., 2024).

We hypothesize that (1) classification of overlapping classes of texts such as Russian fairy tales is a cognitively complex task and (2) its automated classification could be performed using BERT with its enhanced categorizing abilities.

Although the latter were demonstrated on the datasets of well-resourced languages such as English (Tangherlini & Chen 2024), French (see Martin et al., 2019, Bayer et al., 2021), German (Chan et al., 2020, Labusch et al., 2019, Leitner et al., 2020), automation of Russian fairy tales classification, to the best of our knowledge, presents a research problem.

Thus aim of this paper is to demonstrate BERT’s potential in the task of classifying Russian folk tales and verify it on a representative corpus of 743 Russian fairy tales.

LITERATURE REVIEW

Fairy Tales as a Genre

Fairy tales make up a unique genre of literature with specific schemata and style. Nevertheless researchers note that fairy tales often contain recurring motifs, archetypes, and plots, which make them a mysterious black box to investigate. Classifications of fairy tales are numerous and based on various features: “leading conflict”, motif, main characters, etc. The generally accepted ATU or Aarne–Thompson–Uther Index (Aarne, 1910, Uther, 2004) ranges tales into 5 sections ((1)Animal Tales, (2) Ordinary Folk Tales including Fairy Tales, Religious Tales, Realistic Tales or Novellas, Tales of the stupid Ogre, Giant or Devil, (3) Anecdotes and Jokes, (4) Formula Tales, (5) Unclassified Tales) with an AT number for each entry. The definition of a tale type although published by Thompson in 1928, i.e. after releasing the first AT catalogue in 1910, lacks the main classifying principle. Later fairy tales were classified based on the basis of narrative plots, characters, motifs, etc, but in all cases the catalogues contain numerous exceptions, overlaps of the identified classes, and intersections. Even the generally accepted classification of fairy tales of A. Aarne when revised by N. P. Andreev was downsized to three, i.e. Animal Tales, Magic Tales and Household or Realistic Tales (see Tudorovskaya, 1961). Nevertheless, in the preface to his “Index”, Andreev (1929) notes that the accepted classification has a number of shortcomings as the division is always relative and not plausible and the principles of division applied are diverse.

In a bid to overcome the challenges faced, researchers point out the so-called “hard core” and “soft shell” of the fairy tale genre. While the former comprises “classic animal tales” or “tales of magic”, the latter is made up of the fairy-tales which may be categorized differently based on one selected parameter. Besides, a narrative, i. e. a plot, may move across genres, acquiring features from a variety of narratives it encounters along its route. All the above indicates that fairy-tale classification is an interesting though extremely laborious and demanding object for automatic classification and analysis (Pompeu, 2019). The above is probably the reason why classification studies of fairy tale texts are relatively rare, although its number has been growing lately (Tangherlini & Chen, 2024).

Text Classification Analysis

Text classification is one of the classic tasks in computational linguistics with an important practical applications including recommender systems which analyze and categorize texts, scan them for user's specific interests, etc. (Kupriavov et al., 2023, Solnyshkina et al., 2024, Reusens et al., 2024). As early as 1997, Kessler, Numberg and Schütze proposed to classify "genres as bundles of facets, which correlate with various surface cues, and argued that genre detection based on surface cues is as successful as detection based on deeper structural properties" (Kessler et al., 1997, p.32). Samothrakis & Fasli (2015) applied machine learning methods to classify fiction from Project Gutenberg collection into six genres, i.e. "science fiction", "horror", "western", "fantasy", "crime fiction", "mystery". The algorithm comprised extraction of relevant information with the help of Natural Language Toolkit and measurement emotional content in each sentence with Wordnet-Affect. The research emphasis was on the analysis of emotive vocabulary: the authors come to the conclusion that the most distinctive feature discriminating the above-mentioned genres is fear. Three years later Worsham & Kalita (2018) implemented a set of different neural network models and classifiers to identify six genres, i.e. Science fiction, Adventure stories, Historical fiction, Love stories, Detective and mystery stories and Western stories. The authors also used multiple strategies to compensate for the extreme lengths of the documents in the dataset and argued that when trained on the BOW form of the Gutenberg Dataset, XGBoost proved to be "a highly optimized, award winning Gradient Boosting solution" (Worsham & Kalita, 2018: p. 1969). Nowadays data extraction, managing and structuring unstructured data envision using a variety of machine learning techniques (Parida et al., 2021) and deep learning neural networks. BERT marked a new level of research and demonstrated significant improvements over previous models on a variety of NLP tasks including text classification. In 2020 (Batraeva et al., 2020) concluded that convolutional neural networks (CNN) and recurrent neural networks (RNN), have gained the greatest popularity for solving classification problems and are rightly recognized as the most effective. The detailed reviews of implementing neural networks to classification tasks published by (Minaee et al., 2021) and (Reusens et al., 2024) came up with revolutionary findings. Reusens et al. (2024) argue that BiLSTM is the overall best-ranked method which significantly outperforms all other methods except LR TF-IDF, and RoBERTa with a confidence level of 95%.

English has always been the most widely studied and resourced language, however scholars worldwide set themselves the task of conducting in-depth genre classification studies into under-resourced languages such as Russian, Arabic (El-Halees, 2017), Hebrew (Devlin, et al. 2018, Liebeskind et al., 2023) and even non-alphabetic languages, i.e. Chinese (Jin et al., 2020), Korean (Liu et al., 2022) and Japanese (Lippert et al., 2022). As for the choice of text collections,

the research shows that the most studied is news, including fake news. The range of classes comprises topic, emotion, polarity and even sarcasm detection. Although, there is ample research into other text types and discourses, e.g. Barros, Rodriguez and Ortigosa (2013) focus on automatic classification of Spanish poetry by Francisco de Quevedo utilizing emotional content and sentiment categorization.

The past several years, i.e. 2019-2024, witnessed significant advancements in classification of Russian texts that were largely driven by deep learning techniques and transformer-based models (Solovyev et al., 2023, Tomin et al., 2023). BERT is now widely implemented in numerous applications based on Russian datasets such as fiction (detective stories, children's literature, poetry, fantasy, and science fiction), academic discourse (History, Natural Sciences, Medicine and Health, Culture), business, news, research and political discourse, advertisement, tweets, reviews etc. The text collections, tools, and algorithms used for experiments with Russian text classifications differ greatly. For instance, experiments conducted by Dubovik (2017) on texts of four functional styles, i.e. scientific, fiction, business and media with stylometric methods proved extremely successful with F1-measure ranging from 0.7 in media texts to 1.0 in business. Batraeva, Nartsev and Lezgyan (2020) implemented convolutional neural networks (CNN) on the collection of five genres, i.e. history, detective stories, children's literature, poetry, and science fiction, reaching 73.12% of classification accuracy for all 5 classes. Lagutina et al. (2021) report that implementing "rhythmic patterns" to range research articles, advertisement, tweets, novels, reviews and political articles into classes resulted in the highest accuracy (F1=98%) for fiction. Two years later the same group of researchers, using a similar algorithm, accomplished an even more ambitious task classifying novels, articles, reviews, VKontakte posts and OpenCorpora news with even higher accuracy (F1=99%) (Lagutina, 2023). A more challenging task, i.e. taxonomy of ten genres, including Fiction, Fantasy, Detectives, Prose, History. Historical Sciences, Information Technology, Natural Sciences, Medicine and Health, Cooking, Culture. Art was undertaken by (Nikolaev, 2022). The best accuracy of results (F1=71.11%) was obtained after only three epochs of training the neural network.

Fairy Tales Classification

Enabled by available datasets and advances in technology, modern scholars accomplish extremely ambitious tasks of fairytales classification. Among the first in the area were (Nguyen et al., 2012, 2013), who trained classification models, i.e. SVMs (2012) and Learning to Rank methods (2013), for Dutch fairytales. The authors reported a macro-average F_1 score of 0.62 for classifying fairy tales and indicated a high impact of character n-grams. Though the implemented models demonstrated relatively moderate success, they were followed by others. In 2013, Nguyen, Trieschnigg, Meder & Theune designed and developed a fairy tale classifi-

er using Learning to Rank and BM25 queries. The features employed in the study were lexical and story similarity, information retrieval measures, as well as subject-verb-object triplets. The results indicated the highest mean reciprocal rank accuracy of 0.82. In the same year, 2013, Karsdorp and Van den Bosch published “Identifying Motifs in Folktales using Topic Models” in which they argued that that Labeled LDA and Big Document Model produce representations that match relatively well to a manually constructed motif classification system used in folktale research.

Six years later, in 2019, based on the Hierarchical Attention Network (HAN), Pompeu successfully evaluated a cross-language neural network approach on the biggest collection in his dataset, i.e. English subset of the folktales. In 2022 Ostrow reports on the unique model with an overall F1 score of 0.77 able to parse fairy tale characters into Proppian archetypes by tracking their probabilistic association with linguistic occurrences such as adjectives and verbs. The researcher argues that the classification schema enables a broader classification of fairy tales into the types identified by Propp (1984). Thus, the research performed on fairy tales automatic classification fail to develop a reliable and accurate taxonomy achieved in various other tasks of computer linguistics. Besides, the studies conducted in the area use different philological classifications of fairy tales and as such lack a unifying foundation theory. As for Russian fairytales, to the best of our knowledge, they were never utilized for type or genre classification. All the above holds great promises for going beyond the unhelpful traditional approaches to the study of fairy-tales as a genre.

METHOD

Dataset

We selected a dataset of Russian fairytales from the collection of Afanasyev (1982) and sourced it from nukadeti.ru and www.rodon.org/other/rnsoj.htm. The collection comprises three main types of tales, i.e. Magic Tales, Realistic (Household) Tales and Animal Tales, which constitute the core of Afanasyev’s Russian fairy tales collection. The types of tales differ in their plots, themes, and styles thus being suitable for classification tasks. The Dataset information is provided in Table 1 below.

Table 1
Dataset Information

Fairy tales	Words	Sentences	Number of fairy tales
Realistic tales	10,766	1,179	203
Animal Tales	10,754	1,018	342
Tales of Magic	9,371	874	198
TOTAL	30,891	3,071	743

Method

The method of training a neural network implemented in the current research is standard and includes (1) training the network with the texts classified and labeled by experts; (2) reorganization of the network parameters as a result of multiple stages of training and (3) evaluation of accuracy and efficiency on validated datasets.

BERT training involves tuning hyper-parameters, i. e. mini-batch size, number of epochs, learning rate, etc. The loss function (Loss) is viewed as a significant parameter which measures effectiveness of the model to predict the target values compared to the true one. The loss function calculates the model error and is used to update the model parameters during training with the help of gradient descent or other optimization algorithms.

To train the model, the data was split into two sets: the training (df_train) set and the validation (df_val) set in an 80/20 ratio, where 80% of the data is used for training and 20% for validation. The latter enables to test the quality of model generalization functions based on the data which was not used in training.

Below we provide description of parameters, training and testing procedures.

BERT was trained on two simultaneous tasks: generation of missing tokens and prediction of next sentences. BERT also receives tokenized pairs of sentences with masked, i.e. missed, tokens. Thanks to the MLM technique, the network learns a deep bidirectional representation of the language, thus taking into account context of a sentence. The task of predicting the next sentence presents a binary classification task which implies identifying if the second sentence follows the first one. Thanks to this binary classification, network trains to identify relationships between sentences in a text.

Although BERT is generally a bidirectional transformer, in this research we used only the input encoder. The main idea of transformers is to apply an attention mechanism that allows the model to weigh the importance of different parts of the input text for each token processed.

The overall architecture of BERT is illustrated in Figure 1 with a sentence fragment fed as input.

BERT architecture utilizes several types of vector representations (embeddings) that transform text data into numeric vectors.

1. Token Embedding

Each word or sub-word is represented as a unique vector in the embedding space, which is standard practice in modern NLP models. BERT-based models function similarly and imply:

(1) Tokenization Model: BERT implements tokenization using WordPiece algorithm, which splits words into sub-words. This approach helps the model effectively handle rare words and morphological variations. For example, the word “unbelievable” might be split into “un,” “believ,” and “able,” each of which has its own embedding.

(2) Token Vectors: Each token, whether it is a complete word or part of a word, receives a numerical representation—a fixed-length vector (e.g., 768 dimensions for BERT base and 1024 for BERT large). This vector includes semantic information, helping the model understand the meaning and context of words.

(3) Characteristics: The token embedding vector allows the model to understand relationships between words, even if they do not follow each other in a sentence. This is crucial for transformers, which rely on self-attention. Token embeddings allow the model to map out the meaning of words and sub-words, building semantic connections.

Thus, token embedding provides the model with information about the meaning and semantic context of individual tokens in a sentence.

2. Segment Embedding

For BERT, it is of utmost importance to distinguish tokens from different sentences, especially when handling tasks that require understanding of two sentences or text parts. Its works implies the following:

(1) Purpose of Segments: BERT is trained on tasks that require distinguishing the contexts of two sentences, similar to Natural Language Inference (NLI), where it is necessary to determine if two sentences are contradictoty, neutral, or entail each other.

(2) Encoding Segments: Each token receives a special segment embedding that indicates which sentence it belongs to:

- Segment A comprises tokens from the first sentence.
- Segment B comprises tokens from the second sentence (if there is one).

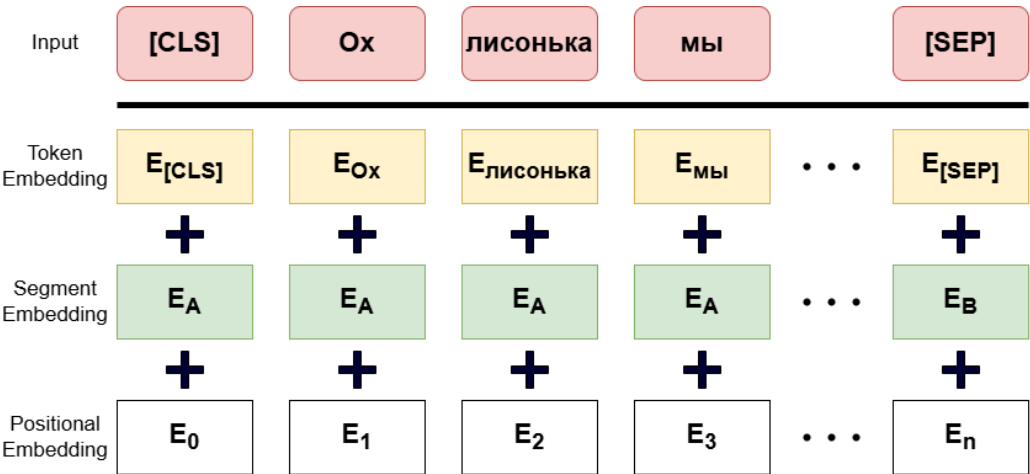
(3) Single-Sentence Input: When the text contains only one sentence, all tokens are assigned to the same segment embedding, meaning they all belong to one sentence. This does not hinder the model’s ability to understand meaning and structure, as segment embeddings simply help in tasks with more than one part of text.

Segment embeddings allow the model to understand not only text-based features of words but also contextual features in two-part tasks, such as question answering and inference tasks.

3. Positional Embedding

Transformers, including BERT, are built on self-attention, where the model can view all tokens at once but does not

Figure 1
BERT architecture



know their positions. Positional embeddings are added to tokens to account for word order.

(1) **Lack of Order Information:** Transformers cannot recognize token sequence on their own as they see all words simultaneously and lack built-in information on sequencing words. This differs BERT from RNNs, which process information sequentially taking order into account.

(2) **Positional Embeddings:** For the purpose of helping the model to differentiate token positions, each token is assigned to a positional vector. Each position is unique and complies with the first, second, etc., positions of tokens in the sentence. These vectors help the model understand the relative position of tokens, which is necessary for accurately capturing structure and order.

(3) **Mathematical Formula:** BERT creates positional embeddings using sinusoidal functions of different frequencies, which allow the model to map out positions at both short and long distances. Each token position has a unique vector based on these sinusoidal functions.

Positional embeddings provide the model with information on each token's position, which is important for maintaining text structure, especially in long sentences.

Input

BERT input is the sum of three embeddings:

Input Embedding = Token Embedding + Segment Embedding + Position Embedding

Each token is represented in the model as the sum of its token, position, and segment embeddings, which together form a fixed-length vector (usually 768 or 1024, depending on the model configuration). Thus, the input vector for each token contains information not only about what the token itself is (Token Embedding), but also about its position in the sentence (Position Embedding) and what segment it belongs to (Segment Embedding). Data classification presupposes proper preparation of the dataset, when each sequence in a text receives a corresponding class label. Hypothetically each document in the collection may belong to more than one class, and as such receives the corresponding number of labels: two in a binary classification or more in a multi-label classification problem.

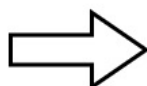
On the next stage, each word is tokenized with PyTorch (pytorch.org/get-started/locally/) (Fig. 2 below) and each sentence is converted into an identifier. Figures 2 and 3 illustrate representation of sentences.

These actions turn the dataset into a list (or Series/Data-Frame object from pandas) of lists. Before BERT processes the dataset, vectors lengths are balanced by padding shorter vectors with an ID of 0.

BERT utilizes a special vocabulary compiled on the pre-training stage. It contains thousands of tokens mapped to a unique identifier. Compiling vocabulary involves using WordPiece algorithm to break words into sub-tokens. The latter allows the model to effectively deal with rare and unknown words by breaking them into smaller parts. BERT also requires specific tokenization which breaks texts into tokens and adds special tokens [CLS] at the beginning and [SEP] at the end of each sequence.

Figure 2
Tokenization

Raw dataset			
category	text		
animal_fairy_tale	Идет волк по лесу. Видит, дятел долбит		
animal_fairy_tale	А дятел волку и говорит: А ты, волк, все		
animal_fairy_tale	А я тебе принесу овец!		
animal_fairy_tale	Согласилась лиса.		
animal_fairy_tale	Вот волк приносит лисе овец: одну, дру		
animal_fairy_tale	Ты нарежь ее и принеси хвост и гриву на		
animal_fairy_tale	Пошел волк и видит лошадь . Подкралс		
animal_fairy_tale	И сейчас по снегу волка косточки блестя		
animal_fairy_tale	Бежала лисица по лесу, увидала на дере		
animal_fairy_tale	Лиса с журавлем подружилась, даже по		
animal_fairy_tale	– Приходи, куманек, приходи, дорогой!		
animal_fairy_tale	Журавль хлоп-хлоп носом, стучал, стуча		
animal_fairy_tale	Лиса начала вертеться вокруг кувшина,		
animal_fairy_tale	Жили курочка с кочетком, и пошли они		
animal_fairy_tale	Вот кинул кочеток орешек, и попал кур		
animal_fairy_tale	Жил кот с кочетком. Кот идет за лыками		

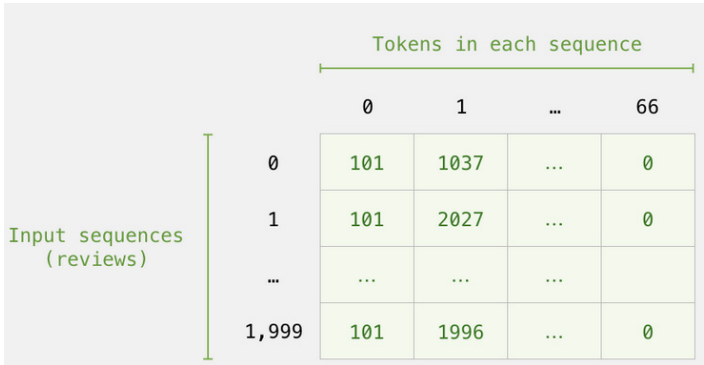


Sequences of Token IDs

[illegible]

Figure 3

Matrix/Tensor for the Neural Network Input



BERT input parameters include the following:

1. Text data: Russian fairy tales.
2. Fairy tales categories: Animal, Realistic (Household), Magical.
3. Tokens: text is tokenized using a pre-trained BERT tokenizer.
4. Token identifiers (input_ids): Numeric representations of words in the text.
5. Attention masks: Specification of tokens to be taken into account.
6. Category labels: Converted to numeric labels for classification.

We operationalize cross-entropy loss or cross-entropy loss function widely used for classification purposes, specifically in neural networks. For multi-class classification we implement the following formula:

$$\text{Loss} = - \sum_{i=1}^n y_i \log(\hat{y}_i)$$

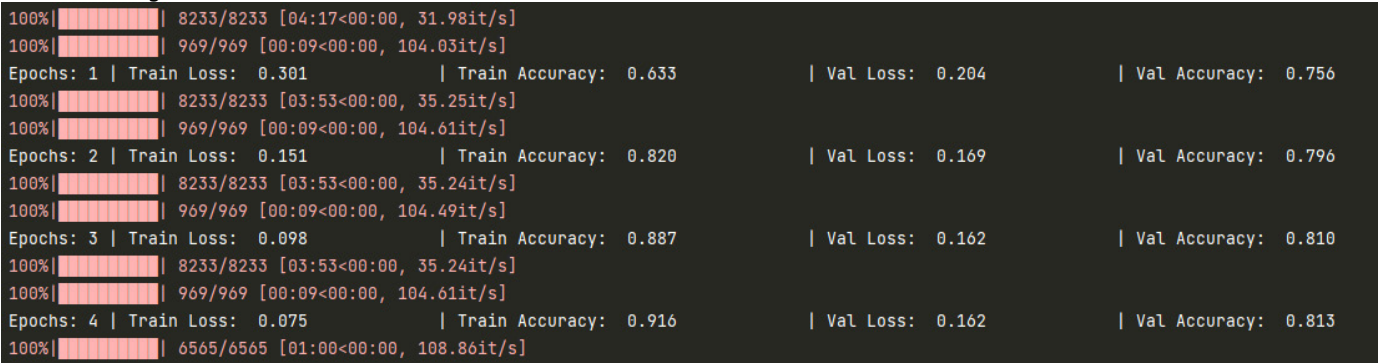
During the training process, the model goes through all the data sets in each epoch. The main steps include the following:

1. *Feeding data to the model:* The data from the training set is fed to the model to make predictions.
2. *Calculating loss and updating model weights:* the loss is measured with the loss function (CrossEntropyLoss), then backpropagation of the error is performed and the model weights are updated using the optimizer.
3. *Printing average loss and accuracy on the training set:* After each epoch, with the purpose to monitor the training process we calculate the average loss and accuracy on the training dataset.
4. *A screenshot of a fragment of the learning process is presented below in Figure 4.*

As a result of all the above steps, we developed an effective system to train the model with BERT neural network aimed at classifying fairy tales. This process includes careful data preparation, tuning the optimizer and scheduler, as well as sequential training and evaluation of the model to achieve high accuracy and performance.

Figure 4

Model Training



Results of the Work and Evaluation of the Model

The models comparison and evaluation were carried out using several generally accepted metrics, i.e. Loss, Accuracy, Precision, Recall. The research algorithm includes comparison of the following models: rubert -tiny, ai -forever/ ruBert -base, DeepPavlov/rubert - base - cased - sentence (cf. Tables 2 – 5). Constituting a characteristic, feature and a parameter, *accuracy* evaluates how correctly the model classifies objects in a set and is measured as the ratio of correct predictions to the total number of predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP (True Positive): Correctly predicted positive classes; TN (True Negative): Correctly predicted negative classes; FP (False Positive): Incorrectly predicted positive classes and FN (False Negative): Incorrectly predicted negative classes.

Precision measures how many of all positive classes predicted by the model are actually positive and is calculated with the following formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where TP (True Positive) is correctly predicted positive classes; FP (False Positive) is incorrectly predicted positive classes.

The results demonstrate, that the optimal number of epochs for the models under study is 5, as it reaches its pick with the batch size of 5 (see Tables 2-5) and when the number of batches increases to 6, the accuracy on the validation set begins to decrease (see Table 5).

In our experiment, the best result received was that by cointegrated / rubert-tiny model with accuracy of 0.875 and minimal losses. However, the model is far from being perfect, and about 12% of the data was classified incorrectly. The erroneous examples from the validation sample are presented in Table 6 below.

Table 2

Models Evaluation with Batch Size 3

Model	Loss	Accuracy	Precision	Recall	Batch size
cointegrated / rubert -tiny	0.098	0.810	0.820	0.815	3
ai-forever / ruBert-base	0.114	0.804	0.810	0.805	3
DeepPavlov / rubert -base-cased-sentence	0.189	0.727	0.735	0.730	3

Table 3

Models Evaluation with Batch Size 4

Model name	Loss	Accuracy	Precision	Recall	Batch size
cointegrated / rubert -tiny	0.075	0.813	0.810	0.815	4
ai-forever / ruBert-base	0.114	0.804	0.800	0.805	4
DeepPavlov / rubert -base-cased-sentence	0.189	0.727	0.730	0.725	4

Table 4

Models Evaluation with Batch Size 5

Model name	Loss	Accuracy	Precision	Recall	Batch size
cointegrated / rubert -tiny	0.054	0.875	0.870	0.870	5
ai-forever / ruBert-base	0.114	0.804	0.800	0.805	5
DeepPavlov / rubert -base-cased-sentence	0.189	0.727	0.730	0.725	5

Table 5

Learning Results in Batch 5 and 6

Batch size	Train Loss	Train Accuracy	Val Loss	Val Accuracy
5	0.008	0.989	0.054	0.875
6	0.008	0.983	0.054	0.863

The model classified the sentences “In the old days there lived a peasant. The peasant had a bee” from the fairy tale “How the Deacon was Treated with Honey” as components of the Class “Fairy Tale about Animals”, although according to the test data set they belong to the Class “Realistic Fairy Tale”. Similarly, the fragment “A passerby says to them: “You, good fellows, instead of nagging and tugging, should get off the cart. Here the horse will ride up the mountain!” from the fairy tale “Seven Stupid Agathons” as a constituent of the Class “Household Fairy Tale” was assigned by the model to the class of “ Fairy Tale of Magic”.

We suggest solving the problem by changing the process of training, since, in the tested version, the model perceives and processes key words and the local not the global context only. It may be also caused either by the class imbalance problem or class overlapping problem in the original dataset. The latter refers to the cases when (in our case) fairy tales from different classes exhibit similar features. The problem is viewed as “one of the toughest problems in machine learning and data mining communities” (see Xiong et al., 2010: p. 491). In situations when texts classification is hampered, it is important and recommended to increase the size of the input data (e.g. for the class of Realistic fairy tales) and retrain the model. The suggested minimum of the size increase is one paragraph.

As it was stated above, we selected 5 batches, but the tensor structure of the input data on everyday fairy tales became 4 – 5 times larger. The results for the best cointegrated / rubert - tiny model with an increase in the input data size to one paragraph only improved Accuracy, precision and Recall significantly (see Table 7).

Table 8 exemplifies class probability of classifying four fairy tales and as we can see, the classification accuracy though high is still below 100%.

DISCUSSION

Classification as one of the main scientific methods applied ubiquitously requires special carefulness when used on works of art. It is caused predominantly by their nature, i.e. the ability to reflect the whole world and encapsulate myriads of ideas. The latter makes the works of art difficult to classify. The current study aims at demonstrating potential of the latest generation of neural networks to solve the abovementioned problem. And the Russian fairy tales embody a model problem as the resulting classification is easy to validate against the indexed catalogue compiled by professional linguists which researchers have at their disposal. The research indicates that BERT-based classification model

Table 6
Classification Errors Generated by Cointegrated / Rubert-tiny Model

oid	text	category
729	V staroye vremya zhil da byl muzhichok. U muzhichka byla pchela Once upon a time there lived a peasant. The peasant had a bee	animal_fairy_tale
594	Vidit, chto ovtsy razbreilis' po polyu, davay ikh lovit' da glaza vydirat'. Vsekh perelovil, vsem glaza vydolbil, sobral stalo v odnu kuchu i sidit sebe radokhonek He saw sheep had wandered off across the field and began catching them and gouging out their eyes. He caught them all, gouged out their eyes, gathered them all into one flock and sat there happily	animal_fairy_tale
497	Da smotri bol'shogo vozu ne nakladyvay, a vpered na menya ne nadeysya: segodnya day da zavtra day, a potom But watch out, don't load a big cart, and don't rely on me in the future: give me today, give me tomorrow, and then	animal_fairy_tale
407	Ne spal, vse barskuyu zagadku otgadyval. Razdumayet, tak i malo li chego na svete ne byvayet, a i to v um pridet: "Mozhet, eto i byvayet, tol'ko ya ne He did not sleep, but tried to guess the master's riddle. Things do happen in the world. And it occurs to him: «Maybe this does happen, but I don't	animal_fairy_tale

Table 7
Rubert-tiny Results for batch size 5

Model name	Loss	Accuracy	Precision	Recall	Batch size
cointegrated / rubert -tiny	0.0 34	0.959	0.915	0.920	5

Table 8*Probability of Belonging to Class*

Genre	Fairy-tale	Class Probability		
		Realistic tales	Animal Tales	Tales of Magic
Realistic Tales	«Porridge from an Axe» by Alexander Afanasyev	0.9281	0.0349	0.0368
Realistic Tales	“Soldier’s Overcoat” by Sergey Saptsov	0.8563	0.0174	0.1260
Animal Tales	“The Crow and the Crawfish” by Konstantin Ushinsky	0.0913	0.7042	0.2045
Tales of Magic	“Geese-Swans” Alexey Tolstoy	0.0667	0.0932	0.8398

demonstrates high accuracy of classifying fairy tales into the three main categories. Below we provide our views of the findings received and research prospects.

We achieved a significantly higher accuracy, i.e. 95%, than Nguyen et al. (2012; 2013) with 82 %. Even though this fact alone does not signify a breakthrough, it marks sustainability and competitiveness of our algorithms. Earlier classifications of other text types resulted in different outcomes largely dependent on the groups categorized. E.g. Lagutina et al. (2021), when ranging into classes such various types of texts as research articles, advertisements, tweets, novels, reviews and political articles, achieved 98% of accuracy. In this regard, classification of fairy tales as one genre into subclasses is a much more difficult task and 98% of accuracy is viewed at the moment as incredible and unattainable. In addition to the main result, i.e. a fairly high percentage of classification accuracy achieved by the neural network on Russian fairy tales, we also obtained a number of auxiliary results potentially useful for further research. Namely: (a) while contrasting three modifications of BERT, cointegrated / rubert -tiny variant performed the best results, b) the optimal number of training epochs proves to be 5 only, c) input data increase, equal to one paragraph only, is prone to higher levels of accuracy. All the above may be viewed as mandatory conditions of elaborated algorithm.

While presenting text classification experimental failures, researchers point out a number of reasons. The first one usually refers either to lack of representativity, insufficiency or misbalance of the training collection, categories or sub-categories of the texts under study (Pompeu, 2019). Similarly, in our case the model performance tends to scale with the number of samples for each category in the collection, which suggests that results may improve if the size of the training data increases. Another reason for misclassification of fairy tales is the above-mentioned “class overlapping problem” (Xiong et al., 2010) when constituents of sub-classes within the class possess very similar characteristics. The latter is very true about fairy-tales “since it is difficult to determine which of the features [in a fairy-tale – *authors’ insert*] is the main one, the task is reduced to assigning the same fairy-tale to two or more classes (groups)” (Andreev, 1929). What we have managed to accomplish in this research is to state

a new problem and a baseline which are open up to further studies.

Prospects of the designed fairy tale classifier lie in the three main directions. First, with a representative dataset and using tales and stories from around the world we plan to pursue comparative classification studies. The latter are of great interest to linguists, historians, cultural scientists and anthropologists and open a vista for further cognitive studies. Second, as fairy tales are viewed in the modern research paradigm as a genre manifesting and transmitting cultural values and as such able of targeting diverse broad audiences, we also plan to implement the designed algorithms into developing a fairy tale profiler with the function of assigning fairy tales for target age and cultural groups. A fairy tale profiler of the kind will provide possibilities to conduct stylometric and multidimensional analysis of fairy tales for specific age groups thus enabling findings and discoveries of cultural and cognitive (dis)similarities of peoples. Third, since there are a number of overlapping genres manifesting features similar to those in fairy tales, i.e. fables, myths, fantasy, etc., experiments with the neural network trained on the three abovementioned types of fairy tales are of the authors’ particular research interest.

Limitations

A standard limitation of neural networks utilization is its dataset or, more specifically, its amount and quality. A relatively small collection of fairy tales used in the current study probably affected accuracy of its classification. Another problem is ambiguity of classification parameters accepted (or ignored) by human experts, but causing fundamental questions: which of the proposed classifications is “correct” (if any), which of them should be used to train the neural network and whether any classification of neural network may be qualified “correct”. The results we obtained are not absolute, though positively relative to the selected classification.

CONCLUSION

Our study highlights significant feasibility of the automatic classification of fairy tales and confirms that further explo-

ration of BERT-based classification model is necessary. BERT represents a substantial advancement in natural language processing due to its ability to provide deep analysis and process context. The present study highlights BERT's significant classification power and effectiveness in developing a taxonomy of Russian folk tales. While pre-trained on representative corpus and fine-tuned for specific tasks BERT is able to accurately classify texts, identifying subtle relationships and contextual features characteristic of Russian folk tales. In particular, models such as cointegrated / ruBERT-tiny, ai forever / ruBERT-base, and DeepPavlov / ruBERT-base-cased-sentence, demonstrated high levels of accuracy, with the best accuracy of 95.9% for the cointegrated / ruBERT-tiny model.

BERT classification power opens up broad prospects for further research and applications, however, despite the progress made, there are still open questions and directions for future research, including quality improvement of tokenization and embeddings, as well as adapting the model to different languages and specific tasks. Overall, BERT demonstrates great potential for enhancing NLP technologies and advancing the time of much more sophisticated and intelligent NLP systems. It is a powerful tool that can significantly improve the quality of automated text analysis and offer new opportunities for research and application in a wide range of areas.

Classification problems regarding fairy tales are caused by numerous factors including topical similarity of classification objects, miscellaneousness of their constituents and the lack of universally accepted genre classification. Two more contributions to the above are fuzzy boundaries of

fairy tale as the concept and their ability to be incorporated into bigger genres, e.g. "Master and Margarita", "Monday Begins on Saturday", "The Lord of the Rings". Further research using increasingly powerful AI systems may result in better understanding and conceptualization of fiction. Our findings signify both challenges and prospects in the area.

ACKNOWLEDGEMENT

The research was supported by the RSF grant 24-28-01355 "Genre-discourse characteristics of the text as a function of lexical range".

DECLARATION OF COMPETING INTEREST

None declared.

AUTHOR CONTRIBUTIONS

Valery Solovyev: conceptualization; investigation; methodology; project administration.

Marina Solnyshkina: formal analysis; writing – original draft; funding acquisition.

Andrey Ten: resources; software; visualization.

Nikolai Prokopyev: writing – review & editing.

REFERENCE

- Aarne, A. (1910). *Verzeichnis der Märchentypen* [List of fairy tale types]. *Folklore Fellows' Communications*, (3). Suomalaisen Tiedeakatemian Toimituksia.
- Andreev, N. P. (1929). *Index of fairy-tale plots according to the Aarne System*. Russian Geographical Society.
- Atagün, E., Hartoka, B. & Albayrak A. (2021). Topic modeling using LDA and BERT Techniques: Teknofest example. *6th International Conference on Computer Science and Engineering* (pp. 660–664). Akdeniz University Publisher. <https://doi.org/10.1109/UBMK52708.2021.9558988>
- Barros, L., Rodriguez, P., & Ortigosa, A. (2013). Automatic classification of literature pieces by emotion detection: A study on Quevedo's poetry. *Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 141–146). IEEE. <https://doi.org/10.1109/ACII.2013.30>
- Batraeva, I. A., Nartsev, A. D., & Lezgyan, A.S. (2020). Using the analysis of semantic proximity of words in solving the problem of determining the genre of texts within deep learning", *Tomsk State University Journal of Control and Computer Science*, 50, 14–22. <https://doi.org/10.17223/19988605/50/2>
- Bayer, M., Kaufhold, M.-A., & Reuter, Ch. (2021). *A survey on data augmentation for text classification*. arXiv preprint. arXiv:2107.03158. <https://doi.org/10.48550/arXiv.2107.03158>
- Chan, B., Schweter, S., & Möller, T. (2020). German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6788–6796). International Committee on Computational Linguistics. <https://doi.org/10.18653/v1/2020.coling-main.598>

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. arXiv:1810.04805, <https://doi.org/10.48550/arXiv.1810.04805>
- Dubovik, A.R. (2017). Automatic text style identification in terms of statistical parameters. *Komp'yuternaya lingvistika i vychislitel'nye ontologii*, 1, 29–45. <https://doi.org/10.17586/2541-9781-2017-1-29-45>
- Fu, Z., Zhou W., Xu J., Zhou H., & Li L. (2022). Contextual representation learning beyond Masked Language Modeling. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 2701-2714). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.193>
- El-Halees, A. M. (2017). Arabic text genre classification. *Journal of Engineering Research and Technology*, 4(3), 105–109.
- Gerasimenko, N.A., Chernyavsky, A.S. & Nikiforova, M.A. (2022) ruSciBERT: A transformer language model for obtaining semantic embeddings of scientific texts in Russian. *Doklady Mathematics*, 106 (Suppl. 1), 95–96. <https://doi.org/10.1134/S1064562422060072>
- Jin, Q., Xue, X., Peng, W., Cai, W., Zhang, Y., Zhang, L. (2020). TBLC-rAttention: A deep neural network model for recognizing the emotional tendency of Chinese medical comment. *IEEE Access*, 8, 96811–96828. <https://doi.org/10.1109/ACCESS.2020.2994252>
- Jwa, H. D. Oh, K. Park, J. M. Kang, & H. Lim (2019). exBAKE: Automatic fake news detection model based on Bidirectional Encoder Representations from Transformers (BERT). *Applied Sciences*, 9(19), 4062. <https://doi.org/10.3390/app9194062>
- Karsdorp, F. & Bosch, Van den A. (2013). Identifying motifs in folktales using topic models. *Proceedings of BENE-LEARN 2013* (pp. 41–49). Radboud University. <https://hdl.handle.net/2066/112943>
- Kelodjoue, E., Gouliau, J., & Schwab D. (2022). Performance of two French BERT models for French language on verbatim transcripts and online posts. *Proceedings of the 5th International Conference on Natural Language and Speech Processing* (pp. 88–94). Association for Computational Linguistics. <https://aclanthology.org/2022.icnl-sp-1.10>
- Kessler B., Numberg G. & Schütze H. (1997). Automatic detection of text genre. *Proceedings of the Eighth Conference on European chapters of the Association for Computational Linguistics*. (pp. 32–38). Association for Computational Linguistics. <https://doi.org/10.3115/976909.979622>
- Kupriyanov, R.V., Solnyshkina, M.I. & Lekhnitskaya, P.A. (2023). Parametric taxonomy of educational texts. *Science Journal of VolSU. Linguistics*, 22(6), 80–94. <https://doi.org/10.15688/jvolsu2.2023.6.6>
- Labusch, K., Kulturbesitz, P., Neudecker, C., & Zellhofer, D. (2019). BERT for named entity recognition in contemporary and historical German. *Proceedings of the 15th Conference on Natural Language Processing* (pp. 9–11). Erlangen.
- Lagutina, K. V., Lagutina, N. S., & Boychuk, E. I. (2021). Text classification by genre based on rhythm features. *Modeling and Analysis of Information Systems*, 28(3), 280–291. <https://doi.org/10.18255/1818-1015-2021-3-280-291>
- Lagutina, K. V. (2023). Genre classification of Russian texts based on Modern Embeddings and Rhythm. *Automatic Control and Computer Sciences*, 57(7), 817–827. <https://doi.org/10.3103/S0146411623070076>
- Lai, Y. A., Lalwani, G. & Zhang, Y. (2020). context analysis for pre-trained masked language models. *Findings of the Association for Computational Linguistics* (pp. 3789–3804). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.338>
- Liebeskind, Ch., Liebeskind, Sh., & Bouhnik, D. (2023) Machine translation for historical research: A case study of Aramaic-Ancient Hebrew translations. *Journal on Computing and Cultural Heritage*, 17(2), 1–23. <https://doi.org/10.1145/3627168>
- Leitner, E., Rehm, G., & Moreno-Schneider, J. (2020). A dataset of German legal documents for named entity recognition. *arXiv preprint*. arXiv:2003.13016. <https://doi.org/10.48550/arXiv.2003.13016>
- Lippert, Ch., Junger, A., Golam R., Md., Mohammad Ya., Hasan Sh., Md, & Chowdhury, Md. (2022). *Kuzushiji (Japanese Text) classification*. Technical Report. <https://doi.org/10.13140/RG.2.2.22416.07680>

- Liu, C., Zhao, Y., Cui X. & Zhao, Y. (2022) A comparative research of different granularities in Korean text classification. In *IEEE International Conference on Advances in Electrical Engineering and Computer Applications* (pp. 486–489). CONF-CDS. Publisher. <https://doi.org/10.1109/AEECA55500.2022.9919047>
- Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., Villemonte de La Clergerie, É., Seddah, D., & Sagot, B. (2019). Camembert: A tasty French language model. *arXiv preprint*. arXiv:1911.03894. <https://doi.org/10.18653/v1/2020.acl-main.645>
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2022). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys*, 54(3), 1–40. <https://doi.org/10.1145/3439726>
- Nikolaev, P.L. (2022) Classification of books into genres based on text descriptions via deep learning. *International Journal of Open Information Technologies*, 10(1), 36–40.
- Nguyen, D., Trieschnigg, D., Meder, Th., & Theune, M. (2012). Automatic classification of folk narrative genres. *Proceedings of the KONVENS 2012* (pp. 378–382). ASAI. http://www.oegai.at/konvens2012/proceedings/56_nguyen12w/
- Nguyen, D., Trieschnigg, D., Meder, Th., & Theune, M. (2013) Folktale classification using learning to rank. *Proceedings of the European Conference on Information Retrieval. Lecture Notes in Computer Science* (vol. 7814, pp. 195–206). Springer. https://doi.org/10.1007/978-3-642-36973-5_17
- Ostrow, R. A., (2022). Heroes, villains, and the in-between: A Natural Language Processing approach to fairy tales. *Senior Projects Spring*, 275.
- Parida, U., Nayak, M., Nayak, A.K., (2021) News text categorization using random forest and naive bayes. In *1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology* (pp. 1–4). IEEE. <https://doi.org/10.1109/ODICON50556.2021.9428925>
- Peters, M., E., Neumann, M., Iyyer, M., Gardner, M., Clark, Ch., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. *ArXiv*, abs/1802.05365. <https://doi.org/10.18653/v1/N18-1202>
- Pompeu, D. P. (2019). *Interpretable deep learning methods for classifying folktales according to the Aarne-Thompson-Uther Scheme* [Master's Thesis]. Instituto Superior Técnico.
- Propp, V. (1984). *The Russian fairy tale*. Izd. LSU.
- Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. (2021) Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 4(1), 86. <https://doi.org/10.1038/s41746-021-00455-y>
- Reusens, M., Stevens, A., Tonglet, J., De Smedt, J., Verbeke, W., Vanden Broucke, S., & Baesens, B. (2024). Evaluating text classification: A benchmark study. *Expert Systems with Applications*, 254, 124302. [10.1016/j.eswa.2024.124302](https://doi.org/10.1016/j.eswa.2024.124302)
- Sabharwal, N. & Agrawal, A. (2021). *BERT model applications: Question answering system in hands-on question answering systems with BERT*. Apress eBooks. <https://doi.org/10.1007/978-1-4842-6664-9>
- Samothrakakis, B. S., & Fasli, M. (2015). Emotional sentence annotation helps predict fiction genre. *PloS One*, 10(11), e0141922. <https://doi.org/10.1371/journal.pone.0141922>
- Santoro, A. & Faulkner, R. & Raposo, D. & Rae, J. & Chrzanowski, M. & Weber, Th. & Wierstra, D. & Vinyals, O. & Pascanu, R. & Lillicrap, T. (2018). *Relational recurrent neural networks*. *arXiv*. <https://doi.org/10.48550/arXiv.1806.01822>
- Solnyshkina, M.I., Kupriyanov, R.V. & Shoeva, G.N. (2024). Linguistic profiling of text: Adventure story vs. Textbook. In *Scientific Result. Questions of Theoretical and Applied Linguistics*, 10(1), 115–132. <https://doi.org/10.18413/2313-8912-2024-10-1-0-7> (In Rus).
- Solovyev, V., Solnyshkina, M., & Tutubalina, E. (2023). Topic modeling for text structure assessment: The case of Russian academic texts. *Journal of Language and Education*, 9(3), 143–158. <https://doi.org/10.17323/jle.2023.16604>
- Sun, F., Liu, J., Wu, J., Pei, Ch., Lin, X., Ou, W. & Jiang P. (2019). BERT4Rec: Sequential recommendation with bi-directional encoder representations from transformer. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 1441–1450). Association for Computing Machinery. <https://doi.org/10.1145/3357384.3357895>

- Tangherlini, T. & Chen, R. (2024). Travels with BERT: Surfacing the intertextuality in Hans Christian Andersen's travel writing and fairy tales through the network lens of large language model based topic modeling. *Orbis Litterarum*, 79(6), 519–562. <https://doi.org/10.1111/oli.12458>
- Tianqi, Ch. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Tomin, E., Solnyshkina, M., Gafiyatova, E. & Galiakhmetova, A. (2023). Automatic text classification as relevance measure for Russian school physics texts. In 2023 *16th International Symposium on Embedded Multicore/Many-core Systems-on-Chip* (pp. 366–370). IEEE. <https://doi.org/10.1109/MCSoc60832.2023.00061>
- Tudorovskaya, E.A. (1961). On classification of Russian folk fairy tales. Specifics of Russian folklore genres. *Specificity of genres of Russian folklore: Theses of the report*. Institute of Russian Literature (Pushkin House).
- Uther, H.-J. (2004). The types of international folktales: A classification and bibliography, based on the system of Antti Aarne and Stith Thompson. *Folklore Fellows' Communications* (vol. 3, pp. 284–286). Suomalainen Tiedekatemia.
- Thompson, S. (1928). The types of the folk-tale: A classification and bibliography. *Folklore Fellows' Communications*, (74). Suomalainen Tiedekatemia.
- Thompson, S. (1977). *The folktale*. University of California Press.
- Wang, Z., Wu, H. Liu, H. & Cai, Q.-H. (2020). BertPair-networks for sentiment classification. *2020 International Conference on Machine Learning and Cybernetics* (pp. 273–278). IEEE Xplore. <https://doi.org/10.1109/ICMLC51923.2020.9469534>
- Worsham, B. J., & Kalita, J. (2018). Genre identification and the compositional effect of genre in literature. *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1963–1973). Association for Computational Linguistics. <https://aclanthology.org/C18-1167>
- Xiong, H. & Wu, J. & Liu, L. (2010). Classification with ClassOverlapping: A systematic study. *1st International Conference on E-Business Intelligence* (pp. 303–309). Atlantis Press. <https://doi.org/10.2991/icebi.2010.43>

DATASET SOURCES

Narodnye Russkie Skazki (Engl. Russian Folklore Tales) from the collection by A. N. Afanasyev, Moscow, Pravda Publishing House, 1982.

Russian folk tales, Moscow, "Fiction", 1965.

Nukadeti.ru

www.rodon.org/other/rnsj.htm

<https://doi.org/10.17323/jle.2024.22272>

Fighting Evaluation Inflation: Concentrated Datasets for Grammatical Error Correction

Vladimir Starchenko¹, Darya Kharlamova¹, Elizaveta Klykova², Anastasia Shavrina¹, Aleksey Starchenko¹, Olga Vinogradova², Olga Lyashevskaya^{1,3}

¹ HSE University, Moscow, Russia

² independent researcher

³ Vinogradov Russian Language Institute, Russian Academy of Sciences, Moscow, Russia

ABSTRACT

Background: Grammatical error correction (GEC) systems have greatly developed over the recent decade. According to common metrics, they often reach the level of or surpass human experts. Nevertheless, they perform poorly on several kinds of errors that are effortlessly corrected by humans. Thus, reaching the resolution limit, evaluation algorithms and datasets do not allow for further enhancement of GEC systems.

Purpose: To solve the problem of the resolution limit in GEC. The suggested approach is to use for evaluation concentrated datasets with a higher density of errors that are difficult for modern GEC systems to handle.

Method: To test the suggested solution, we look at distant-context-sensitive errors that have been acknowledged as challenging for GEC systems. We create a concentrated dataset for English with a higher density of errors of various types, half-manually aggregating pre-annotated examples from four existing datasets and further expanding the annotation of distant-context-sensitive errors. Two GEC systems are evaluated using this dataset, including traditional scoring algorithms and a novel approach modified for longer contexts.

Results: The concentrated dataset includes 1,014 examples sampled manually from FCE, CoNLL-2014, BEA-2019, and REALEC. It is annotated for types of context-sensitive errors such as pronouns, verb tense, punctuation, referential device, and linking device. GEC systems show lower scores when evaluated on the dataset with a higher density of challenging errors, compared to a random dataset with otherwise the same parameters.

Conclusion: The lower scores registered on concentrated datasets confirm that they provide a way for future improvement of GEC models. The dataset can be used for further studies focusing on distant-context-sensitive GEC.

KEYWORDS

Grammatical error correction, L2 errors, ESL, concentrated datasets, cross-sentence GEC

Citation: Starchenko, V., Kharlamova, D., Klykova, E., Shavrina, A., Starchenko, A., Vinogradova, O., & Lyashevskaya, O. (2024). Fighting evaluation inflation: Concentrated datasets for grammatical error correction. *Journal of Language and Education*, 10(4), 112-129. <https://doi.org/10.17323/jle.2024.22272>

Correspondence:
Vladimir Starchenko,
vmstarchenko@edu.hse.ru

Received: August 18, 2024

Accepted: December 16, 2024

Published: December 30, 2024

INTRODUCTION

Grammatical error correction (GEC) is an important task of applied natural language processing (NLP). It involves identifying and correcting errors in word spelling and punctuation, modifying syntactic patterns, as well as suggesting the right word and word order to improve the readability and clarity of text. The definition of the task includes not only detection, classification, and correction of forms and structures that are “strictly grammatical in nature” (Bryant et al.,

2023) but also broader contextual analysis and fluency enhancement that ensure that the correction is consistent with the intended meaning and style of the text (Du & Hashimoto, 2023). GEC technologies can be used to assist children or second language (L2) learners, they can save language teachers’ time, as well as optimize the work of proofreaders, editors, and other specialists dealing with texts.

GEC systems have greatly developed over recent decades. Qorib and Ng



(2022) note that the state-of-the-art GEC models GECToR (Omelianchuk et al., 2020) and T5 (Rothe et al., 2021) exhibit better results than human experts do from the point of view of common metrics, and yet, these systems still fail to detect and/or correct some errors that are easily handled by an educated native speaker. GEC is thus facing the crisis of metric resolution limit: while there is room for growth regarding the observed quality for various types of errors, metrics appear to have reached the ceiling.

At the moment, no solution to this problem has been implemented in the research field. The practice that could pave the way to the solution is to give the scores of a model for various types of errors separately. It allows GEC systems to reveal the more challenging types of errors, but it does not overcome the problem of challenging errors being underrepresented in the existing datasets. Additionally, this practice is scarce in the research (see Yuan & Bryant, 2021; Zhang et al., 2022 as examples) and, crucially, has not been used for further comparison and tuning of models. The present study makes a step towards solving the resolution limit problem.

One of the types of errors which are affected by the resolution limit problem are errors that require information from a distant context (i.e., context broader than one clause) for detection or correction. There is a consensus in the literature that such errors are particularly challenging for models to correct, both for technical reasons (such as the common practice of training models at the sentence level rather than the text level) and due to the difficulty of taking into consideration long-distance dependencies (Chollampatt et al., 2019; Yuan & Bryant, 2021; Qorib & Ng, 2022). Resolution limit makes the advancement of GEC systems with respect to the context-sensitive errors problematic, if common benchmarks and metrics are used.

The present study suggests using evaluation datasets with a higher ratio of the errors which still cause problems for GEC systems. Such datasets are expected to lower the scores of models and allow tuning them for challenging errors. We have selected context-sensitive errors as the material for testing the suggested approach to the resolution limit problem. The concentrated dataset we created comprises 1,014 examples collected from widely used GEC datasets. It consists of manually selected and additionally annotated examples, each containing at least one error that requires distant context for correction. To verify that the concentrated dataset provides higher resolution, we applied two neural networks, BART and T5, to solve the GEC task on the created dataset. We showed that the two GEC systems produce low scores when evaluated across the concentrated dataset, despite the fact that they show competitive results for GEC in general. Thus, creation of the concentrated dataset paves the way for GEC results to grow, as lower (but more accurate) scores make evaluation more distinctive and leave room for improvement. The present study also contributes

to the area of applying machine learning approaches to the problem of wide-context dependency, providing a tool for the evaluation and comparison of models with respect to context-sensitive errors.

LITERATURE REVIEW

GEC Task

Researchers have been trying to improve the results of error correction in texts since the beginning of the computer era. Initially, the practically-oriented studies focused on spelling correction (Cargill, 1980; Bentley, 1985), while GEC in a wider sense was mostly discussed as a preprocessing step for NLP systems that failed to process grammatically incorrect input (Kwasny & Sondheimer, 1981; Jensen et al., 1983). The first GEC tools created for practical use emerged later (Burstein et al. 2003; Leacock et al. 2009, among others), primarily relying on rule-based approaches.

Practically-oriented systems quickly moved to data-driven supervised machine learning designs relying on classification (e.g., Lee 2004; Rozovskaya & Roth, 2010; Dahlmeier & Ng, 2011) and statistical machine translation (SMT) architectures (e.g., Brockett et al., 2006; Yuan & Felice, 2013). A detailed survey of studies dedicated to GEC at this stage can be found in Leacock et al. (2014).

Since that survey, GEC systems have rapidly advanced with the development of deep learning and large language models (LLMs). GEC systems based on various architectures were implemented, including Recurrent Neural Networks (RNN – cf. Yuan & Briscoe, 2016; Xie et al., 2016; Wang et al., 2017), Convolutional Neural Networks (CNN – cf. Chollampatt & Ng, 2018), and Transformers (Edunov et al., 2018; Wang et al., 2019, and many subsequent studies). More detailed discussions of the recent advancement of GEC systems are presented by Wang et al. (2021) and Bryant et al. (2023).

Resolution limit in GEC validation

Modern GEC models seem to have reached the resolution limit: they have previously received even higher scores in terms of common metrics ($F_{0.5}$ proposed for the GEC task by Ng et al., 2014) than human experts (Qorib & Ng, 2022). However, more recent studies (e.g., Zhou et al., 2023; Li & Wang, 2024) claim even further improvements of GEC systems.

We must emphasize two crucial notes at this point. Firstly, when we refer to the “low” scores of annotators, we do not mean those that are caused by the inaccuracies in their annotation. Imperfect annotators’ agreement mostly results from the fact that they choose different options equally suitable for the correction of errors in the original text.

Secondly, the output of GEC systems cannot be claimed to be perfect either. Qorib and Ng (2022, pp. 2795–2797) list several types of errors that GEC models recurrently fail to locate and correct. Among them are such common classes of errors as inaccuracies in syntactic patterns (e.g., subject-verb agreement); errors in long sentences; sentences with high error rates; cross-sentence errors; errors that require paraphrasing of a sentence segment (like correcting phrases that do not sound authentic); and some others. Another notable error type that does not challenge a native speaker, but is repeatedly discussed as being problematic for modern GEC systems, is spelling errors (Chollampatt & Ng, 2018; Starchenko & Starchenko, 2023).

A likely explanation for these pitfalls of GEC systems is that the number of challenging errors in training and evaluation datasets is not large enough to noticeably affect the metrics. Yet the percentage of challenging errors is quite high both in manual data processing and in the application of GEC systems (see, e.g., the discussion of character-level errors in Starchenko & Starchenko, 2023). A possible reason for this is that the corpora used for the training and evaluation of GEC systems are created on the basis of non-native speakers' texts, which are often overloaded with basic grammatical errors.

Some researchers report the evaluation results for different types of errors separately (e.g., Yuan & Bryant, 2021; Zhang et al., 2022), providing special procedures for evaluating the efficiency of the model for errors identified worse than others. This practice is becoming more common, especially after the emergence of the ERRANT scorer (Bryant et al., 2017), which implements separate evaluations for various error types. There are currently no approaches that directly leverage such breakdown evaluation statistics in model enhancement. They are usually presented as a hindsight observation rather than used for tuning a GEC system, while they constitute the material which can be directly used for training and evaluation.

Concentrated Datasets in NLP

Concentrated datasets are successfully applied in various domains of NLP and not only in GEC. One example is the handling of ethics-related biases by LLMs. While these biases are not frequent in the natural data, even a singular appearance in the output greatly impacts the use of such models in commercial practices. As a result, models are additionally fine-tuned and evaluated on concentrated datasets containing biased data (e.g., Nangia et al., 2020; Zhao et al., 2023).

In the GEC domain, concentrated datasets are currently not a widespread tool in evaluation or training. Starchenko (2024) created a concentrated synthetic dataset for fine-tuning an LLM for the GEC task, while Starchenko and Starchenko (2023) proposed a synthetic evaluation dataset. Both

studies are, however, restricted to spelling errors, which are arguably the most basic type of aforementioned challenging errors, naturally allowing for wide-scale synthesizing of “error — correction” pairs. Chollampatt et al. (2019) generated a synthetic dataset with tense errors. To the best of our knowledge, no concentrated dataset of natural language production has been used for GEC. The present study fills this gap by creating and applying a concentrated evaluation dataset consisting of annotated examples from several learner corpora.

Context Dependency in GEC

The problem of context dependency is crucial for GEC. Identifying a grammatical error and suggesting a correction for it is highly dependent on the context, such as parts of speech of neighboring words, their lexical semantics, and word order. Discourse type and the general intentions of the author are also relevant to the way an error is corrected.

Since early on, GEC systems have greatly relied on the context, which has been achieved either by passing some of its features to a model (for classifiers) or by using architectures incorporating context-sensitivity (SMT, RNN, CNN, Transformers). It is, however, the local context around the error that usually receives more attention. Models are often trained for correcting sentences out of context (e.g., the state-of-the-art (SOTA) model by Rothe et al., 2021, pp. 703–704). Moreover, some of the commonly used GEC datasets contain sampled sentences, rather than paragraphs or full texts (cf. Napoles et al., 2017 for a relatively recently released dataset JFLEG). As a result, even the most powerful modern GEC models often fail to correct some types of errors that are more sensitive to the wider context, e.g., pronouns, verb tenses, modality, and usage of discourse markers.

Only a few studies pay special attention to the broader context in GEC. This problem is usually formulated in terms of cross-sentence errors, or “errors that require cross-sentence context to [be] correct[ed]” (Qorib & Ng, 2022). Chollampatt et al. (2019) created a CNN model that includes an additional encoder, preserving information from the previous sentences, and incorporated the encoding in the decoder via attention and gating mechanisms. Yuan & Bryant (2021) compared various Transformer-based architectures by measuring the performance on longer-context-sensitive errors.

METHOD

Working Definition of Context-Sensitive Errors

The narrower practical scope of this paper concerns errors that require taking into account distant context. The most straightforward case of distant-context-sensitive errors are

cross-sentence errors. Consider the discourse presented in (1):

(1) *I go for a walk to a park every day with my two lovely Corgi dogs. I met → [meet] many people in the park.*

The second sentence in (1) is correct when regarded on its own, but from the first sentence in the given context it is clear that the verb in the second sentence cannot be used in Past Simple and must be given in Present Simple. As discussed, it is such examples that are problematic for modern neural networks. Henceforth, we call such errors context-sensitive, subsuming distant rather than local context by the single term “context”.

Notably, context-sensitive errors may also emerge within one sentence. The modification (1)' minimally differs from (1), and the context required for correcting the tense error is practically the same as in (1).

(1)' *I go for a walk to a park every day with my two lovely Corgi dogs, and I met → [meet] many people in the park.*

As discussed later, we show that it is not the sentence borders that make errors challenging for GEC systems; cases like (1)' are also problematic for them. Thus, it is important to include errors that depend on a context which is distant yet is located within the same sentence. This removes a clear-cut border between local and distant context-sensitive GEC, which cannot be set up at the sentence border.

In order to operationalize the annotation of examples for the dataset, we use the following working definition:

Definition: *Errors that cannot be detected or corrected without access to the material from another clause headed by a finite verb are context-sensitive.*

Our dataset thus includes not only cross-sentence errors, but also cross-clause errors. Clearly, this definition excludes some possible (and arguably more debatable) cases of context-sensitive GEC, with the distant context located within the same clause or in a different non-finite clause. What is crucial for the present study is that the suggested definition allows us to include only uncontroversial cases of context-dependent GEC, while not limiting ourselves to cross-sentence examples. We leave a more theoretically-grounded definition of context-sensitive errors for further research.

Creation of a Concentrated Dataset with a Higher Ratio of Context-Sensitive Errors

The concentrated dataset with a higher rate of cross-clause errors is built with the data extracted from existing error-annotated datasets. In this section, we focus on the algorithm of its creation and the characteristics of the four datasets that have formed it, while the resulting features of the dataset are presented in the Results section.

Characteristics of the Non-Concentrated Datasets Used

The concentrated dataset comprises examples annotated for grammatical errors from the following four datasets.

The First Certificate in English (FCE) dataset (Yannakoudakis et al., 2011) contains texts of B1–B2 English learners in the style of a short essay, letter, or description, with each text corrected by one annotator. It is split into training, development, and evaluation subsets.

The CoNLL-2014 dataset (Ng et al., 2014) is a part of the National University of Singapore Corpus of Learner English (NUCLE; Dahlmeier et al., 2013). It was created as an evaluation dataset for the CoNLL-2014 shared task and contains essays of C1 English learners. Different versions of CoNLL-2014 present annotations by 18 different experts.

The Write & Improve (W&I) and LOCNESS (BEA-2019) dataset (Bryant et al., 2019) was created for the BEA-2019 shared task and includes essays by A1–C2 English learners and by undergraduate native speakers. It is split into training, development, and evaluation subsets, with the latter annotated by 5 experts.

These datasets are frequently used for training and evaluation in GEC studies, including various shared tasks (Dale et al., 2012; Ng et al., 2014; Bryant et al., 2019).

The Russian Error-Annotated Learner English Corpus (REALEC) dataset (Vinogradova & Lyashevskaya, 2022) comprises essays of university English learners, most of them at B1–B2 levels of English proficiency. A single annotation approach for each type of error is described in the annotation guide, which has been used by four experts. While this corpus has been released recently and has only been used once in large-scale GEC studies (Volodina et al., 2023), it is particularly useful for the present research, as its annotations include discourse-related error types that are highly relevant for context-dependent GEC.

More detailed information about the datasets is summarized in Appendix A.

For FCE and BEA-2019, only evaluation subsets are taken into consideration. As CoNLL-2014 is an evaluation dataset, and since REALEC has not been actively used for GEC model training yet, there is no expectation that models could learn relevant examples from them during training. Therefore, the concentrated dataset should not be problematic for evaluation in this respect.

Annotation of the Concentrated Dataset

Context-sensitive errors do not have any common features that allow for their easy automatic extraction, have not been

annotated in most existing datasets, and are not frequent. As a result, their extraction from the corpora requires a substantial amount of manual annotation.

To ensure that all the annotations conformed to the same standard and all sentences could be compared regardless of their source, we normalized the annotations for all the datasets by processing the corrections and automatically applying the Error Annotation Toolkit (ERRANT) tags (Bryant et al., 2017) to them. Additionally, we preserved the annotation of discourse-related errors from REALEC.

To make the annotation feasible, we focused on several error types that were expected to be context-sensitive more often (building on suggestions in Bryant et al., 2021). We chose to consider errors with ERRANT tags CONJ (conjunctions), DET (determiners), NOUN:INFL (nominal inflection), PRON (pronouns), PUNCT (punctuation), VERB:SVA (subject-verb agreement), VERB:TENSE (verbal tense), WO (word order), and REALEC tags *Inappropriate_register* (stylistic errors), *Linking_device* (discourse linking tools), and *Ref_device* (usage of anaphoric expressions). For most of these errors, sensitivity to the information in the preceding and/or subsequent context does not have to be explained and is demonstrated by the examples in Appendices B and C. To mention just a couple of the types of such errors: the use of definite article for the first mentioning or of indefinite article for any further mentioning (annotated with DET); the use of predicates in present tenses when there is a reference to the specific time in the past in the context (annotated with VERB:TENSE); etc.

For each of the types, around 50 examples were annotated, and the tags with the highest ratio of context-sensitive errors were chosen for further annotation: ERRANT tags PRON, PUNCT, VERB:TENSE, and REALEC tags *Inappropriate_register*, *Linking_device*, and *Ref_device*. Next, 140–260 examples of each of these error types were annotated. The number of annotated examples and the ratio of context-sensitive errors for each tag are presented in Appendix B. The description of less frequent tags is provided in Appendix C.

For each sentence, the initial tag assigned either by ERRANT or by an annotator (for REALEC) was displayed. An expert from the team of authors had to examine the sentence in context and decide whether it is necessary to take into account information from other clauses or sentences to locate and/or correct the error. For such examples, additional annotation had to be provided:

- whether context from another clause/sentence is required to locate the error;
- whether context from another clause/sentence is required to correct the error;
- the type of context required for locating or correcting the error, namely, whether it is a cross-clause or cross-sentence error;

- the distance in sentences or clauses (if the context is within the same sentence) from the one containing the error (i.e., the number of sentences or clauses that need to be considered to locate and correct the error);
- the direction in which this context is located: to the left, to the right, to any direction or to both directions from the erroneous sentence or clause;
- the type of error (see Appendices B and C).

As a result, we processed and annotated 3,403 errors in the extended context from four corpora of English learner texts and selected a total of 1,014 context-sensitive errors.

Inter-Annotator Agreement

To get a better understanding of the validity of our results, we calculated inter-annotator agreement. For this, we randomly chose 100 sentences representing all the initial error types. All of the sentences were marked up by the four annotators who worked on the whole dataset. We used this subset (henceforth called the agreement dataset) to calculate the inter-annotator agreement for the column “whether context from another clause/sentence is required to locate the error”.

Since the agreement dataset did not have empty values and there were more than two annotators, we used Krippendorff’s Alpha (Krippendorff, 2011), Fleiss’ Kappa (Fleiss, 1971), and Randolph’s Kappa (Randolph, 2005; Warrens, 2010). The main challenge connected with the metric calculation was that the two classes in the dataset were extremely unbalanced: the proportion of context-sensitive errors was relatively small compared to the whole body of errors. This disrupted the estimation of annotator agreement by random chance and resulted in an underestimation of agreement by the commonly used Krippendorff’s Alpha and Fleiss’ Kappa.

The rapid degradation of Krippendorff’s Alpha for the annotation of a small and unbalanced dataset (Marzi et al., 2024) is illustrated in Table 1. The table shows the application of Krippendorff’s Alpha to an imaginary dataset, annotated by three groups of experts. The first group shows perfect agreement; in the second group there is one error in the annotation; and in the third group two experts made one error each. One can see that even one error causes the score to drop to 0.429, and the second error makes it zero, despite the fact that intuitively the annotator’s agreement is relatively high.

To compensate for this, we resorted to using Randolph’s Kappa, which is less affected by class imbalance. Additionally, we provided a custom estimation of agreement: we compared each annotation in the agreement dataset to the annotation that ended up in the main dataset and calculated the percentage of annotators that agreed with the label from the main dataset. After that, we calculated the mean of this percentage across the subset we were working with:

Table 1*Illustration of Rapid Degradation of Krippendorff's Alpha on Unbalanced Datasets*

Expected value ^a	Case of annotation 1 ^b			Case of annotation 2 ^b			Case of annotation 3 ^b		
	A1.1	A1.2	A1.3	A2.1	A2.2	A2.3	A3.1	A3.2	A3.3
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	1	1	2	1
1	1	1	1	1	1	1	1	1	1
K's α^c	1.000			0.429			0.000		

Notes. ^a The correct annotation value expected for the imaginary dataset. ^b Annotations by three groups of experts. The gray cells show the cases of incorrect annotation by an expert. ^c Krippendorff's Alpha.

$$(2) \quad \frac{100}{n} \times \sum_{i=1}^n \frac{\sum_{j=1}^m [item_{i,j} = base_i]}{m}, \quad \text{where}$$

n – the number of datapoints in the dataset,

m – the number of annotators,

$item_{i,j}$ – the annotation by j th annotator for i th datapoint,

$base_i$ – the annotation in the main dataset corresponding to the i th datapoint.

While this method is unconventional, it provides a rough estimate of how well the experts agreed with the annotations that were used in the main dataset, helping to put other inter-rater agreement matrices into perspective.

As the most commonly used scores for Krippendorff's Alpha and Fleiss' Kappa were rather low, we calculated the inter-rater agreement scores for each initial error category to demonstrate what categories were the most and the least reliable. The scores for separate error categories, as well as for the whole agreement dataset, can be found in Table 2.

All four metrics indicate perfect agreement at 1 (or, in our case, 100, since we use percentages). However, each metric is interpreted slightly differently.

Krippendorff's Alpha can be either negative (indicating higher-than-chance disagreement among annotators) or positive (but not exceeding 1). Typically, inter-annotator agreement above 0.67 is considered high enough to be able to draw cautious conclusions based on the annotated data, while agreement above 0.8 is considered robust enough to consider the data reliable. While for our agreement dataset the metrics are not high enough, one should keep in mind that due to the small size of the dataset this metric is likely to show lower agreement than there actually is. Taking this into account, it can be assumed that the "real" agreement score is at least as high as 0.67.

Fleiss' Kappa and Randolph's Kappa are interpreted in almost the same way as Krippendorff's Alpha, with agreement above 0.6 considered substantial and agreement above

0.8 – almost perfect. For our dataset, we are close to the 0.6 threshold for Fleiss' Kappa and above it for Randolph's Kappa. Keeping in mind the fact that this metric is also sensitive to dataset size, it can be assumed that the real agreement is substantial.

As for the custom metric, while there is no conventional interpretation, we can see that its values are quite high, with an average of 9 out of 10 annotations conforming to those found in the main dataset.

Evaluation of Context-Sensitive Errors

The evaluation procedure is crucial for context-sensitive GEC, because its standard implementation in the GEC task leads to consistently lower scores for longer texts, independently of their content.

Score Calculation

The most common measure used for the evaluation of predictive performance is F_β -score. In the GEC task, $F_{0.5}$ -score is used most often, following Ng et al. (2014). It (arguably) represents the judgments of human experts about the quality of text correction (Grundkiewicz et al., 2015; Napoles et al., 2015; Chollampatt & Ng, 2018).

F_β -score is a complex measure that takes into account True Positives (TP, cases in which a model made a correct prediction), False Negatives (FN, cases in which a model did not correct an error it was supposed to), and False Positives (FP, cases in which a model changed the text that does not contain errors). Precision = TP / (TP + FP) and Recall = TP / (TP + FN) are calculated as an intermediate step, and the multiplier $\beta = 0.5$ weights Precision twice as much as Recall. The straightforward interpretation of this metric is as follows: the higher the score, the better the corresponding model performs. That is, a model with a higher $F_{0.5}$ -score accurately corrects more errors and/or does not introduce correction in the fragments of the text that were not annotated as erroneous.

The most recent implementation of an $F_{0.5}$ -scorer is in ER-RANT by Bryant et al. (2017). One of the advantages of this

Table 2
Inter-Annotator Agreement Metrics

Error type	Krippendorff's Alpha	Fleiss' Kappa	Randolf's Kappa	Custom metric ^a	Share of context-sensitive errors in the dataset
All types	55.9	55.9	73.6	89.5	9/100
DET	N/A	N/A	N/A	100	0/18
Inappropriate register	0.0	-5.3	80.0	95.0	0/5
Linking device	60.4	58.3	60.0	90.0	2/5
PRON	35.0	34.0	56.9	82.4	0/17
PUNCT	53.2	52.6	55.0	81.2	2/20
Ref_device	54.8	52.4	60.0	90.0	2/5
VERB:TENSE	74.0	73.5	77.8	90.0	3/15
WO	-1.7	-3.4	86.7	96.7	0/15

Note. ^aMean percentage of annotations conforming to those found in the main dataset.

tool is that it allows for the calculation of scores for each error type separately. The evaluation by ERRANT includes the following steps.

- (1) Preparation. The tool accepts as input a text with errors and a set of versions of this text corrected by experts. It calculates the sets of corrections that must be applied to the original text to obtain the experts' versions of the text, thus yielding the reference corrections. Likewise, the output of a model is compared to the original text, calculating the set of predicted corrections.
- (2) Calculation of $F_{0.5}$ -scores. For each pair of a reference correction set and the prediction correction set, True Positives (TP), False Negatives (FN), and False Positives (FP) are calculated. Based on them, Precision, Recall, and $F_{0.5}$ -score for each expert's annotation is found.
- (3) Choice of the closest annotator. Among the annotations provided by all experts, the one that has the highest $F_{0.5}$ -score is chosen, and TP, FN, and FP of this annotation are selected for this text.
- (4) Iteration over texts. Steps 1–3 are repeated for every text in the dataset, meaning that different texts can be evaluated with respect to different annotators. By default, each sentence in the dataset is treated as a separate text.
- (5) Calculating the final score. TP, FN, and FP received for each text are summarized and used to calculate the $F_{0.5}$ -score for the whole dataset.

A non-trivial property of the described algorithm is the built-in possibility to have more than one annotator for a dataset and the fact that the scorer relies on the closest possible annotation. This property reflects that language allows various ways of expressing the same thoughts, and that there can be various accurate corrections of the same errors. As a result, evaluating just one annotation (without the possibility for one annotator to suggest various corrections) is insufficient for the decision on the model's efficiency. A more substantial discussion may be found in Bryant and Ng (2015).

Relationship between Text Length and $F_{0.5}$ -Score

The outlined algorithm and the way it solves the problem of variability in accurate corrections highly affects the evaluation of context-sensitive errors.

In L2 texts, the density of errors is relatively high (regularly more than one error per sentence). Due to this and to the fact that each error introduces possible variation, the combinatorics of accurate corrections may become complex. Some correct versions of a text generated by a GEC system may not be found in the reference annotations and would be unfairly claimed to be wrong.

To minimize this effect, the texts are split into sentences: the smaller a text fragment fed to a scorer and the fewer errors it contains, the greater the variation accounted for, meaning that the score is more accurate. If text fragments are small enough, one can expect that a large number of annotators cover all combinations of variable corrections within it.

To demonstrate this, we used ERRANT to evaluate the model BART (Katsumata & Komachi, 2020) on the CoNLL-2014 dataset. We calculated two measurements for the same output provided by the model applied at the full text level. The first measurement followed the regular ERRANT workflow, including splitting the dataset into sentences (note that the model is still applied at the text level). The second measurement differed from the first one in that it was conducted for whole texts. The results of these measurements are presented in Table 3. Additionally, we provide the measurement for the model applied to sentences rather than texts, showing that the model handles longer texts more poorly, which is one of the main focuses in this study.

The two measures calculated for the same prediction differ: as discussed, when longer units are considered, the score becomes lower. Notably, provided that the annotations are totally correct, it is the higher score that characterizes the

performance of the model better, making it reasonable to split texts into sentences for evaluation.

This solution, however, is problematic from the point of view of context-sensitive GEC, which by definition requires processing longer contexts. In some cases, context-sensitive errors are located within one sentence, and even though they require context for correction, it does not really affect the measurement. However, this is not always the case.

Firstly, there are errors located at the edge of sentences. The most straightforward example is a punctuation error such as the replacement of a period with a comma or vice versa, but more complicated cases are also possible.

Secondly, some errors are dependent on each other: it may be the case that two errors must be corrected in agreement with each other – like capitalizing the initial letter in the segment that follows the change of a comma for a period. Another regularly occurring example of this kind is sequences of coordinated verbs used in an incorrect tense with the whole sequence depending on the distant left context. For such examples, treating sentences separately is problematic: one annotator could make use of one form (e.g., Past Simple) throughout the whole sequence, while another annotator might choose a different form (e.g., Present Simple). If the sequence is separated into sentences, switching between Past and Present Simple would be erroneously evaluated by the model as correct.

In order to account for these problems, we perform the evaluation of the concentrated dataset in the following way:

- (1) To balance between the necessity of evaluating the shortest text fragments and the possibility of the incorrect treatment of context-sensitive errors, we split the texts into the smallest spans in which sentences with errors dependent on each other are not separated. That is, if an error occurred at the sentence border or its correction required merging two or more sentences, all sentences involved were taken for evaluation.
- (2) We only evaluate the annotated context-sensitive errors, which allows minimizing the distortions caused by enlarging the accessed contexts.
- (3) We preserve only one annotation for context-sensitive errors, provided that our manual annotation did not

reveal large-scale variation (corrections with variation were found for PRON errors, but there are only few such exceptions).

Setup of the Experiment

To test the concentrated dataset, we use it to evaluate two SOTA GEC models: BART (large, Katsumata & Komachi, 2020) and T5 (base, Rothe et al., 2021). We chose these two models over more recently released GEC systems (e.g., Zhou et al., 2023) because the latter generally use one or more LLMs from a standard set, adding supplementary pre- or post-processing components. Provided that the overall result is comparable, we select less complex constructions to obtain a more interpretable result.

RESULTS

Concentrated Dataset of Context-Sensitive Errors

One practical result of the study is the creation of a concentrated dataset with a higher ratio of context-sensitive errors¹. The dataset contains 1,014 context-sensitive errors with additional annotation.

Tables 4–6 present general information about the dataset: the distribution of error types, the representation of the original datasets in the concentrated dataset, and the type of context required for correcting an error.

As shown in Table 4, the concentrated dataset contains 5 main types of context-sensitive errors: pronouns (PRON), punctuation (PUNCT), referential device (REF), verb tense (VERB:TENSE), and linking device (LINK). Other types either have a low ratio of context-sensitive errors and were not extensively annotated (e.g., WO – word order) or emerged accidentally as a result of manually correcting inaccurate tag attribution by ERRANT.

Table 6 demonstrates that in most cases it is the left context that determines how the error must be corrected. Most commonly, only one sentence is enough for correction, but a range of larger distances is represented as well. Only the

Table 3
BART Evaluation, Measurements for Sentences and Texts

Prediction	Measurement	FP	FN	TP	Precision	Recall	F _{0.5}
for texts	for sentences	507	1,589	1,111	68.67	41.15	60.56
	for texts	620	2,131	1,012	62.01	32.2	52.32
for sentences	for sentences	216	978	1,367	86.36	58.29	78.77

¹ <https://huggingface.co/datasets/startc/doc-gec>.

Table 4
Error Types in the Concentrated Dataset

Error type	Number
PRON	259
PUNCT	202
REF	201
VERB:TENSE	171
LINK	140
DET	19
VERB:MODAL	8
Other types	14
Sum	1,014

Table 5
Representation of the Four Datasets in the Concentrated Dataset

Dataset	Number of extracted errors
REALEC	633
BEA-2019	218
FCE	135
CoNLL-2014	28
Sum	1,014

Table 6
Context Required for Correcting an Error

Type of context	# of sentences required for detection or correction	Errors
Left	1	810
	2	64
	3	22
	4	17
	>4	21
Left, sum		934
Right	1	59
	2	1
	>4	1
Right, sum		61
Left or right	1 to left, 1 to right	16
Left and right	1 to left, 1 to right	2
	3 to left, 1 to right	1
Left and right, sum		3
Sum		1,014

right context is required to correct an error in about 1 out of 20 cases, and it is almost exclusively the neighboring sentence. On rare occasions, either left or right context suffices,² and in very few cases, both left and right contexts are required.

Performance of GEC Systems on the Concentrated Dataset

To test the concentrated dataset, we measure the performance of SOTA GEC models BART and T5. Table 7 presents the results of the evaluation.

Before discussing the patterns in the data, we must comment on a feature of the measurement that highly affects the results. The number of False Positives in the table is zero for every row. Consequently, for every row, the Precision equals 100. This directly follows from the measurement

procedure described in the Materials and Methods section: we only evaluate context-sensitive errors and therefore do not access other types of errors, to which False Positives are automatically assigned. To evaluate the number of context-sensitive False Positives properly, one would have to manually process all the False Positives in the output of every model and annotate whether they are dependent on distant context.

As a result, $F_{0.5}$ -scores presented in the table must be treated as the upper bound estimate. The total absence of False Positives in the output of even the best-performing model is outstandingly unlikely, so the actual $F_{0.5}$ -scores are lower. Provided that the 0.5 coefficient of the $F_{0.5}$ -score weighs Precision twice as much as Recall, the $F_{0.5}$ -estimates in Table 7 are significantly more optimistic than they should be, and considering the raw True Positives, False Negatives, and Recall is more relevant.

² Note that if the material necessary for correction can be found in both left and right contexts, but is closer on one side, only the closest context is used. For example, if a clause can be corrected based on the previous sentence or the one located in four sentences to the right, we only take into account the left context.

Table 7*Evaluation of the Concentrated Dataset by BART and T5*

Error type	BART						T5					
	FP	FN	TP	Prec	Rec	F _{0.5}	FP	FN	TP	Prec	Rec	F _{0.5}
All types	0	893	121	100	11.93	40.39	0	986	24	100	2.38	10.85
PUNCT	0	147	55	100	27.23	65.17	0	192	10	100	4.95	20.66
VERB:TENSE	0	142	29	100	16.96	50.52	0	165	4	100	2.37	10.81
PRON	0	239	20	100	7.72	29.5	0	197	4	100	1.99	9.22
REF	0	192	9	100	4.48	18.99	0	256	3	100	1.16	5.54
LINK	0	138	2	100	1.43	6.76	0	140	0	100	0	0

Even with these caveats, the scores in Table 7 are definitely lower than the scores for errors that are not sensitive to distant context. The overall scores of BART are 40.39 for context-sensitive errors vs. 78.04 for non-context-sensitive errors (the latter measured on the CoNLL-2014 dataset); for T5, the measurements are 10.85 and 74.38, respectively.

If we look at the Recall, the poor performance of the models on context-sensitive errors becomes even more noticeable. For PUNCT as the best-handled error, and with BART as the best-performing model, only 27.23% of errors are properly corrected. The other results are even lower, and most values (all of them for T5; REF and LINK for BART) are close to noise. At the same time, the dataset provides distinctive power to observe the difference between the quality of the models' performance: BART consistently shows higher results than T5. This fact confirms that the poor results obtained on the concentrated dataset do not boil down to its inner properties, but reveal imperfections of GEC systems with respect to selected types of errors.

Lastly, it is interesting to point out the pattern in the difference of metrics for different error types. The highest scores are attributed to punctuation, which presents an artificially regulated construction above the writing system, and tense, which is the only purely grammatical type in the sample. Pronouns as anaphoric means represent a more discourse-oriented language domain, yet they are usually discussed as a part of the grammar, while the type Referential_device contains lexically encoded (and less grammar-related) anaphoric means. Lastly, linking phrases are purely in the discourse realm. Thus, one could claim that the quality of error correction decreases with the shift of the error type from grammar to discourse.

DISCUSSION

Studying the evaluation scheme in distant-context-sensitive GEC tasks, we have been able to make several observations. First, we have proved that the dataset with a distribution bias of error types helps to realistically assess the model performance. In fighting inflation in evaluation metrics

obtained on conventional GEC datasets, the concentrated datasets may serve as additional indicators of the models' failures, along with breakdown per-type evaluation reports (Bryant et al., 2017).

Second, we have noticed that the evaluation metrics differ significantly across four subsets stratified according to the data source. As shown in Figure 2, the F_{0.5}-score ranges from 0.61 in BEA-2019 to 0.26 in REALEC, while Recall ranges from 0.24 in BEA-2019 to 0.07 in REALEC. This is in line with other comparative GEC studies based on full test or evaluation datasets (Zhang et al., 2023; Volodina et al., 2023, among others), confirming that the observed variance can be attributed to many factors such as L2 proficiency level, text register, writing task type, text length, sentence length, annotation strategy, and associated differences in the distribution of error tags. Yet, it is necessary to note that the drop in performance across subsets in the concentrated data is clearly more pronounced compared to results observed on non-concentrated datasets.

Third, even though the concentrated dataset is relatively small to be able to draw decisive conclusions, we have observed that error types are associated with the amount of context needed to detect and correct errors. The latter information can be extracted from the dataset annotations as the number of context units (sentences or clauses). For instance, the vast majority of VERB:TENSE errors require no more than one clause (see example (3)), whereas errors tagged as LINK tend to be associated with one or more sentences in the left or right window (see example (4)). Obviously, this affects the overall metrics.

(3) When I was little I *had* → [Ø] tried a lot of sports...

(4) From 2000 the percentage of elderly people in Sweden began to rise to 20 per cent. *Moreover* → [Contrary to that], from 2000 the percentage in the USA was at the same level of 14 per cent.

Further applications of the received results will involve more experiments with different GEC architectures and methods to understand the metric variability across datasets and the role of the available context in models' performance.

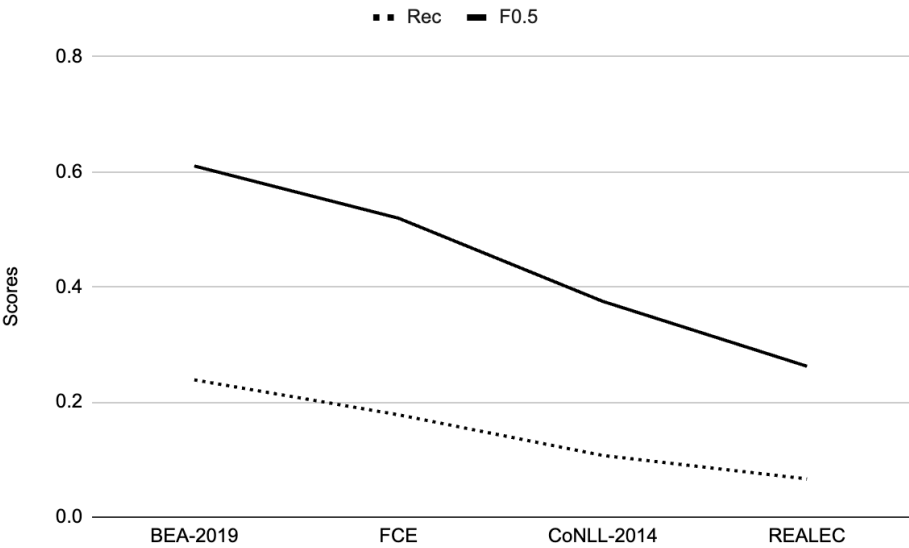
While the difference in the $F_{0.5}$ -scores for the concentrated and non-concentrated datasets are evident, the suitability of this metric for the GEC task remains an open question. With recent advances in generative models prompting, Recall is reported to be equal to, or even greater than, Precision. In this regard, Zeng et al. (2024) suggest using F_1 and F_2 scores as representative metrics in GEC results. As we have shown, $F_{0.5}$, Precision, and Recall calculated for the same model applied to texts vs. separate sentences and measured in text-based vs. sentence-based conditions (see Table 3 above) do not directly correspond to each other. The harmonization of metrics is necessary to establish a consistent benchmark for distant-context-sensitive GEC in various settings.

LIMITATIONS

The nature and key properties of the corpora have to be assessed in the task of compiling the concentrated dataset. Future work may focus on increasing the size of the dataset, balancing the examples with regard to the proficiency level of the authors and to error types, and involving more experts to ensure the robustness of the annotations.

Another limitation of our approach is that the dataset presented in this article is just a preliminary step towards detailed surveys in data curation, evaluation techniques, and model training in the field. We only used off-the-shelf models for evaluation. It is clear that future experiments with training models using concentrated (training) datasets are needed to improve the overall understanding of the role of the error-type bias methods in distant-context-sensitive GEC.

Figure 2
Evaluation Results for BEA-2019, FCE, CoNLL-2014, and REALEC Subsets of the Concentrated Dataset (BART Model)



CONCLUSION

In this study, we propose using a concentrated dataset with a high ratio of context-sensitive errors as a way to solve the resolution limit problem in GEC. This problem arises because the metrics commonly used for evaluating GEC systems may overestimate the model performance, even though certain types of errors are frequently overlooked by these models. By manually annotating examples of various error types (those related to punctuation, verb tense, determiners, pronouns, referential tools, and linking constructions), we have created a dataset containing 1,014 errors that require distant context for identification and/or correction. We have evaluated two GEC models on this dataset and demonstrated that their performance is significantly lower on a concentrated dataset compared to a non-concentrated one. This finding confirms that GEC systems still require substantial improvement and highlights the potential of concentrated datasets as a tool for both training and evaluation.

Based on the performance of the two models across different error types, we hypothesize that error correction becomes more challenging as the error type shifts from the realm of grammar to discourse. For instance, errors in punctuation and verb tense are corrected more successfully than those related to referential and linking devices.

Overall, this article demonstrates the potential of using concentrated datasets with a high ratio of context-sensitive errors to further enhance GEC systems and improve their applicability to real-world tasks. As a practical contribution, we publish the dataset³.

³ <https://huggingface.co/datasets/startc/doc-gec>

ACKNOWLEDGMENTS

This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

CONFLICT OF INTERESTS

None declared.

AUTHOR CONTRIBUTIONS

Vladimir Starchenko: Conceptualization; Data mining; Data curation (automatization); Investigation; Methodology; Model testing; Project administration; Statistics; Supervision; Writing – original draft; Writing – review & editing.

Elizaveta Klykova: Conceptualization; Data curation; Investigation; Methodology; Project administration; Writing – original draft; Writing – review & editing.

Anastasia Shavrina: Conceptualization; Data curation; Investigation; Methodology; Writing – review & editing.

Olga Vinogradova: Conceptualization; Data curation; Investigation; Methodology; Writing – original draft.

Olga Lyashevskaya: Conceptualization; Data curation; Investigation; Supervision; Writing – original draft; Methodology; Project administration; Writing – review & editing.

Darya Kharlamova: Conceptualization; Data curation; Investigation; Methodology; Resources; Writing – original draft; Writing – review & editing.

Aleksey Starchenko: Conceptualization; Investigation; Methodology; Project administration; Statistics; Supervision; Writing – original draft; Writing – review & editing.

REFERENCES

- Bentley, J. (1985). Programming pearls: A spelling checker. *Communications of the ACM*, 28(5), 456–462. <https://doi.org/10.1145/3532.315102>
- Brockett, C., Dolan, B., & Gamon, M. (2006). Correcting ESL errors using phrasal SMT techniques. *21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL* (pp. 249–256). Association for Computational Linguistics. <http://dx.doi.org/10.3115/1220175.1220207>
- Bryant, C., Felice, M., Andersen, Ø. E., & Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications* (pp. 52–75). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/W19-4406>
- Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 793–805). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/P17-1074>
- Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., & Briscoe, T. (2023). Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3), 643–701. http://dx.doi.org/10.1162/coli_a_00478
- Bryant, C., & Ng, H. T. (2015). How far are we from fully automatic high quality grammatical error correction? *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (vol. 1: Long Papers, pp. 697–707). Association for Computational Linguistics. <http://dx.doi.org/10.3115/v1/P15-1068>
- Burstein, J., Chodorow, M., & Leacock, C. (2003). CriterionSM online essay evaluation: An application for automated evaluation of student essays. *Proceedings of the Fifteenth Conference on Innovative Applications of Artificial Intelligence* (pp. 3–10). American Association for Artificial Intelligence.
- Cargill, T. A. (1980). The design of a spelling checker's user interface. *ACM SIGOA Newsletter*, 1(3), 3–4. <https://doi.org/10.1145/1017923.1017924>
- Chollampatt, S., & Ng, H. T. (2018). A multilayer convolutional encoder-decoder neural network for grammatical error correction. *Proceedings of the AAAI conference on artificial intelligence* (vol. 32(1), pp. 5755–5762). Association for the Advancement of Artificial Intelligence. <http://dx.doi.org/10.1609/aaai.v32i1.12069>
- Chollampatt, S., Wang, W., & Ng, H. T. (2019). Cross-sentence grammatical error correction. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 435–445). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/P19-1042>

- Dahlmeier, D., & Ng, H. T. (2011). Grammatical error correction with alternating structure optimization. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 915–923). Association for Computational Linguistics.
- Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a large annotated corpus of learner English: The NUS corpus of learner English. *Proceedings of the eighth workshop on innovative use of NLP for building educational applications* (pp. 22–31). Association for Computational Linguistics.
- Dale, R., Anisimoff, I., & Narroway, G. (2012). HOO 2012: A report on the preposition and determiner error correction shared task. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 54–62). Association for Computational Linguistics.
- Du, Z., & Hashimoto, K. (2023). Sentence-level revision with neural reinforcement learning. *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)* (pp. 202–209). Association for Computational Linguistics.
- Grundkiewicz, R., Junczys-Dowmunt, M., & Gillian, E. (2015). Human evaluation of grammatical error correction systems. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 461–470). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/D15-1052>
- Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 489–500). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/D18-1045>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. <https://doi.org/10.1037/h0031619>
- Jensen, K., Heidorn, G., Miller, L., & Ravin, Y. (1993). Parse fitting and prose fixing. *Natural Language Processing: The PLNLP Approach* (pp. 53–64). Springer. https://doi.org/10.1007/978-1-4615-3170-8_5
- Katsumata, S., & Komachi, M. (2020). Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (pp. 827–832). Association for Computational Linguistics.
- Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. https://repository.upenn.edu/asc_papers/43
- Kwasny, S. C., & Sondheimer, N. K. (1981). Relaxation techniques for parsing grammatically ill-formed input in natural language understanding systems. *American Journal of Computational Linguistics*, 7(2), 99–108.
- Lee, J. S. (2004). Automatic article restoration. *Proceedings of the Student Research Workshop at HLT-NAACL 2004* (pp. 31–36). Association for Computational Linguistics.
- Nangia, N., Vania, C., Bhalariao, R., & Bowman, S. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1953–1967). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2020.emnlp-main.154>
- Omelianchuk, K., Atrasevych, V., Chernodub, A., & Skurzzhanskyi, O. (2020). GECToR-Grammatical Error Correction: Tag, not Rewrite. *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 163–170). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2020.bea-1.16>
- Leacock, C., Gamon, M., & Brockett, C. (2009). User input and interactions on Microsoft Research ESL assistant. *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 73–81). Association for Computational Linguistics.
- Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). *Automated grammatical error detection for language learners* (2nd ed.). Morgan & Claypool Publishers. <https://doi.org/10.1007/978-3-031-02153-4>
- Li, W., & Wang, H. (2024). Detection-correction structure via general language model for grammatical error correction. *arXiv preprint arXiv:2405.17804*. <http://dx.doi.org/10.48550/arXiv.2405.17804>
- Marzi, G., Balzano, M., & Marchiori, D. (2024). K-Alpha Calculator-Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. *Methods X*, 12, 102545. <https://doi.org/10.1016/j.mex.2023.102545>
- Napoles, C., Sakaguchi, K., Post, M., & Tetreault, J. (2015). Ground truth for grammatical error correction metrics. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Vol. 2: Short Papers, pp. 588–593). Association for Computational Linguistics. <http://dx.doi.org/10.3115/v1/P15-2097>
- Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the eighteenth conference on computational natural language learning: Shared task* (pp. 1–14). Association for Computational Linguistics. <http://dx.doi.org/10.3115/v1/W14-1701>

- Qorib, M. R., & Ng, H. T. (2022). Grammatical error correction: Are we there yet? In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 2794–2800). International Committee on Computational Linguistics.
- Randolph, J. J. (2005). Free-marginal multirater kappa (multirater K[free]): An alternative to fleiss' fixed-marginal multirater kappa. *Presented at the Joensuu Learning and Instruction Symposium 2005* (October 14–15, 2005). <http://files.eric.ed.gov/fulltext/ED490661.pdf>
- Rothe, S., Mallinson, J., Malmi, E., Krause, S., & Severyn, A. (2021). A simple recipe for multilingual grammatical error correction. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (vol. 2: Short Papers, pp. 702–707). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2021.acl-short.89>
- Rozovskaya, A., & Roth, D. (2010). Training paradigms for correcting errors in grammar and usage. *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 154–162). Association for Computational Linguistics.
- Rozovskaya, A. & Roth, D., (2021). How good (really) are grammatical error correction systems? *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2686–2698). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.eacl-main.231>
- Sakaguchi, K., Napoles, C., Post, M., & Tetreault, J. (2016). Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4, 169–182. <http://dx.doi.org/10.18653/v1/P18-1020>
- Starchenko, V. M., & Starchenko, A. M. (2023). Here we go again: modern GEC models need help with spelling. *Proceedings of ISP RAS*, 35(5), 215–228. [http://dx.doi.org/10.15514/ISPRAS-2022-35\(5\)-14](http://dx.doi.org/10.15514/ISPRAS-2022-35(5)-14)
- Starchenko, V. M. (2024). No need to get wasteful: The way to train a lightweight competitive spelling checker. *Computación y Sistemas*, 28(3), 1–12. <https://doi.org/10.13053/CyS-28-4-5068>
- Vinogradova, O., & Lyashevskaya, O. (2022). Review of practices of collecting and annotating texts in the learner corpus REALEC. *International Conference on Text, Speech, and Dialogue* (pp. 77–88). Springer International Publishing. http://dx.doi.org/10.1007/978-3-031-16270-1_7
- Volodina, E., Bryant, C., Caines, A., De Clercq, O., Frey, J., Ershova, E., Rosen, A., & Vinogradova, O. (2023). MultiGED-2023 shared task at NLP4CALL: Multilingual grammatical error detection. *Linköping Electronic Conference Proceedings* (pp. 1–16). LiU Electronic Press. <https://doi.org/10.3384/ecp197001>
- Wang, C., Li, R., & Lin, H. (2017). Deep context model for grammatical error correction. *SLaTE* (pp. 167–171). International Speech Communication Association. <http://dx.doi.org/10.21437/SLaTE.2017-29>
- Wang, Y., Xia, Y., He, T., Tian, F., Qin, T., Zhai, C., & Liu, T. Y. (2019). Multi-agent dual learning. *Proceedings of the International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations.
- Wang, Y., Wang, Y., Dang, K., Liu, J., & Liu, Z. (2021). A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology*, 12(5), 1–51. <http://dx.doi.org/10.1145/3474840>
- Warrens, M. J. (2010). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification*, 4(4), 271–286. <https://doi.org/10.1007/s11634-010-0073-4>
- Xie, Z., Avati, A., Arivazhagan, N., Jurafsky, D., & Ng, A. Y. (2016). Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727*. <http://dx.doi.org/10.48550/arXiv.1603.09727>
- Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 180–189). Association for Computational Linguistics.
- Yuan, Z., & Bryant, C. (2021). Document-level grammatical error correction. *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 75–84). Association for Computational Linguistics.
- Yuan, Z., & Felice, M. (2013). Constrained grammatical error correction using statistical machine translation. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task* (pp. 52–61). Association for Computational Linguistics.
- Yuan, Z., Briscoe, T., & Felice, M. (2016). Candidate re-ranking for SMT-based grammatical error correction. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 256–266). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/W16-0530>
- Zeng, M., Kuang, J., Qiu, M., Song, J. and Park, J. (2024). Evaluating prompting strategies for grammatical error correction based on language proficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 6426–6430). ELRA and ICCL. <https://doi.org/10.48550/arXiv.2402.15930>

- Zhang, Y., Zhang, B., Li, Z., Bao, Z., Li, C., & Zhang, M. (2022). SynGEC: Syntax-enhanced grammatical error correction with a tailored GEC-oriented parser. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2518–2531). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2022.emnlp-main.162>
- Zhao, J., Fang, M., Pan, S., Yin, W., & Pechenizkiy, M. (2023). GPTBIAS: A comprehensive framework for evaluating bias in large language models. *arXiv preprint arXiv:2312.06315*. <http://dx.doi.org/10.48550/arXiv.2312.06315>
- Zhou, H., Liu, Y., Li, Z., Zhang, M., Zhang, B., Li, C., Zhang, J., & Huang, F. (2023). Improving Seq2Seq grammatical error correction via decoding interventions. *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 7393–7405). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/2023.findings-emnlp.495>

APPENDIX A

GENERAL INFORMATION ABOUT THE SOURCE DATASETS USED FOR THE COLLECTION OF THE CONCENTRATED DATASET

Dataset	Size, tokens	# of annotations per document ^a	Error types	Language proficiency
FCE, evaluation part	41.9k	1	71	B1–B2
CoNLL-2014	30.1k	2–18	28	C1
BEA-2019, evaluation part	85.7k	5	55	A1–Native
REALEC	1550.6k	1	48	B1–B2

Note. ^a The number of annotation sets (by different annotators) provided for each document.

APPENDIX B

ERROR TAGS USED IN THE DATASET AND THE RATIO OF CONTEXT-SENSITIVE ERRORS

Original tag ^a	New tag ^b	Ratio of distant-con-text-sensitive errors ^c	Description	Example ^d
Linking_device	LINK	59,05%	The linking device is either wrong or erroneously absent	Secondly, the majority of the population will use other kinds of public transport, for example, trains, cars, or ships. So → However , we cannot say that these types of transport harm our environment less than planes do.
Ref_device	REF	50,83%	The wrong referential device is used	We should not create barriers for ambitious people and accept persons → those who don't have interest in education just because of sex equality.
VERB: TENSE	VERB: TENSE	45,35%	The wrong verb tense is chosen	When I was small, we lived in the country. I remembered → remember , we used to have oil lamps which used a cotton string dipping in the oil in the small bottle and made it burn the tip of the cotton string to give us light during the night.
PUNCT	PUNCT	37,61%	The wrong punctuation mark is used	In Sweden the level fell from 84% to 15%, a similar situation was in France. The → : the level changed from 90% to 50%.
PRON	PRON	36,72%	The personal pronoun is either wrong or erroneously absent	Also, he is very funny and I laugh a lot with him. Both → We both like to travel around the world and to do some sports, for example, tennis, running or trekking.
Inappropriate_register	REF, PRON4	15,50%	Errors related to style and appropriateness	When a child begins learning, for example, English in primary school, he → they get the necessary basis for further studying. (Tagged as PRON) Unfortunately, watching sports doesn't teach us → viewers anything and people don't get any information about the surrounding world from it. (Tagged as REF)
DET	DET	9,45%	The determiner is either wrong or erroneously absent	This situation creates a lot of pollution for ∅ → the environment, so we have to be more concerned about the planet's health.

Notes. ^a Original tag is the tag used in the original dataset. ^b New tag is the tag used in the concentrated dataset.

^c Ratio of distant-context-sensitive errors denotes the percentage of such errors among all annotated errors marked with the original tag.

^d For clarity purposes, all the other mistakes present in the example sentences were corrected in accordance with corrections suggested by the annotators of the source datasets.

⁴ During the annotation process, we concluded that other tags (such as PRON or REF) were suitable for the context-sensitive examples tagged as Inappropriate_register in REALEC.

APPENDIX C

OTHER TAGS USED IN THE CONCENTRATED DATASET

Tag	Description	Example
LEX	Lexical choice error	Also, it is a good way to get some positive emotions. All of this → Watching sports can even promote future productivity at work.
NOUN:NUM	The noun is used in the wrong number	By the way, there is an opposite tendency with young people, their num-ber → numbers are the largest at the science courses and the smallest in the sports and health courses. Additionally, students of the health and sports course → courses are mostly middle-aged.
SPELL	Spelling error	To sum up, both characteristics are important in our life. We need to know how to operate with once → ones we were born with and know how to develop knowledge gained from our experience to have a successful life and reach goals we set for ourselves.
SYN	Wrong choice or erroneous change of syntactic structure	Although the grandparents are in most cases ready to help, they can not transfer the values of the new world to the kids, and their → this results in the wrong choice of paths of life for the grown-up adults in future.
VERB:MODAL	The modal verb is erroneously absent, unnecessarily present, or used incorrectly	In addition, to decrease the risk of negative comments or posts, Facebook and Twitter would → should improve their futures by solving the personal privacy problem.
VERB:SVA	Errors related to subject-verb agreement	Today, public transport still play → plays an important role in the transport system and it will keep on doing so in the future.
WO	Errors in word order, e.g., the subject and verb are not inverted in the necessary contexts	But when I was a teenager, I began to experience situations that I did not like, for instance, girls said to me bad things → bad things to me or they talked unkindly about me.

<https://doi.org/10.17323/jle.2024.22224>

Facilitating Large Language Model Russian Adaptation with Learned Embedding Propagation

Mikhail Tikhomirov^{ORCID}, Daniil Chernyshev^{ORCID}

Lomonosov Moscow State University, Moscow, Russia

ABSTRACT

Background: Recent advancements in large language model (LLM) technologies have introduced powerful open-source instruction-tuned LLMs that match the text generation quality of leading models like GPT-4. Despite accelerating LLM adoption in sensitive-information environments, the lack of disclosed training data hinders replication and makes these achievements exclusive to specific models.

Purpose: Given the multilingual nature of the latest iteration of open-source LLMs, the benefits of training language-specific LLMs diminish, leaving computational efficiency as the sole guaranteed advantage of this computationally-expensive procedure. This work aims to address the language-adaptation limitations posed by restricted access to high-quality instruction-tuning data, offering a more cost-effective pipeline.

Method: To tackle language-adaptation challenges, we introduce Learned Embedding Propagation (LEP), a novel method with lower training data requirements and minimal disruption of existing LLM knowledge. LEP employs an innovative embedding propagation technique, bypassing the need for instruction-tuning and directly integrating new language knowledge into any instruct-tuned LLM variant. Additionally, we developed Darumeru, a new benchmark for evaluating text generation robustness during training, specifically tailored for Russian adaptation.

Results: We applied the LEP method to adapt LLaMa-3-8B and Mistral-7B for Russian, testing four different vocabulary adaptation scenarios. Evaluation demonstrates that LEP achieves competitive performance levels, comparable to OpenChat 3.5 and LLaMa-3-8B-Instruct. Further improvements were observed through self-calibration and additional instruction-tuning steps, enhancing task-solving capabilities beyond the original models.

Conclusion: LEP offers a viable and efficient alternative to traditional language-specific instruction-tuning, significantly reducing the costs associated with language adaptation while maintaining or surpassing the performance benchmarks set by contemporary LLMs.

KEYWORDS

large language model, llama, language adaptation, natural language generation

Citation: Tikhomirov, M., Chernyshev, D. (2024). Facilitating Large Language Model Russian adaptation with learned embedding propagation. *Journal of Language and Education*, 10(4), 130-145. <https://doi.org/10.17323/jle.2024.22224>

Correspondence:
Mikhail Tikhomirov,
tikhomirov.mm@gmail.com

Received: August 15, 2024

Accepted: December 16, 2024

Published: December 30, 2024

INTRODUCTION

Emergence of universal instruct-tuned large language models (LLM) such as ChatGPT (Ouyang, 2022) has substantially accelerated the development of natural language processing technologies. However, despite the remarkable achievements in zero-shot task solving, the close-source nature of such models prevented their adoption in the areas with sensitive or exclusive information where any risk of data-leak jeopardizes the integrity of the business process.

As a result the rising demand for open-source alternatives drove the researchers to derive methods for knowledge distillation of state-of-the-art LLMs. One of the first approaches was Alpaca (Taori, 2023) which used ChatGPT to synthesize the instruct-tuning data for open-source foundation LLM LLaMA (Touvron, 2023a). While Alpaca was far from state-of-the-art this inspired the creation of more advanced schemes like BactrianX (Li, 2023) that augmented the synthesis process with cross-lingual machine translation which in turn enabled training of open-



source multilingual chatbots. However, with release of GPT-4 (Achiam, 2023) which excelled in multilingual setting it became possible to integrate the explicit translation step into instruction synthesis pipeline thus increasing accessibility of knowledge distillation. This has led to creation of series language-specialized instruction-tunes of open-source LLMs such as Saiga (Gusev, 2023), PolyLM (Wei, 2023), Vikhr (Nikolich, 2024), LLAMMAS (Kuulmets, 2024).

With increasing instruction synthesis quality the open-source language-specific LLMs were closing the gap with the state-of-the-art closed-source solutions eventually hitting the performance ceiling of conventional instruction-tuning (Cui, 2023) due to low utilization of inherent English contextual knowledge which is dominant in state-of-the-art pre-trained open-source LLMs (Touvron, 2023b; Jiang, 2023; Dubey, 2024). As a possible solution researchers (Zhu, 2023; Li, 2024; Chai, 2024) proposed enriching the instruction-tuning datasets with translation tasks which are designed to align new language knowledge with the existing English semantic representations. However, it was shown by Ranaldi (2023) and Husain (2024) that the cause of alignment issue is likely to lie with the inefficiency of tokenization algorithm which can be addressed either by building a new language-specific token vocabulary or by recycling the English tokens for Romanized language representation.

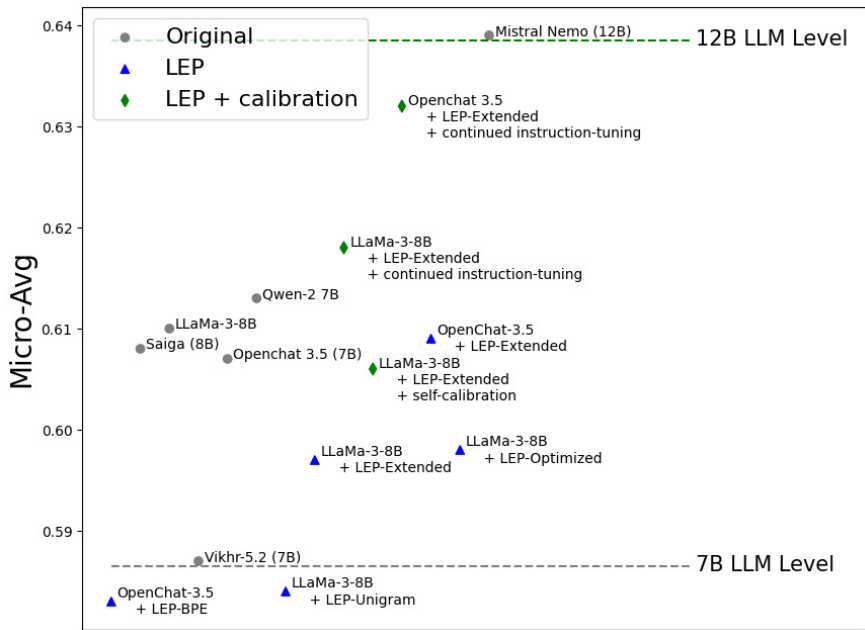
Inspired by works of Lakew (2018), Kuratov (2019), Rust (2021) & Yang (2022) on vocabulary adaptation for encoder models Cui et al. (2023) proposed language-specific continued pre-training pipeline for full LLM language adaptation which paired with instruct-tuning on synthesized examples allowed to create Chinese LLaMa, the first open-source model to reach the performance level of ChatGPT

with substantially improved computation efficiency thanks to Chinese-adapted tokenization vocabulary. This approach was studied in detail by Tikhomirov (2023) for LLaMa-2 (Touvron, 2023b) adaptation to Russian language and it was shown that semantic alignment efficiency can be further improved with morphologically accurate tokenization algorithm. Moreover, the full LLM language adaptation pipeline was shown by Nguyen (2023) to outperform state-of-the-art closed-source counterparts on low-resource languages due to their bias towards popular languages.

While the current iteration of language adaptation algorithm is relatively cost-efficient, the benefit of developing language adapted LLMs is falling amid the rapid development of LLM technology and multilingual specialization of open-source options. At the same time it becomes common to release instruction-tuned models (Jiang, 2023; Dubey, 2024) that perform on par with closed-source state-of-the-art counterparts without disclosing the instruction-tuning data the quality of which is the major factor of resulting LLM task-solving capabilities (Zhou 2024). Collecting data of such quality requires a considerable investment in human annotation to an extent that only large organizations can afford creation of such datasets (Dubey 2024). If a language specific counterpart of a high quality instruction dataset is unavailable the result of full language adaptation will only have the benefit of higher computational performance as an inferior instruction-tuning data will lead to inferior task-solving performance.

To cut the language adaptation costs and enable direct language adaptation of instruction-tuned LLM we propose an updated pipeline for language adaptation, Learned Embedding Propagation. Unlike the original full LLM language ad-

Figure 1
Performance Comparison of Proposed Adaptation Method on Darumeru Benchmark



adaptation pipeline (Cui, 2023), our method requires less data and computational resources due to limited pre-training impact on model parameters which is compensated by novel ad-hoc embedding propagation procedure that allows to skip the instruction-tuning step and instead implant the new language knowledge directly into any existing instruct-tuned variant. To further facilitate the Russian adaptation we developed a new lightweight benchmark for train-time evaluation of LLM text generation robustness, Darumeru. We test Learned Embedding Propagation pipeline on Mistral-7B and LLaMa-3-8B LLMs for 4 Russian tokenization variants. The evaluation results (Figure 1) demonstrate that despite lower parametrization our language-adaptation method manages not only to regain the original quality of the instruction tune but in some cases even outperform it by a significant margin. Additional case-study experiments on improving the best language-adapted models with continued instruct-tuning and self-calibration also confirm the superiority of our language-adapted models, pushing their performance beyond existing counterparts.

METHOD

Model Language Adaptation

Following the previous work on LLM lingual adaptation (Cui, 2023; Tikhomirov, 2023) we first optimize model vocabulary for better alignment with Russian language morphology and then continue the pre-training process on a large corpora of Russian texts of various genres and topics.

Formally the model adaptation consists of 3 steps:

1. Tokenization training;
2. Model embedding initialization;
3. Continued pre-training of new embeddings (both input and output).

Tokenization training

Since there are no best practices for vocabulary optimization we consider 4 options for tokenization training:

BPE - fully substituting the tokenization vocabulary by rebuilding the BPE tokenization algorithm (Vries, 2021), which is used in the majority of state-of-the-art LLMs.

Unigram - fully substituting the tokenization vocabulary with morphologically accurate tokenization obtained with Unigram algorithm (Tikhomirov, 2023).

Extension - extending the original BPE vocabulary by first building a new BPE vocabulary for Russian corpora and then merging it with the original (Cui, 2023).

Optimization - refactoring the existing BPE vocabulary by reducing it to the most common 50% tokens of Russian corpora and then subsequent Extension to the original size. (considered only for LLMs with extensive English vocabulary).

Embedding Initialization

Previous work on LLM language adaptation (Cui, 2023; Tikhomirov, 2023; Nguyen, 2023) found simple averaging of embeddings of overlapping subtokens to be a sufficient solution for embedding initialization. Formally, given embedding vectors of old and new tokenization vocabularies the new embeddings are initialized as the following:

$$v_{new}(t_i^n) = \frac{1}{K} \sum_{j=1}^K v_{old}(t_j^o); \quad (1)$$

$$tokenize_{old}(t_i^n) = [t_1^o, \dots, t_K^o]. \quad (2)$$

where t_i^n is the original tokenization function, t_i^n is token in new vocabulary, t_j^o is a token in original vocabulary.

While there are more advanced initialization techniques, recent studies on design choices for LLM language adaptation (Tejaswi, 2024) concluded that embedding averaging has the best expected adaptation quality and the performance gap with task-tailored methods is within standard deviation of task evaluation protocol. Therefore for all experiments we use the described subtoken averaging embedding initialization strategy.

Continued Pre-Training

The main issue with embedding initialization is that despite introduction of new tokens the LLM retains the habit to use the tokens that were present in the original tokenization. As a result the model computational performance of text generation remains the same as the model tends to use more tokens per word than it is expected while also misinterpreting the new tokens due to homonymy of token context.

To alleviate the issue the common tactic is to train the newly initialized embeddings on adaptation language corpora using the same pre-training task as LLM, which is causal language modeling. In this task the input text is broken into sequences of tokens of increasing size all of which start from the beginning and the model is asked to predict for each sequence the next possible token. The model optimization is done using simple cross-entropy loss thus any text corpora can be used for the pre-training task.

Continued pre-training of embeddings only allows the model to tailor those embeddings for inner semantics thus redistributing the existing language knowledge among the newly introduced tokens. However, some researchers (Cui, 2023; Tikhomirov 2024) argued that pre-training embeddings only may be insufficient for proper model-vocabulary alignment

and intermediate model layers must be also trained. On the other hand, increasing the number of trained model parameters reduces the training process stability which in turn substantially raises the data size requirements and computational costs of training procedure. As the middle ground we complement embedding pre-training with a post-training layer alignment procedure that recycles existing fine-tunes of the adapted model.

Learned Embedding Propagation

The issue of cost-efficient knowledge transfer for language adapted models has been studied before in the context of encoder models. To solve the absence of task-tuning dataset in the target language Artetxe et al. (2019) proposed a simple algorithm for transferring task-solving knowledge to BERT models:

1. Pre-train the full language model from scratch on available large monolingual text corpora (e.g English) using language modeling training objective (for BERT it is masked language modeling);
2. Create a copy of the pre-trained model and replace the embeddings of the original with new embeddings for the target language;
3. Continue the pre-training of the modified original on target language monolingual corpora for model embeddings while freezing (not updating) all other layers using the same training objective;
4. Fine-tune the copy on the downstream task dataset while keeping the embeddings frozen;
5. Swap the embeddings of the fine-tuned copy with embeddings of the original model obtained after continued pre-training on the target language corpora.

The major advantage of the described algorithm is that the continued pre-training step requires much less data than initial pre-training from scratch as it requires training only a fraction of model parameters which reduces model optimization task complexity and thus has faster convergence (Kaplan, 2020). The main hypothesis is that task-solving knowledge is language agnostic and it was confirmed in the original experiments (Artetxe, 2019) for natural language understanding and document classification tasks. However, the authors noted that fine-tuning on downstream tasks with frozen embeddings is not enough for proper embedding swap alignment and additional embedding transformations or special embedding utilization penalties are required to maximize the efficiency of target language vocabulary processing. As a possible solution to the embedding alignment problem Chen et al. (2023) proposed using a special pre-training regime with active embedding forgetting to force the language model to accumulate the knowledge in intermediate layers. The downside of such an approach is that we must have full control on the initial pre-training which is not possible for state-of-the-art LLMs obtained by

training on high quality proprietary datasets with immense computational budget.

We argue that embedding swap alignment can be achieved without special training procedures by leveraging the fine-tuning parameter update trajectory. Ilharco et al. (2023) showed that the fine-tuning trajectory may be approximated with linear transformations of base model parameters which can be derived from parameter decomposition of fine-tuned variants. Therefore, by finding appropriate linear transformations for embedding parameters we can approximate the results of a full language adaptation pipeline without involving the instruction-tuning dataset.

Formally, let I , O be the input and output embeddings of LLM and W a pseudo-linear approximation of composition of intermediate LLM layers:

$$LLM_{base} = I_{base} W_{base} O_{base} \quad (3)$$

Denote D , U as linear embedding transformations that align original embeddings with the fine-tuned layers:

$$\begin{aligned} LLM_{base \rightarrow inst} &= I_{inst} W_{inst} O_{inst} = \\ &= I_{base} D_{inst} W_{inst} U_{inst} O_{base} \end{aligned} \quad (4)$$

Since our target language embedding initialization strategy averages the embeddings of overlapping tokens in I_{base} and O_{base} we can formalize the initialization process with vocabulary transformation operation :

$$LLM_{base \rightarrow ru} = T_{ru} I_{base} W_{base} O_{base} T_{ru}^T = I_{ru} W_{base} O_{ru} \quad (5)$$

Following the logic described above the fine-tune of language adapted base model $LLM_{ru \rightarrow inst}$ can be represented as the following:

$$LLM_{ru \rightarrow inst} = T_{ru} I_{base} D_{inst}^{ru} W_{ru \rightarrow inst} U_{inst}^{ru} O_{base} T_{ru}^T \quad (6)$$

Now by assuming that the optimal $W_{ru \rightarrow inst} \approx W_{inst}$ we arrive at the final equation for propagation of continued pre-trained embeddings $I_{ru/cpt}$, $O_{ru/cpt}$:

$$LLM_{ru/cpt \rightarrow inst} = I_{ru/cpt} D_{inst}^{ru} W_{inst} U_{inst}^{ru} O_{ru/cpt} \quad (7)$$

The remaining variables D_{inst}^{ru} , U_{inst}^{ru} are determined by chosen assumptions about embedding alignment properties. In our experiments we consider 3 options:

1. Direct embedding swap
2. Overlapping token correction
3. Vocabulary conversion projection

Direct embedding swap

Considering that most state-of-the-art LLMs are trained on multilingual datasets, it can be expected that their inner representations are tailored for language-agnostic text processing. Similarly to the original works on embedding-based knowledge transfer for encoder models we assume that the embedding layer carry only conceptual information i.e. we suppose $D_{inst}^{ru} = U_{inst}^{ru} = E$ where E is an identity matrix.

Overlapping Token Correction

Since the considered LLMs are initially designed for multilingual text generation they have a basic set of the most common tokens for popular languages such as russian. The idea is to find the union $C = \text{tokens}^{old} \cap \text{tokens}^{new}$ of the original tokens^{old} and language-adapted tokens^{new} vocabularies and use this subset to reduce I_X, O_X to the common components of embedding initialization $I_{X/com}, O_{X/com}$ where $X \in \{base, inst\}$. This allows to approximate the embedding projections as $D_{inst}^{ru} \approx D_{inst}$ and $U_{inst}^{ru} \approx U_{inst}$:

$$D_{inst}^{ru} = I_{base/com}^{-1} I_{inst/com}, \quad (9)$$

$$I_{X/com} = [I_X^{idx(t)}]_{t \in C}, \quad (10)$$

$$O_{X/com} = [O_X^{idx(t)}]_{t \in C}. \quad (11)$$

where $idx(t)$ is a function that maps token t to its respective position in the embedding matrix. It must be noted that $I_{X/com}, O_{X/com}$ matrices are likely to be not invertible and thus their inversion must be approximated with least squares problem solvers.

Vocabulary Conversion Projection

Since embedding initialization transformation T_{ru} is universal for both base and fine-tuned models we can derive an alternative equation for obtaining language-adapted instruction-tuned LLM:

$$LLM_{inst \rightarrow ru} = T_{ru} I_{inst} W_{inst} O_{inst} T_{ru}^T \quad (12)$$

By assuming that both variants of instruction-tune adaptation are equivalent $LLM_{ru \rightarrow inst} = LLM_{inst \rightarrow ru}$ we obtain the following formulae for embedding alignment:

$$D_{inst}^{ru} = (T_{ru} I_{base})^{-1} T_{ru} I_{inst} \quad (13)$$

$$U_{inst}^{ru} = O_{inst} T_{ru}^T (O_{base} T_{ru}^T)^{-1} \quad (14)$$

Similarly to the previous alignment method the calculation of transformation matrices involves least square problem solvers for finding the pseudo-inversion of non-invertible

matrices. This is the main reason why vocabulary transformation T_{ru} should not be isolated. The pilot experiments showed that such simplification increases the error margin of alignment transformations which lowers the quality of embedding propagation procedure.

Darumeru Benchmark

Existing LLM benchmarks for Russian language (Fenogenova, 2024) do not expose the testing data labels for local evaluation. On one hand such an initiative is reasonable amid the rising trend of training on test data which renders the LLM ranking results meaningless. On the other hand hidden test labels means that the evaluation requires having an online connection to the benchmark system which prevents evaluation in offline computational environments thus postponing the evaluation until the end of training session. Moreover lack of access to test labels makes it impossible to classify the type of prediction errors thus limiting the post-training quality analysis.

To address the issue we developed a new benchmark framework that focuses on quick and informative LLM text generation quality evaluation. This benchmark consists of combinations of open splits of datasets from MERA (Fenogenova, 2024), mmlu_ru / mmlu_en, RuCoLA (Mikhailov, 2022), as well as new datasets for text generation assessment - 17 datasets total. A more detailed description of each dataset is given in the following sections.

Framework

The evaluation framework utilizes message format to ensure compatibility with both pre-trained and instruction-tuned LLMs. This means that all task data for the models is converted into a sequence of "user role"- "message content" pairs, from which the final prompt is constructed. The framework supports tasks that require estimating the probability of the next token, generation, or logsoftmax for the entire generated sequence. The evaluation can be carried out directly in a conventional Transformers model training environment or via VLLM specialized model inference servers.

DaruMERA and DaruMMLU

We composed **DaruMERA** from the following MERA datasets: MultiQ, PARus, RCB, RWSD, USE, , ruOpenBookQA, ruWorldTree. For better language understanding evaluation we also added validation split of RuCoLA dataset.

For **DaruMMLU** part we separated ruMMLU (MERA) and complemented it with MMLU datasets from the NLP-Core-Team repository¹.

¹ https://github.com/NLP-Core-Team/mmlu_ru

There are several changes to the original datasets:

1. MultiQ version was augmented with additional gold answers. The existing labels do not correspond in form to the questions, as they were extracted from the text without proper preprocessing. The augmentation process consisted of passing the question and reference answer pairs to LLaMa-3-70B-Instruct model to rephrase the answer in accordance with the question.
2. The ruMMLU version differs from the similar one in NLP-Core-Team repository in that it has few-shot examples common to all queries, regardless of the domain, and also uses not one fixed template, but several options as instructions.
3. When calculating PARus, for each example the same example was generated, but with a different order of options, and only the case when the model predicts the correct option for both the direct and reverse order was considered a success.

To measure the performance on PARus, RWSD, MMLU datasets we used accuracy metric. For RCB, ruOpenBookQA and ruWorldTree we averaged accuracy and F1-macro. For RuCoLa we used average of accuracy and Matthews Correlation Coefficient (MCC). For MultiQ we used the average of F1 and exact match metrics. For USE the normalized total grade was used.

DaruSum

Most of the evaluation tasks aim to measure the model's text comprehension capabilities and global contextual knowledge which is required for proper prompt processing. However for text generation the model must be also capable of filtering the input text for the query relevant content to ensure that the user would receive the desired answer regardless of input format or size. Text summarization is the perfect evaluation task for such a case as it requires both filtering the input content and composing the answer from the salient fragments.

There are two summarization settings: extractive and abstractive. Extractive summarization is a task of sentence saliency ranking where the summary is obtained by taking top-k ranked sentences. Abstractive summarization on the other hand is a text generation task where saliency ranking is integrated in the token sampling process as the model guides itself toward the most concise summary. While the abstractive setting has the higher preference it is hard to distinguish automatically the suboptimal content filtering from the text generation errors. At the same time constraining the text generation process to input fragments such as sentences basically reduces the task to extractive summarization. Thus to evaluate content filtering accuracy and text

generation quality it is sufficient to evaluate the abstractive summarization in free and constrained generation settings.

For the summarization dataset we chose Gazeta (Gusev, 2020) which has established itself as the standard for Russian automatic summarization evaluation. To improve the accuracy of evaluation procedure we derived an example filtering protocol that all reference summary content can be inferred from the input document. Since LLaMa-3-70B showed high human agreement in LLM evaluation² we employed it as the example correctness evaluator and tasked it to find all citations that support the summary sentence. We filtered out all examples that had more than 20% of unsupported summary sentences and mapped found citations to document sentences, thus producing accurate extractive labels. To adapt the task for a few-shot setting which is limited by context window limitations we compressed the documents by dropping the paragraphs that had no extractive summary labels. To account for LLM text generation length variance (Dubois, 2024) as the metric for abstractive and extractive settings we chose average of ROUGE-1 and ROUGE-2 recall and R-precision respectively.

DaruCopy

When replacing the LLM vocabulary it is important that it learns to fully utilize new tokens. The input token embeddings are responsible for conveying the text meaning which can be evaluated by natural language understanding tasks such as MMLU. In contrast, the output token embeddings are used to find the closest semantic meaning to the current neural network state which depends on contextual history. As a consequence, in creative tasks this state is unstable and LLM tends to generate rarer tokens. At the same time, in the tasks where the LLM is required to reuse the input context the network state is expected to fall into semantic clusters of tokens that are present in the input sequence. Following that logic by prompting the LLM to produce a copy of the input text we can evaluate its token generation efficiency.

We used Wikipedia articles of different genres to collect copy task datasets for English and Russian languages involving 2 copy settings: sentence-wise and paragraph-wise. The former setting assesses the LLM alignment with tokenization algorithm which is calculated as the ratio of the length of the original text to the generated text in tokens. In paragraph setting we evaluate the overall text generation stability by measuring the percentage of generations in which the ratio of longest common subsequence (lcs) tokens to all paragraph tokens is greater than 99% (1% is left for spacing errors). Deviation from 99% amid the high sentence copy scores indicates that the model tends to confuse tokens and thus can hallucinate context in creative tasks which is the major reliability concern for practical applications.

² https://github.com/tatsu-lab/alpaca_eval

Benchmark Parameters

When calculating the benchmark metrics, the following parameters were set: batch size 8, sequence length 4096, 5-shot for foundation models and zero-shot for instruct models.

Experiment Setting

We conducted adaptation experiments with two models: Mistral-7B-v0.1 (Jiang, 2023) and LLaMa-3-8B (Dubey, 2024).

Continued Pre-Training

Training dataset for tokenization and continued pre-training consists of documents from the following domains: Russian Wikipedia, English Wikipedia, Habrahabr, Pikabu, Fiction, News, Educational literature.

The documents were deduplicated using Locality Sensitive Hashing Minhash algorithm. We removed metadata, links, comment sections and badly formatted documents to improve vocabulary distribution and reduce the number of grammatically incorrect examples. To reduce the semantic noise we restricted the vocabulary to Cyrillic and Latin languages and stripped non-standard symbols like emoji or logograms (e.g. Chinese characters) using UTF-8 normalization.

For training, texts were sampled with increased weights for Wikipedia, educational and scientific literature. Additionally, to feed texts into the language model, we ensured that each sample began either with a new document or with a new paragraph.

Tokenization parameters. We trained BPE and Unigram tokenizers with 32000 and 128000 tokens for Mistral-7B and LLaMa-3-8B respectively. For Extended tokenizer, we extended the original tokenizers to 55328 and 174816 tokens using new Russian-adapted BPE vocabularies for corresponding models. Since LLaMa-3-8B tokenization vocabulary is likely to be extensive we created an Optimized version, where we shrunk the original BPE vocabulary to 64000 tokens and then merged with top 64000 most common tokens from new BPE vocabulary, resulting in 114504 tokens.

Hyperparameters. During continued pre-training we used the following hyperparameters: Total Batch Size: 256; Block Size: 1024; Weight Decay: 0.1; Scheduler: Cosine; Warmup Steps: 100; Epochs: 1.

We tested 4 different learning rates: 2e-5, 5e-5, 1e-4, 2e-4 for each model and tokenization on 20% of all continued pre-training dataset. Based on benchmark results, we chose a learning rate equal to 1e-4 for all Mistral-7B models, and

learning rate equal to 2e-4 for LLaMa-3-8B models. It is important to note that the efficiency of model adaptation showed a significant dependence on the learning rate, especially for LLaMa-3-8B based models.

Case Study: Self-Calibration

For the cases of full vocabulary substitution where the model learns to rewire all new embeddings virtually from scratch the propagation process may have lower efficiency as the difference between instruct-tuned and language-adapted embeddings may be dramatic. The logical solution is to synthesize self-instruct data using the original instruct-tuned LLM and then use it to calibrate the language-adapted version. To generate the examples, we used prompts from Saiga instruction dataset and used greedy decoding to get the most likely answer from instruct-tuned LLM viewpoint. Then we asked LLaMa-3-70B to evaluate the quality of synthesized pairs in terms of grammar and relevance on a 5-point grading scale. All examples that received a score less than 4 were discarded which left us 13531 calibration examples.

Since calibration examples are native for LLM inner semantic representations there is a risk that instead of alignment the model may revert back to the original tokenization behavior which prioritizes smaller but more familiar tokenization chunks. To avert such a scenario we leverage the fact that all modern LLMs are pre-trained on Wikipedia articles in such a manner that their embedding representations are aligned with Wikipedia concepts. By asking the fine-tuned model to repeat a Wikipedia article token by token we force the model to recall its pre-training memory and thus to propagate the activation signals respective to the concepts in the article to embeddings of optimal tokens of new tokenization. Following that logic we supplemented the self-instruct dataset with 10000 article-copy task examples, obtained from the part of Wikipedia that has no overlap with our pre-training or benchmark datasets.

We found the following LoRA-tuning settings to be optimal for calibration procedure: Rank: 8; Alpha: 1; Learning Rate: 2.5e-5; Weight Decay: 0.1; LoRa target modules: first and last transformer layers; LoRa modules to save: lm_head, embed_tokens; Max Sequence Length: 8096 (i.e. max context length); Total Batch Size: 64; Epochs: 1.

Case Study: Continued Instruction-Tuning Calibration

In addition to the self-calibration experiments, we decided to test how continued instruction-tuning on the high-quality Russian instruction dataset would affect the final performance. For this experiment we choose Saiga³ dataset which is considered to be the best open-source option for Russian language. We also investigated the impact of adding a small

³ https://huggingface.co/datasets/IlyaGusev/saiga_scored

number (2000) of special instructions to the dataset, the purpose of which is to copy a large text from Wikipedia

To fine-tune the models we used LoRA adapters with Saiga-recommended hyperparameter settings which is the following: Rank: 32; Alpha: 16; Learning Rate: 5e-5; Weight Decay: 0.05; LoRa target modules: attention, mlp; LoRa modules to save: lm_head; Max Sequence Length: 4096; Total Batch Size: 128; Epochs: 1.

RESULTS

Open-source LLM Benchmark

To establish a baseline we benchmarked popular instruct-tuned LLMs (see Table 1): Openchat 3.5, LLaMa-3 (instruct) (Dubey, 2024), Saiga (Gusev, 2023), Vikhr (Nikolich, 2024), Qwen-2, Mistral Nemo (Jiang, 2023). As expected the largest model, Mistral Nemo, has the highest zero-shot performance. Smaller counterparts have the same score margin. However, Qwen-2 7B manages to outperform Mistral Nemo in MMLU tasks while falling behind on text generation robustness tests of DaruSum and DaruCopy. Vikhr-5.2 similarly has the same score on DaruMERA as Mistral Nemo. Considering the LLM scaling laws (Kaplan, 2020) and the performance gap with state-of-the-art sub-10B parameter LLM, LLaMa-3, this observations suggest that some parts of MMLU and MERA datasets were leaked to training data of Vikhr-5.2 and Qwen-2 7B.

Table 1
Darumeru Zero-Shot Evaluation Results for Popular Open-Source Instruct-Tuned Models

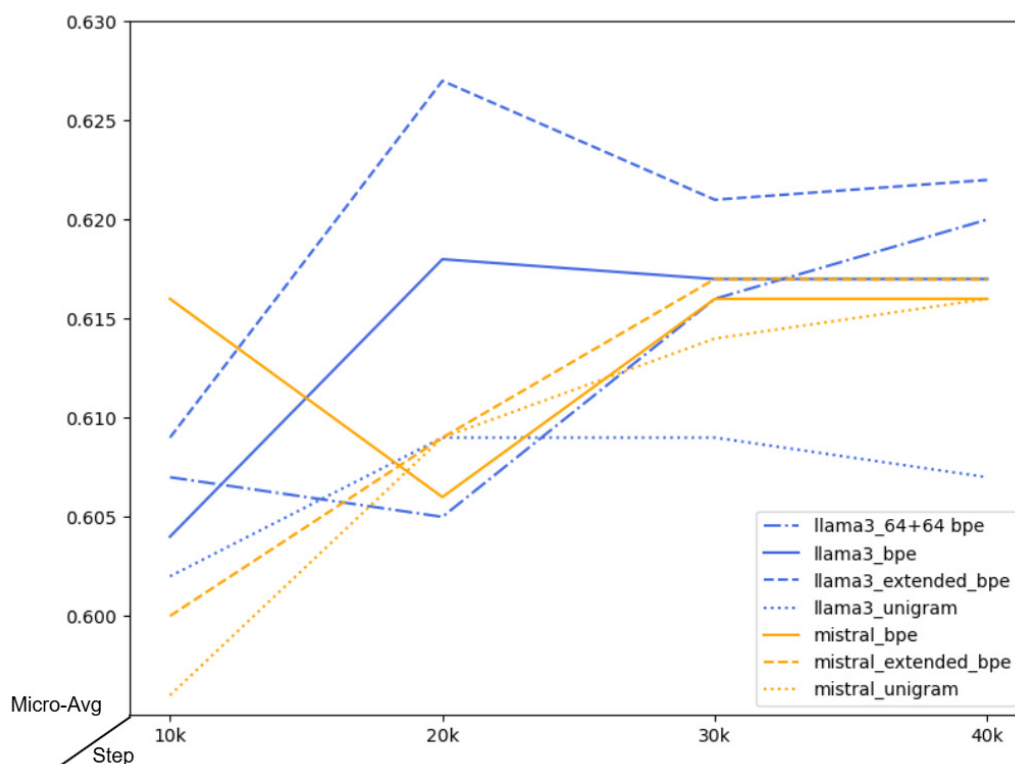
Model	Micro-Avg	DaruMMLU	DaruMERA	DaruSum	DaruCopy (EN)	DaruCopy (RU)
Openchat 3.5 (Mistral-7B)	0,607	0,543	0,526	0,322	0,999	0,917
LLaMa-3-8B (Instruct)	0,610	0,571	0,510	0,322	1,000	0,972
Saiga (LLaMa-3-8B)	0,608	0,574	0,514	0,320	0,995	0,939
Vikhr-5.2 (Mistral-7B)	0,587	0,494	0,573	0,308	0,959	<u>0,693</u>
Qwen-2 7B	0,613	0,624	0,548	0,300	0,938	<u>0,842</u>
Mistral Nemo (12B)	0,639	0,592	0,576	0,320	0,998	0,924
Ours						
Openchat 3.5 + LEP-Extended + calibration (best)	0,632	0,541	0,563	0,321	1,000	0,989
LLaMa-3-8B (Instruct) + LEP-Extended + calibration (best)	0,618	0,565	0,521	0,339	1,000	0,984

Vocabulary Adaptation and Continued Pre-Training

Following our initial benchmark results we focused on Russian adaptation of the foundation models of the most performant instruct-tunes: Mistral-7B and LLaMa-3-8B. To evaluate the language-adaption results we used few-shot in-context-learning as the models are not used to interpreting the instructions directly.

Figure 2 shows the Darumeru score dynamic throughout the continued pre-training process. In case of Mistral-7B the vocabulary substitution methods such as BPE and Unigram almost exhaust the training examples converging to the optimum at the final 10k training steps. In contrast LLaMa-3-8B is more robust to vocabulary adaptation methods as they all tend to converge in the middle of a training session at 20-30k steps. Since the full dataset size is 96 GB we can conclude that 40 GB of texts is the minimum required for the good performance of Russian adapted embeddings.

In Table 2 we report the detailed results of the best performing checkpoints. As expected, vocabulary extension methods such as Extended and Optimized have the lowest optimization difficulty as they show the highest language-adaptation scores. For Mistral-7B all language adaptations significantly outperform the original foundation, however the difference between their tokenization efficiency (symbols per token) and average task-performance may be considered marginal. For LLaMa-3-8B only Extended variants managed to reach

Figure 2*Micro Average Benchmark Score Dynamic throughout Training*

the original LLM benchmark scores mainly falling behind on DaruMMLU tasks. Most tokenization-efficient variants, BPE and Unigram, considerably lag behind, losing in DaruMERA and DaruSum. We assume that vocabulary substitution in case of BPE and Unigram has a major impact on language understanding and that in their case continued pre-training of embeddings-only is not sufficient for proper semantic alignment and additional tuning procedures are required.

Learned Embedding Propagation

The results of complete Learned Embedding Propagation (LEP) are reported in Table 3. For each adapted vocabulary construction option (BPE, Unigram, Extended and Optimized) we test 3 methods: Direct Embedding Swap (**Swap**), Overlapping Token Correction (**Overlap**) and Vocabulary Conversion (**Conversion**). For embedding donor model we used best continued pre-training checkpoints (see Table 2).

For Mistral-7B and OpenChat 3.5 the embedding propagation results have large variance depending on the chosen tokenization algorithm for Russian vocabulary. In case of BPE, which is the same algorithm used for the original, the trained embedding for new vocabulary has the highest alignment with instruct-tuned counterpart in case of direct embedding swap. In case of more morphologically correct Russian tokenization, Unigram, overlap projection has the highest average task performance. However, if we look at

group-wise scores it becomes evident that conversion is a better option as it leads in every task but DaruCopy (Ru) where all unigram conversion variants are experiencing issues. The conventional vocabulary extension also leans towards conversion projection and has the best overall task performance among all vocabularies even outperforming the original OpenChat 3.5.

For LLaMa-3-8B embedding conversion is more straightforward. For all tokenization variants the conversion projection yields the best results, however, unlike the Mistral-7B LEP none of embedding propagations manage to reach the original LLaMa-3-8B (instruct) quality. The significant performance degradation is observed among all task groups with DaruCopy taking the biggest hit. Moreover, despite being the original tokenization algorithm, BPE-build Russian vocabulary has the lowest embedding compatibility with instruction-tune having the largest score gap. While the vocabulary Optimized variant has lower vocabulary size limit it maintains the same quality level as Extended and comparing the conversion projections the former has better DaruSum and DaruCopy solving capabilities.

There are several implications of the observations. First is that the Vocabulary Conversion LEP algorithm is likely to be the most efficient solution for the majority of embedding projection scenarios in some cases even being sufficient for recovering the original instruction-tuned performance. Sec-

Table 2
Darumeru Few-Shot Evaluation Results for Best Language-Adaptation Checkpoints

Model	Vocab	Symbols per token	Micro-Avg	DaruMMLU	DaruMERA	DaruSum	DaruCopy (EN)	DaruCopy (RU)
Mistral-7B	original	2,44	0,604	0,545	0,504	0,307	1,000	1,000
	BPE	3,76	0,616	0,528	0,537	0,316	0,995	0,984
	Unigram	3,78	0,614	0,516	0,544	0,311	0,995	0,960
	Extended	3,77	0,617	0,538	0,532	0,314	1,000	0,995
LLaMa-3-8B	original	2,89	0,629	0,582	0,547	0,326	0,980	0,982
	BPE	4,40	0,618	0,561	0,532	0,321	1,000	0,963
	Unigram	4,35	0,609	0,560	0,517	0,316	1,000	0,951
	Extended	3,78	0,627	0,560	0,550	0,325	0,980	0,983
	Optimized	3,40	0,620	0,552	0,536	0,323	0,981	0,989

Table 3
Darumeru Zero-Shot evaluation Results for Learned Embedding Propagation Methods

Vocab	LEP method	Micro-Avg	DaruMMLU	DaruMERA	DaruSum	DaruCopy (En)	DaruCopy (Ru)
OpenChat-3.5							
BPE	Swap	0,587	0,528	0,526	0,277	0,988	0,829
	Overlap	0,584	0,525	0,523	0,281	0,986	0,818
	Conversion	0,583	0,526	0,524	0,284	0,993	0,791
Unigram	Swap	0,556	0,517	0,517	0,282	0,985	0,614
	Overlap	0,572	0,514	0,534	0,297	0,981	0,68
	Conversion	0,565	0,515	0,519	0,301	0,999	0,651
Extended	Swap	0,608	0,535	0,540	0,298	0,999	0,907
	Overlap	0,607	0,535	0,539	0,307	0,999	0,898
	Conversion	0,609	0,535	0,541	0,306	0,999	0,909
LLaMa-3-8B (instruct)							
BPE	Swap	0,565	0,544	0,486	0,317	0,999	0,729
	Overlap	0,569	0,546	0,489	0,314	0,999	0,753
	Conversion	0,570	0,546	0,490	0,318	0,999	0,754
Unigram	Swap	0,582	0,545	0,488	0,313	0,999	0,865
	Overlap	0,580	0,545	0,482	0,314	0,999	0,876
	Conversion	0,584	0,545	0,488	0,315	0,994	0,889
Extended	Swap	0,592	0,557	0,498	0,319	0,969	0,921
	Overlap	0,597	0,556	0,504	0,321	0,964	0,936
	Conversion	0,597	0,556	0,501	0,318	0,994	0,921
Optimized	Swap	0,594	0,554	0,499	0,327	0,970	0,928
	Overlap	0,586	0,553	0,495	0,323	0,925	0,925
	Conversion	0,598	0,555	0,500	0,324	0,995	0,928

only, while Unigram tokenization vocabulary may be considered morphologically correct for Russian language it is inferior to Extended and Optimization options as it requires full vocabulary substitution, which, considering unstable BPE performance, creates the largest disparity between embedding and inner-layer semantic representation. The tokens removed in the Optimized variant seem to be unimportant for Russian task-solving capabilities as it manages to outperform Extended tokenization which completely retains the original vocabulary. The performance gap in LEP LLaMa-3-8B (instruct) is likely to be the consequence of proprietary instruction-tuning dataset which was large enough to align the embedding semantics with instruction-following tasks (Dubey, 2024). Another hypothesis is that the original LLM underwent human preference alignment procedure which aims to block text generation of harmful answers at the cost of necessary reasoning limitations and as a consequence has a habit of blocking potential malicious semantics originating from input embeddings which in turn inhibits text comprehension capabilities.

Case Study: Self-Calibration

In self-calibration experiments we focused on closing the gap of best LEP LLaMa-3-8B instruct models (Table 4, self-calibration). As expected the performance of DaruCopy tasks improved substantially, practically reaching the perfect reliability levels. DaruSum also saw the improvements as the improved citation capabilities are beneficial for composing concise summaries. Other tasks however did not improve much and in the case of the weakest vocabulary adaptation, Unigram, saw a significant decline in benchmark scores.

We suspect that the self-calibration data promotes closed-mind reasoning as training on the most probable answers biases the model towards generic vocabulary which had the highest frequency in the training data. As a consequence the comprehension of rarer domain-specific concepts which are present in MMLU and MERA datasets may be inhibited due to increased tendency of using more common language. The issue can be alleviated by more complex example sampling procedures such a beam search or multi-candidate generation with post-generation ranking with larger state-of-the-art LLMs such as GPT-4 or LLaMa-3-405B.

Case Study: Continued Instruction-Tuning Calibration

Our experiments on continued instruction-tuning calibration approach, presented in Table 4, showed that the additionally fine-tuned LEP adapted models achieve and in some cases outperform the original models. Adding 2000 instructions for copying long texts to the instructional dataset has a positive effect in almost all cases. Moreover, the obtained

models are more effective when used in the Russian language, and the loss of initial knowledge in the case of our method is minimal, compared to conventional instruct-tuning.

Examples

We also investigated how the models' responses changed depending on the stage: original model, LEP, LEP + calibration (Figure 3).

From the example, it can be seen that the original model did not correctly perceive the question at all. The LEP model already answers more correctly, but does not take into account that this is a phraseological unit. The calibrated model already answers the question most correctly among the three versions of the model, paying attention to the true meaning of the phrase.

DISCUSSION

LLM Benchmark Results for Russian Language

Results presented in Table 1 demonstrate that fine-tuning of open-source state-of-the-art LLMs on Russian focused instruction datasets commonly leads to performance drops in language understanding. This phenomenon was initially observed within Ru-Arena-General⁴ and Chatbot Arena⁵ benchmarks, however, due to their open-question format it was hard to separate generation errors from bad user prompting. Closed-question benchmarks such as MERA (Fenogena, 2024), which was used as the basis of Darumeru, can not reliably detect language processing degradation due to the possibility of benchmark hacking. Benchmark hacking is a procedure of fine-tuning on benchmark solutions or similar data which is viewed as a variant of cheating in the context of LLM benchmarks. Usually developers of LLM models do not intend to resort to such poor practice and on the contrary make an additional effort to remove any possible benchmark data from the overall LLM training data pool. At the same time detecting benchmark related data-leaks is a labor-intensive task as it requires checking training data not just for exact matches but also for any possible paraphrases which includes translating examples to other languages.

Our Darumeru benchmark addresses the limitation of closed-question format with newly introduced tasks for text summarization (DaruSum) and tokenization diagnostic (DaruCopy). DaruSum requires two crucial task-solving elements, proper text analysis and good text writing skills. Any performance drops in this benchmark subset indicate problems with text understanding or text generation. DaruCopy distinguishes between the two by exclusively evaluating the

⁴ <https://huggingface.co/spaces/Vikhrmodels/arenahardlb>

⁵ <https://lmarena.ai/>

Table 4*Benchmark Results for Model Calibration Schemes of Conversion LEP Models*

Model	Fine-tuning data	Micro-Avg	DaruMMLU	DaruMERA	DaruSum	DaruCopy (EN)	DaruCopy (RU)
OpenChat 3.5							
Original model	-	0,607	0,543	0,526	0,322	0,999	0,917
	saiga d7	0,611	0,540	0,528	0,325	0,999	0,945
	+copy task	0,615	0,541	0,524	0,324	1,000	0,995
Unigram	-	0,565	0,515	0,519	0,301	0,999	0,651
	saiga d7	0,599	0,532	0,556	0,316	0,999	0,754
	+copy task	0,630	0,530	0,559	0,321	1,000	0,999
Extended	-	0,609	0,535	0,541	0,306	0,999	0,909
	saiga d7	0,616	0,543	0,566	0,319	0,999	0,845
	+copy task	0,632	0,541	0,563	0,321	1,000	0,989
LLaMa-3-8B instruct							
Original model	-	0,610	0,571	0,510	0,322	1,000	0,972
	saiga d7	0,615	0,576	0,512	0,329	1,000	0,983
	+copy task	0,616	0,575	0,513	0,332	1,000	0,995
Extended	-	0,597	0,556	0,501	0,318	0,994	0,921
	self-calibration	0,606	0,552	0,512	0,321	1,000	0,958
	saiga d7	0,614	0,568	0,519	0,338	0,995	0,961
	+copy task	0,618	0,565	0,521	0,339	1,000	0,984
Optimized	-	0,598	0,555	0,500	0,324	0,995	0,928
	self-calibration	0,601	0,550	0,501	0,325	1,000	0,95
	saiga d7	0,611	0,555	0,515	0,336	1,000	0,971
	+copy task	0,617	0,555	0,522	0,339	1,000	0,989

latter by reducing the task to explicitly broadcasting the original context without any analysis or paraphrasing. Consequently, lower DaruCopy scores indicate a reasoning conflict within the LLM logic as the model fails to follow simplest task directive of text copying. These two subsets of Darumeru benchmark show that LLaMa-3-8B is a more reliable choice for Russian processing tasks than Saiga or Vikhr-5.2 despite their Russian language specialization which contrasts with the results of MERA benchmark (Fenogenova, 2024). While MERA results of Saiga lie within standard deviation the results of Vikhr-5.2 clearly suggest the case of benchmark hacking.

Language Adaptation Strategy

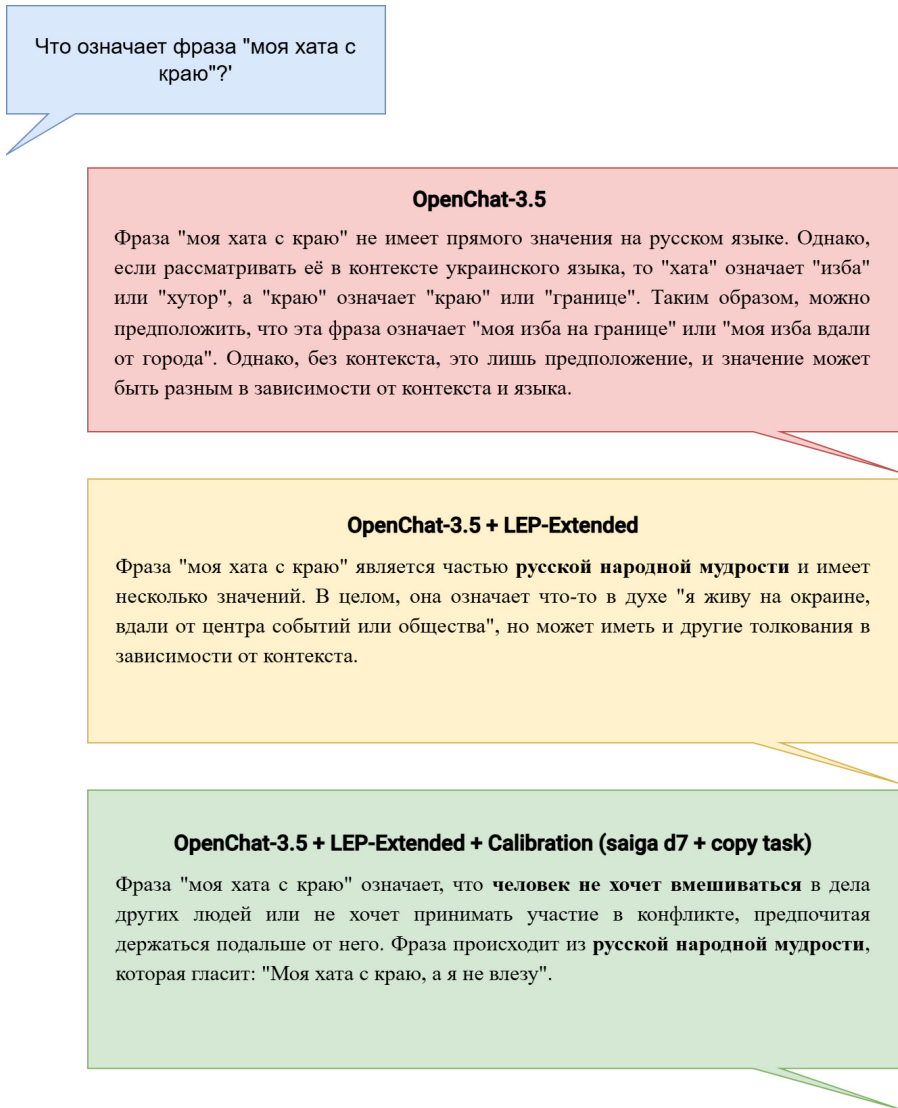
During development of our LLM Russian adaptation pipeline we made several design choices which were explored in previous works. First of all, we assumed that tokenization knowledge and the ability to use new tokens is stored in input embeddings and LM head layers of LLM. Several works (Cui, 2023; Tikhomirov, 2023; Nikolich 2024; Nguyen, 2024) demonstrated that language-adaptation of these subset of

layers only is insufficient for proper language understanding and thus subsequent instruction-tuning of such models leads to suboptimal results. At the same time it was shown (Tikhomirov, 2024) that there is no significant difference between language-adaptation of all-layers and dual-stage approach, when embedding and LM-head training process is complemented with subsequent training of other layers. Results reported in Table 2 reinforce this claim as the first stage of dual-stage approach proves to be efficient enough to substantially improve Russian language comprehension of Mistral-7B model. However, LLaMa-3-8B post-adaptation scores suggest that the necessity of inner-layer training is dictated by the original LLM Russian linguistic skills which are effectively captured by the DaruMERA subset of our benchmark. Learned Embedding Propagation procedure results (see Table 3) also reflect this observation as Mistral-7B showed highest language-knowledge transfer efficiency.

Whether layer discrepancy can be alleviated by instruction-tuning we explored in our calibration experiments. Instruction-tuning on target language often improves token utilization and boosts language comprehension (Gusev,

Figure 3

An Example of Generation Using the OpenChat-3.5 Model and Its Adapted Versions



2023; Wei, 2023; Nikolich, 2024). We see a similar trend in Table 4. By training the original non-adapted instruction-tuned versions of LLMs on Saiga dataset (Gusev, 2023) we enhanced Russian task-solving capabilities which boosted benchmark scores. Applying the same procedure to our LEP models (saiga d7) we retain the positive effect at increased rates with the scores higher than of the original models which were the subjects of LEP knowledge transfer. The drawback of instruction-tuning on Russian instruction datasets is that we inevitably disturb the original knowledge that was gained in prior training (Tejaswi, 2024). We attempted to address the issue by training on the answers generated by the original LLM (self-calibration) rather than using the original references from the Saiga dataset. However for LLaMa-3-8B instruct we did not see noticeable improvement in any LLM capabilities besides tokenization utilization (Daru-Copy). This result is likely due to lack of generation quality of our self-calibration synthesized examples which during our manual inspection revealed to carry much simpler Russian

language logic and vocabulary. Considering that Saiga is a prime example of GPT-4 reference synthesis (Taori, 2024) we hypothesize that by utilizing more advanced sampling techniques and better example quality evaluation protocols we may collect a reference dataset with the similar features without employment of other datasets or third-party models.

LIMITATIONS

Despite the broad applicability of our method, this study has several limitations. First, the method requires that not only instructional versions of LLM but also their foundational versions be available, which is not always the case. Secondly, in the case of languages using hieroglyphs, initialization after tokenizer replacement can be quite weak due to lack of shared tokens and it is not known how much adaptation of embeddings can help with this. Another important point

is that the focus of the knowledge transfer procedure was on preserving the original knowledge of the target model which is why the possible volume of transferred knowledge may be insufficient. However, since the methodology effectively adapts the model to the language, it is always possible to conduct an additional stage of continuous pretraining to acquire new knowledge.

CONCLUSION

In this paper, we proposed Learned Embedding Propagation (LEP), an improved approach to large language model (LLM) language adaptation that has minimal impact on LLM inherent knowledge while enabling transferring the language-adaptation knowledge directly to any instruct-tuned version, including the proprietary. Focussing on cost-efficiency of our method we derived 3 ad-hoc approaches for the embedding propagation: Direct Embedding Swap, Overlapping Token Correction and Vocabulary Conversion. To facilitate the development process of optimal Russian adaptation we introduced Darumeru, a train-time benchmark which focuses on text generation reliability. By analyzing the benchmark performance of popular instruction-tune LLMs and 4 vocabulary adaptation options we derived a recipe for the most cost-efficient procedure. Using the recipe and the proposed LEP methods we built language-adapted variants of sub-9B parameter state-of-the-art instruction-tuned LLMs, Openchat-3.5 and LLaMa-3-8B (Instruct). The evaluation results demonstrated that the Vocabulary Conversion LEP variants reproduce the performance levels of the original instruction-tuned LLM and in the case of OpenChat-3.5 even outperform while having all benefits of improved computational efficiency. To close the remaining gaps in task-solving performance we conducted case-study experiments on self-calibration and continued instruct-tuning alignment approaches which concluded with further language com-

prehension improvements and new benchmark records. The obtained results open new prospects for LLM language adaptation enabling cost-efficient utilization of any instruction-tuned models regardless of openness of their fine-tuning data with all the merits of the original version.

All our models, benchmark and framework are open source and available under the original model licenses.

ACKNOWLEDGEMENT

The work of Mikhail Tikhomirov was supported by Noncommercial Foundation for Support of Science and Education "INTELLECT". The work of Daniil Chernyshev was supported by Noncommercial Foundation for Support of Science and Education "INTELLECT". The research was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHOR CONTRIBUTIONS

Mikhail Tikhomirov: Conceptualization; Investigation; Methodology; Project administration; Software; Writing – original draft.

Daniil Chernyshev: Conceptualization; Data curation; Formal analysis; Methodology; Validation; Visualization; Writing – original draft; Writing – review & editing.

REFERENCES

- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). *Stanford alpaca: An instruction-following llama model*. https://github.com/tatsu-lab/stanford_alpaca
- Li, H., Koto, F., Wu, M., Aji, A. F., & Baldwin, T. (2023). Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. arXiv preprint arXiv:2305.15011. <https://doi.org/10.48550/arXiv.2305.15011>
- Wei, X., Wei, H., Lin, H., Li, T., Zhang, P., Ren, X., Li, M., Wan, Y., Cao, Z., Xie, B., Hu, T., Li, S., Hui, B., Yu, B., Liu, D., Yang, B., & Xie, J. (2023). Polylm: An open source polyglot large language model. arXiv:2307.06018. <https://doi.org/10.48550/arXiv.2307.06018>
- Gusev, I. (2023). *ruLM: A toolkit for training neural language models*. <https://github.com/IlyaGusev/ruLM>.
- Kuulmets, H. A., Purason, T., Luhtaru, A., & Fishel, M. (2024, June). Teaching Llama a new language through cross-lingual knowledge transfer. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3309-3325). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-naacl.210>
- Zhu, W., Lv, Y., Dong, Q., Yuan, F., Xu, J., Huang, S., Kong, L., & Li, L. (2023). *Extrapolating large language models to non-english by aligning languages*. arXiv:2308.04948. <https://doi.org/10.48550/arXiv.2308.04948>
- Ranaldi, L., Pucci, G., & Freitas, A. (2023). *Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations*. arXiv:2308.14186. <https://doi.org/10.48550/arXiv.2308.14186>

- Li, C., Wang, S., Zhang, J., & Zong, C. (2024, June). Improving in-context learning of multilingual generative language models with cross-lingual alignment. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (vol. 1: Long Papers, pp. 8051-8069). Association for Computational Linguistics. 10.18653/v1/2024.naacl-long.445
- Chai, L., Yang, J., Sun, T., Guo, H., Liu, J., Wang, B., Liang, X., Bai, J., Li, T., Peng, Q., & Li, Z. (2024). *xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning*. arXiv preprint arXiv:2401.07037. <https://doi.org/10.48550/arXiv.2401.07037>
- Husain, J. A., Dabre, R., Kumar, A., Puduppully, R., & Kunchukuttan, A. (2024). *RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models via Romanization*. arXiv:2401.14280. <https://doi.org/10.48550/arXiv.2401.14280>
- Lakew, S. M., Erofeeva, A., Negri, M., Federico, M., & Turchi, M. (2018). Transfer learning in multilingual neural machine translation with dynamic vocabulary. *Proceedings of the 15th International Conference on Spoken Language Translation* (pp. 54-61). International Conference on Spoken Language Translation. <https://doi.org/10.48550/arXiv.1811.01137>
- Kuratov, Y., & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. *Komp'uternaja Lingvistika i Intellektual'nye Tehnologii* (pp. 333-339). Komp'uternaja Lingvistika i Intellektual'nye Tehnologii. <https://doi.org/10.48550/arXiv.1905.07213>
- Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2021, August). How good is your tokenizer? On the monolingual performance of multilingual language models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (vol. 1: Long Papers, pp. 3118-3135). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.243>
- Yang, Z., Xu, Z., Cui, Y., Wang, B., Lin, M., Wu, D., & Chen, Z. (2022, October). CINO: A Chinese minority pre-trained Language Model. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 3937-3949). International Committee on Computational Linguistics. <https://doi.org/10.48550/arXiv.2202.13558>
- Vries, W., & Nissim, M. (2021, August). As good as new. How to successfully recycle English GPT-2 to make models for other languages. *Findings of the Association for Computational Linguistics: ACL/JC/NLP 2021* (pp. 836-846). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.findings-acl.74>
- Tikhomirov, M., & Chernyshev, D. (2023). Impact of tokenization on LLaMa Russian adaptation. *2023 Ivannikov Ispras Open Conference* (pp. 163-168). IEEE. <http://dx.doi.org/10.1109/ISPRAS60948.2023.10508177>
- Tikhomirov, M., & Chernyshev, D. (2024). Improving Large Language Model Russian adaptation with preliminary vocabulary optimization. *Lobachevskii Journal of Mathematics*, 45, 3211-3219. 10.1134/S1995080224604120.
- Cui, Y., Yang, Z., & Yao, X. (2023). *Efficient and effective text encoding for Chinese llama and alpaca*. arXiv:2304.08177. <https://doi.org/10.48550/arXiv.2304.08177>
- Nguyen, X. P., Zhang, W., Li, X., Aljunied, M., Tan, Q., Cheng, L., Chen, G., Deng, Y., Yang, S., Liu, C., Zhang, H., & Bing, L. (2023). *SeaLLMs-Large Language Models for Southeast Asia*. arXiv:2312.00738. <https://doi.org/10.48550/arXiv.2312.00738>
- Nikolich, A., Korolev, K., & Shelmanov, A. (2024). *Vikhr: The family of open-source instruction-tuned Large Language Models for Russian*. arXiv preprint arXiv:2405.13929. <https://doi.org/10.48550/arXiv.2405.13929>
- Artetxe, M., Ruder, S., & Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (pp. 4623-4637). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.421>
- Chen, Y., Marchisio, K., Raileanu, R., Adelani, D., Saito Stenetorp, P. L. E., Riedel, S., & Artetxe, M. (2023). Improving language plasticity via pretraining with active forgetting. *Advances in Neural Information Processing Systems*, 36, 31543-31557. <https://doi.org/10.48550/arXiv.2307.01163>
- Tejaswi, A., Gupta, N., & Choi, E. (2024). *Exploring design choices for building language-specific LLMs*. arXiv preprint arXiv:2406.14670. <https://doi.org/10.48550/arXiv.2406.14670>
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., & Levy, O. (2024). Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36. <https://doi.org/10.48550/arXiv.2305.11206>
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... & Ganapathy, R. (2024). *The Llama 3 Herd of Models*. arXiv:2407.21783. <https://doi.org/10.48550/arXiv.2407.21783>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models*. arXiv:2001.08361. <https://doi.org/10.48550/arXiv.2001.08361>
- Ilharco, G., Ribeiro, M. T., Wortsman, M., Schmidt, L., Hajishirzi, H., & Farhadi, A. Editing models with task arithmetic. *The Eleventh International Conference on Learning Representations*. International Conference on Learning Representations. <https://doi.org/10.48550/arXiv.2212.04089>

- Gusev, I. (2020). Dataset for automatic summarization of Russian news. In *Artificial Intelligence and Natural Language: 9th Conference* (Proceedings 9, pp. 122-134). Springer International Publishing. https://doi.org/10.1007/978-3-030-59082-6_9
- Dubois, Y., Galambosi, B., Liang, P., & Hashimoto, T. B. (2024). *Length-controlled alpacaeval: A simple way to debias automatic evaluators*. arXiv:2404.04475. <https://doi.org/10.48550/arXiv.2404.04475>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744. <https://doi.org/10.48550/arXiv.2203.02155>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Roziere, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023a). *Llama: Open and efficient foundation language models*. arXiv:2302.13971. <https://doi.org/10.48550/arXiv.2302.13971>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023b). *Llama 2: Open foundation and fine-tuned chat models*. arXiv:2307.09288. <https://doi.org/10.48550/arXiv.2307.09288>
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B*. arXiv:2310.06825. <https://doi.org/10.48550/arXiv.2310.06825>
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). *Gpt-4 technical report*. arXiv:2303.08774. <https://doi.org/10.48550/arXiv.2303.08774>
- Fenogenova A. et al. (2024). *Mera: A comprehensive LLM evaluation in Russian*. arXiv:2401.04531. <https://doi.org/10.48550/arXiv.2401.04531>
- Mikhailov, V., Shamardina, T., Ryabinin, M., Pestova, A., Smurov, I., & Artemova, E. (2022). RuCoLA: Russian corpus of linguistic acceptability. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 5207-5227). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-main.348>

<https://doi.org/10.17323/jle.2024.22221>

Predictions of Multilevel Linguistic Features to Readability of Hong Kong Primary School Textbooks: A Machine Learning Based Exploration

Zhengye Xu , Yixun Li , Duo Liu 

The Education University of Hong Kong, Tai Po, N.T., Hong Kong, China

ABSTRACT

Introduction: Readability formulas are crucial for identifying suitable texts for children's reading development. Traditional formulas, however, are linear models designed for alphabetic languages and struggle with numerous predictors.

Purpose: To develop advanced readability formulas for Chinese texts using machine-learning algorithms that can handle hundreds of predictors. It is also the first readability formula developed in Hong Kong.

Method: The corpus comprised 723 texts from 72 Chinese language arts textbooks used in public primary schools. The study considered 274 linguistic features at the character, word, syntax, and discourse levels as predictor variables. The outcome variables were the publisher-assigned semester scale and the teacher-rated readability level. Fifteen combinations of linguistic features were trained using Support Vector Machine (SVM) and Random Forest (RF) algorithms. Model performance was evaluated by prediction accuracy and the mean absolute error between predicted and actual readability. For both publisher-assigned and teacher-rated readability, the all-level-feature-RF and character-level-feature-RF models performed the best. The top 10 predictive features of the two optimal models were analyzed.

Results: Among the publisher-assigned and subjective readability measures, the all-RF and character-RF models performed the best. The feature importance analyses of these two optimal models highlight the significance of character learning sequences, character frequency, and word frequency in estimating text readability in the Chinese context of Hong Kong. In addition, the findings suggest that publishers might rely on diverse information sources to assign semesters, whereas teachers likely prefer to utilize indices that can be directly derived from the texts themselves to gauge readability levels.

Conclusion: The findings highlight the importance of character-level features, particularly the timing of a character's introduction in the textbook, in predicting text readability in the Hong Kong Chinese context.

KEYWORDS

Chinese, linguistic features, Random Forest, readability models, Support Vector Machine

Citation: Xu, Z., Li, Y., & Liu, D. (2024). Predictions of multilevel linguistic features to readability of Hong Kong primary school textbooks: A machine learning based exploration. *Journal of Language and Education*, 10(4), 146-158. <https://doi.org/10.17323/jle.2024.22221>

Correspondence:
Duo Liu,
duoliu@eduhk.hk

Received: August 14, 2024

Accepted: December 16, 2024

Published: December 30, 2024

INTRODUCTION

Text readability refers to the ease with which a text can be read and understood (Crossley et al., 2019). A number of studies across languages have found that reader-level characteristics, such as linguistic knowledge and motivation, can influence text readability (Stutz et al., 2016; Zhang et al., 2014). At the same time, text-level linguistic features, such as word frequency and sentence length,

also play pivotal roles in text readability (Crossley et al., 2023; Mesmer, 2005). While reader-level characteristics have been extensively explored in reading research (McBride-Chang et al., 2005; Stutz et al., 2016), text-level features in the context of text readability, particularly in the Chinese language, have received less attention (Crossley et al., 2023; Fitzgerald et al., 2015; Sung et al., 2015). To address this gap, this study examined how text-level features affect text readability



in Hong Kong primary school Chinese textbooks. The findings aim to improve the alignment between text and children's reading abilities, thereby enhancing learning efficiency in Chinese.

Text Readability and Linguistic Features at Different Levels

Text readability can be quantified by constructing a readability formula (Crossley et al., 2019), which provides an overview of text difficulty. It shows promise in benchmarking children's text-difficulty ability levels more accurately, thus allowing them to read texts at target readability levels. These formulas typically result in an absolute score or a grade level that indicates the level of text an average reader in that grade is expected to be able to read and understand successfully (Kincaid et al., 1975; Solnyshkina et al., 2017). For example, one of the most well-known readability formulas, the Flesch-Kincaid grade level formula (Kincaid et al., 1975), $(0.39 \times \text{the average number of words used per sentence}) + (11.8 \times \text{the average number of syllables per word}) - 15.59$, is designed to result in a grade level that indicates a text's readability. For example, a score of 5.3 indicates that the text is appropriate for fifth graders.

These formulas usually consider a few linguistic features, however, research has shown that features relating to word, syntax, and discourse levels significantly affect text comprehension in various languages, such as English and Chinese (Crossley et al., 2019; Liu et al., 2024; Pinney et al., 2024; Solnyshkina et al., 2017). At the word level, word length, i.e., the number of characters per word, is a key indicator of text readability. Longer words typically signify more challenging texts, while shorter words suggest easier comprehension (Crossley et al., 2023; Mesmer & Hiebert, 2015). Word diversity, which reflects the range of different words used in a text, also influences readability (Sung et al., 2015). Word frequency and psycholinguistic-related indexes, particularly reaction time and error rate in lexical decision tasks, have been associated with text readability (Tsang et al., 2018; Tse et al., 2017). In addition to being recognized, the meanings of words are required for successful understanding, resulting in a role for semantic information of words in text reading (Mesmer & Hiebert, 2015). Also, part-of-speech of words (the grammatical category or classification of words in a language based on their functions and roles within a sentence, e.g., nouns and verbs) influence text readability since higher readability levels of texts are generally associated with higher proportions of conjunction words and adverbs, whereas lower readability text levels are linked to higher proportions of adjectives and modal words (Liu et al., 2024).

At the syntax level, sentence length is important. Longer sentences and greater distances between related words in a sentence imply higher syntactic complexity (Crossley et al., 2023). Word dependency, the average distance between two related words in a sentence, has also been found to be related to syntactic complexity (Crossley et al., 2023). A sentence can

be easier if the average distance between two related words is shorter (Crossley et al., 2019). Sentence grammar, which encompasses logical relationships within a sentence, can also contribute to the complexity of syntax (Graesser et al., 2011).

Discourse-level factors, primarily the relationships between the sentences of a text, also impact readability (Pinney et al., 2024). These discourse structures, tied to text cohesion, can influence how clearly a noun, pronoun, or noun phrase can be linked to another element (Givón, 1995). Causal cohesion, related to connective indices, can reduce text readability by building relationships between words, concepts, and paragraphs (Graesser et al., 2011).

Text Readability Research in Chinese

Text readability research in Chinese incorporates word, syntax, and discourse-level features as in alphabetic languages, but also considers character-level features due to that the character is the basic writing unit in the Chinese language (Cheng et al., 2020; Sung et al., 2015). Specifically, a character can stand alone to form a one-character word (e.g., 筆/bat1/pen) or can be combined with others to form two-character words (e.g., 筆記/bat1-gei3/ note), or three- or more-character words (e.g., 筆記本/bat1-gei3-bun2/ notebook). Each Chinese character has its own form, sound, and meaning(s); therefore, related linguistic features attached to characters can influence text readability (Sung et al., 2015). Traditional Chinese text readability formulas include character-level features, like the average number of characters, but often overlook other influential factors, particularly at the discourse level, such as text cohesion (Cheng et al., 2020; Jing, 1995). They also assume a linear relationship between readability level and linguistic features, limiting the accuracy of the model (Rodriguez-Galiano et al., 2015).

To address these issues, machine learning techniques have been employed to improve readability estimation. Unlike traditional formulas, machine learning can handle a large number of linguistic features and identify complex relationships among them (Rodriguez-Galiano et al., 2015). This approach presents the predicted readability level as a category (e.g., Grade 5), indicating the appropriate reading level for readers in that grade. In addition to aiding reader-text matching, the machine learning approach can enhance our understanding of text readability by identifying key linguistic features (Rodriguez-Galiano et al., 2015). For instance, Fitzgerald et al. (2015) analyzed 238 features and determined that nine features related to word structure, semantics, and cohesion were crucial for understanding English text complexity. Therefore, the current study employed machine learning approaches to provide insights into text readability.

Machine Learning Based Text Readability Formulas in Chinese

Machine learning techniques have been utilized to explore text readability in Chinese, with studies primarily focusing

on the support vector machine (SVM) algorithm (e.g., Chen et al., 2011; Sung et al., 2015; Wu et al., 2020). These studies, mostly conducted in Taiwan, used traditional Chinese writing systems and multilevel linguistic features to train SVM models for classifying text readability, achieving high accuracy rates for lower (first and second, 95%) and middle-grade levels (third and fourth grades, 84%; Chen et al., 2013). For a more nuanced classification, using the grade level (i.e., Grades 1–6) as the indicator for text readability, Sung et al. (2015) combined SVM with 31 linguistic features from lexical, semantic, syntax, and discourse levels. Sung et al. (2015) found that models incorporating features from multilevel offered higher accuracy in predicting text readability (71.75%) than models using features from a single level (43.97%–65.13%).

Research on text readability in simplified Chinese, predominantly used in Mainland China, has also been conducted. Wu et al. (2020) utilized SVM models to examine the impact on text readability of 104 linguistic features from character, word (comprising two or more characters), syntax, and discourse. The findings of Wu et al. (2020) indicated that among the models with single-level features, the word-level features (accuracy: 62.1%) performed the best. Moreover, the inclusion of character (accuracy: 63.8%) and syntax (accuracy: 63.1%) level features improved prediction accuracy more than the word-level model did.

A recent study (Liu et al., 2024) examined linguistic features on simplified Chinese text readability using a detailed semester-level scale (i.e., 1–12). They used the random forest (RF) and SVM algorithms along with numerous lexical and discourse features, confirming that models using features from multiple levels outperformed those using features from a single level, with higher accuracies (RF: 27%; SVM: 28%) and lower mean absolute error, the average absolute difference between the true and predicted readability levels (MAE, RF: 1.24; SVM: 1.25). Furthermore, Liu et al. (2024) identified that character and word frequency, semantic features, lexical diversity, syntactic categories, and referential cohesion were the most important features.

However, compared to the situations in Mainland China and Taiwan, less attention has been paid to text readability in Hong Kong, where the Chinese community has unique text-related features that differ from those in Mainland China and Taiwan (McBride-Chang et al., 2005). To address this gap, the current study aimed to develop an appropriate model for approximating text readability in Hong Kong.

The Present Study

The present study focuses on text readability in Hong Kong, where the traditional writing system is used, and texts are processed in Cantonese, differing from Mainland China and Taiwan. Cantonese possesses some unique features, such as additional tones and vocabulary, specific spoken

language terms, and regional variations. For example, the character 是/si6/ is used in written language, while 係/hai6/ is more commonly used in spoken language to express the meanings of *yes*. Moreover, in the spoken language, Cantonese has some words to indicate the ends of utterances, such as 啊/aa3/, 㗎/gaa3/, and 囉/lo1/, which are not commonly used in formal books. Also, some terms used in Hong Kong differ from those used in Mainland China and Taiwan. For instance, the concepts of *bus* and *taxi* are often represented by 巴士/baa1-si6/ and 的士/dik1-si6/, respectively, in Hong Kong. However, they are expressed as 公交/gong1-jiao1/ and 出租车/chu1-zu1-che1/, respectively, in Mainland China, and 公車/gong1-che1/ and 計程車/ji4-cheng2-che1/, respectively, in Taiwan. These features of Cantonese make it necessary to develop readability formulas using a corpus developed with locally used texts (McBride-Chang et al., 2005).

This study uses a corpus of articles from Chinese language arts textbooks commonly used in Hong Kong, particularly for primary school students. Following previous studies in the Chinese language (e.g., Liu et al., 2024; Sung et al., 2015), the study incorporates linguistic features from character, word, syntax, and discourse levels to estimate text readability. It employs a more nuanced scale based on semesters, with a readability level scale of 1–24. The study also uses a subjective indicator, teacher-rated semesters, for each selected text. SVM and RF were adopted, and the importance of linguistic features was analyzed to comprehensively understand text readability in the Chinese language. The current study sought to answer two research questions: 1) Whether and to what extent do the levels of features affect text readability models' performance? 2) What are the features that are most important to the current best model(s)?

METHOD

Study Design

This study utilized a text corpus from Chinese language arts textbooks for primary school students published by three major Hong Kong publishers. Due to copyright issues, the text materials can not be publicly shared. Each publisher contributed four textbooks per grade level, divided into two textbooks per semester, yielding a total of 72 textbooks. Two research assistants meticulously digitalized and proofread the texts three times to ensure accuracy. The study considered 723 texts after excluding non-passage elements such as ancient Chinese prose, illustrations, tables of contents, bibliographies, and indexes. Then, linguistic features to represent character-, word-, syntax-, and discourse-level characteristics of each text were extracted and calculated using the CKIP Chinese word segmentation system (Ma & Chen, 2005). This study was approved by the Human Research Ethics Committee of The Education University of Hong Kong and conformed to the Declaration of Helsinki.

The study used machine learning models built with scikit-learn version 1.1.2 in Python 3.10 to explore the predictive roles of multiple levels of linguistic features in text readability (Pedregosa et al., 2011). Text readability was represented by two indicators: publisher-assigned semester (Y1) and teacher-rated semester (Y2). The teacher-rated semester was the average ratings of text readability levels of 11 experienced primary school teachers (whose written informed consent to participate in the study was obtained), using a 1-9 scale tailored for an average reader in the corresponding grade for these 11 teachers. We then assigned a teacher-rated semester to the texts in each publisher-assigned semester based on the rearranged average ratings within each grade. Both Y1 and Y2 ranged from 1-24, with higher values indicating greater text readability. A total of 15 combinations of linguistic features (referred to as Xs) at different levels were developed: character (C), word (W), syntax (S), and discourse (D). These included single-level Xs (C, W, S, D), two-level Xs (C_W, C_S, C_D, W_S, W_D, S_D), three-level Xs (C_W_S, C_W_D, C_S_D, W_S_D), and a four-level X (Xall).

According to Liu et al. (2024) and Sung et al. (2015), two machine learning algorithms, SVM and RF, were employed. A five-fold cross-validation approach was used to evaluate the performance of the machine learning models. The 723 texts were randomly divided into five subsets, with each subset containing an equal percentage of texts from each semester. Four subsets were used for training, and one subset was used for testing in each iteration. The predicted Y values from the models were compared to the actual Y values to assess accuracy and MAE. Linear mixed models (LMMs) were constructed with the *lmer* package in R 4.0.3 to compare the prediction performance (Baayen et al., 2008). The LMMs used z-score transformations to address collinearity and included X, the machine learning algorithm, and their interaction as fixed factors. RF and Xall were used as reference levels. Random intercepts and slopes were included, and a more complex model was accepted if it improved the fit.

After comparing the readability models, the best model(s) for predicting publisher-assigned and teacher-rated semesters were chosen. Then, the importance of each feature was ascertained using *permutation importance* in the Python *ELI5* package, in which feature importance is estimated by measuring how predictor power decreases when a feature is not available (Korobov & Lopuhin, 2019). The loss of predictive power was evaluated by both accuracy and MAE. The data and analysis code are openly available in Open Science Framework.¹

Linguistic Features

Character-Level Features

In total, 110 character-level features, relating to four aspects: 1) character diversity (N = 4), 2) character structural complexity (N = 8), 3) character frequency (N = 40), and 4) psycholinguistic information for characters (N = 58), were calculated for each text.

Four indicators were considered for character diversity: the raw number of the token (for all characters) count, the raw number of type (for different characters) count, the ratio of type count to token count, and the proportion of characters that only occur once. For example, in one of the texts, «我和妈妈玩捉迷藏,» there are eight token characters. Since the third and fourth characters of the sentence are the same, i.e., 妈, there are seven types of characters.

Character structural complexity was measured with four indicators: the average number of strokes, the proportion of characters with less than ten strokes, between 10 and 20 strokes, and more than 20 strokes. Token count and type count were calculated for each indicator, resulting in eight lexical features for character structural complexity.

Character frequency was measured using five corpora: The Balanced Corpus of Modern Chinese, CNCORPUS (Jin et al., 2005), The SUBTLEX-CH corpus (Cai & Brysbaert, 2010), Sinitic Corpus (Huang, 2006), Chinese text computing (Da, 2004), and Hong Kong, Mainland China & Taiwan: Chinese character frequency (this corpus will be identified as HK-MCT hereafter²). Characters that were not found in a given corpus were considered difficult characters. Four indicators were considered: the average frequency scores for frequent characters, the standard deviation of frequency scores for frequent characters, the raw number of difficult characters, and the proportion of difficult characters. Token count and type count were used to calculate these indicators.

Psycholinguistic information for characters was assessed using 26 indicators from previous studies (i.e., Liu et al., 2007; Su et al., 2023). These indicators were: the age at which it is expected a particular character can be learned, character familiarity, the ease of describing the meaning of a character, the ease of creating an image of a character, the grade and semester in which a character is first introduced in the textbook, the number of meanings of a character, the number of homophones of a character, the summed frequency of

¹ https://osf.io/adqw7/?view_only=c4343a96bd86419b88a8e11d1e0c4426

² <https://humanum.arts.cuhk.edu.hk/Lexis/chifreq/>

all characters that share the same pronunciation (calculated based on the five aforementioned corpora), the number of words that a character can form, the summed frequency of all words that contain a character (calculated based on the five aforementioned corpora), the availability of pronunciation cues in a character (1 for a reliable cue, 0 for the absence of cues, and -1 for unreliable/misleading cues), and the reaction times and error rates for character naming by Chinese adults.

Two indicators concerning pronunciation cues in characters were included: the proportion of characters with reliable pronunciation cues (the pronunciation cue has the same pronunciation as the character) and those with unreliable/misleading pronunciation cues (the sounds of the pronunciation cue and its corresponding character are different; Su et al., 2023). Semantic radical transparency of characters, which refers to the degree of meaning correspondence between the semantic radical and the whole character, was also involved. For instance, while both 海/hoi2/ (sea) and 测/cak1/ (measure) contain the semantic radical 水/seoi2/ (water), the former is semantically transparent and the latter opaque. Token count and type count were calculated for each indicator, resulting in 58 lexical features for this category.

Word-Level Features

A total of 105 word-level features was calculated for each text, covering six aspects: word length ($N = 12$), word diversity ($N = 4$), word frequency ($N = 32$), psycholinguistic information for words ($N = 22$), set structure ($N = 1$), and part-of-speech syntactic categories ($N = 34$).

For word length, six facets were considered: the average word length and the percentages of one-character, two-character, three-character, four-character, and five-or-more-character words in a text. Token count and type count were calculated for each facet, resulting in 12 linguistic features.

Word diversity refers to the richness of words in a text. Four indicators were considered: the raw number of words (both token count and type count), the ratio of type count to token count, and the proportion of words that only occurred once.

Word frequency was measured based on the frequency of a word in five corpora, except for HKMCT, as it does not have statistics for word frequency. Four indicators were considered: average frequency scores for frequent words, standard deviation of frequency scores for frequent words, raw number of difficult words, and proportion of difficult words. Token count and type count were used to calculate these indicators, resulting in 32 lexical features.

Psycholinguistic information for words was calculated based on the corpus MELD-SCH (Tsang et al., 2018), which provided reaction times and error rates for Chinese adults. Twelve features were calculated based on the mean and standard deviation of reaction times and error rates. Words that are not included in MELD-SCH were considered low-frequent words, which were identified by their proportions and the raw numbers. The semantic radical transparency of two-character words was extracted based on the work of Su et al. (2023). Three indicators were considered for each character and for the word as a whole. For all psycholinguistic features, both token count and type count were calculated for each facet.

The set structure was measured by calculating the raw number of named entities (e.g., the names of people, organizations, and locations) in a text using HanLP³. Part-of-speech syntactic categories were determined by assigning one of 16 categories to each word using the Natural Language Processing & Information Retrieval Sharing Platform (Liu et al., 2004). The 16 categories include nine types of content words (nouns, verbs, adjectives, numerals, quantifiers, pronouns, time words, place words, and position words) and seven types of function words (adverbs, prepositions, conjunctions, particles, interjections, differentiators, and state words). The raw number and proportion of words for each category were calculated, resulting in 32 features. Additionally, the number and proportion of all content words were calculated, resulting in 34 discourse features.

Syntax-Level Features

At the syntax level, 15 features were considered, focusing on sentence length ($N = 4$), word dependency ($N = 3$), and grammar ($N = 8$). Sentence length was represented by four features: the average numbers of characters and words were calculated separately for each sentence and each clause. Individual sentences and clauses were identified based on the punctuation. A sentence ends with a full stop, exclamation mark, question mark, ellipsis, or dash, whereas a clause ends with a comma, colon, or semicolon (Wang & Wu, 2020).

Word dependency was analyzed on a per-sentence basis. Three indicators were considered within each sentence: the numbers of characters and words before the main verb and the average word distance between any pairs of related words. Related words refer to words that are syntactically governed or dependent on another word.

Four grammar-related indicators reflecting the presence of complex Chinese grammar were considered: negative, metaphorical, passive, and contrastive sentences, based on the Baidu Open Platform (<https://cloud.baidu.com>). We calculated the number and the proportion of each of these four sentence patterns in a text and, therefore, identified eight discourse features concerning grammar.

³ <https://hanlp.hankcs.com/>

Discourse-Level Features

For the discourse-level features, there were 24 features of referential cohesion and 20 features of causal cohesion. Referential cohesion features were included following the work of Graesser et al. (2011). We tracked four types of words: the overlap of all words, content words, nouns, and verbs, with six indicators for each. The six indicators were the proportion of adjacent sentence/paragraph pairs that shared the same words, the proportion of all possible sentence/paragraph pairs that shared the same words, and the weighted proportion of all possible sentence/paragraph pairs that shared the same words (i.e., the distance of two sentences/paragraphs was quantified into the number of sentences between them, and a score of $1/[L + 1]$ was granted when the distance was L sentence). The same calculations were carried out separately for sentences and paragraphs.

Causal cohesion features were adopted from Graesser et al. (2011). These were the raw numbers of precedents (e.g., at first), causes (e.g., because), adversatives (e.g., however), coordinations (e.g., and), additives (e.g., furthermore), successors (e.g., then), inferences (e.g., only if), conditions (e.g., unless), suppositions (e.g., if), concessions (e.g., even though), purposes (e.g., in order to), frequencies (e.g., always), parentheses (e.g., as everyone knows), abandonments (e.g., would rather not), results (e.g., so), comparatives (e.g., rather than), preferences (e.g., instead of),

summaries (e.g., in sum), recounts (e.g., for example) and temporal (e.g., when) connectives.

RESULTS

The Roles of Linguistic Features in Predicting Text Readability in Chinese

The means and standard deviations for accuracy and MAE of the five-fold cross-validation are shown in Table 1. The results of the Y1 and Y2 models were similar (see Figure 1). Across all four LMMs, a significant effect of the machine learning algorithm was observed. RF outperformed SVM in predicting text readability with higher accuracy (Y1: Estimate = -1.27, $SE = 0.25$, $t(145) = -5.08$, $p < .001$; Y2: Estimate = -1.38, $SE = 0.24$, $t(145) = -5.87$, $p < .001$) and lower MAE (Y1: Estimate = 0.76, $SE = 0.19$, $t(139.99) = 3.90$, $p < .001$; Y2: Estimate = 0.56, $SE = 0.18$, $t(145) = 3.08$, $p = .003$). In terms of X, Xall demonstrated superior performance compared to the Xs without character-level features, except for the W_S_D in the Y2 models ($ps > .05$). This was evident in terms of accuracy (Estimates = -1.25 – -3.03, $SEs = 0.24 - 0.25$, $ts = -12.86 - -4.99$, $ps < .001$) and MAE (Estimates = 0.50 – 2.72, $SEs = 0.18$, $ts = 2.76 - 15.21$, $ps < .01$). However, there were no significant differences between Xall and Xs that include the character-level features ($ps > .05$).

Table 1

Means and Standard Deviations of Accuracy (ACC) and Mean Absolute Error (MAE) of All Machine Learning Models

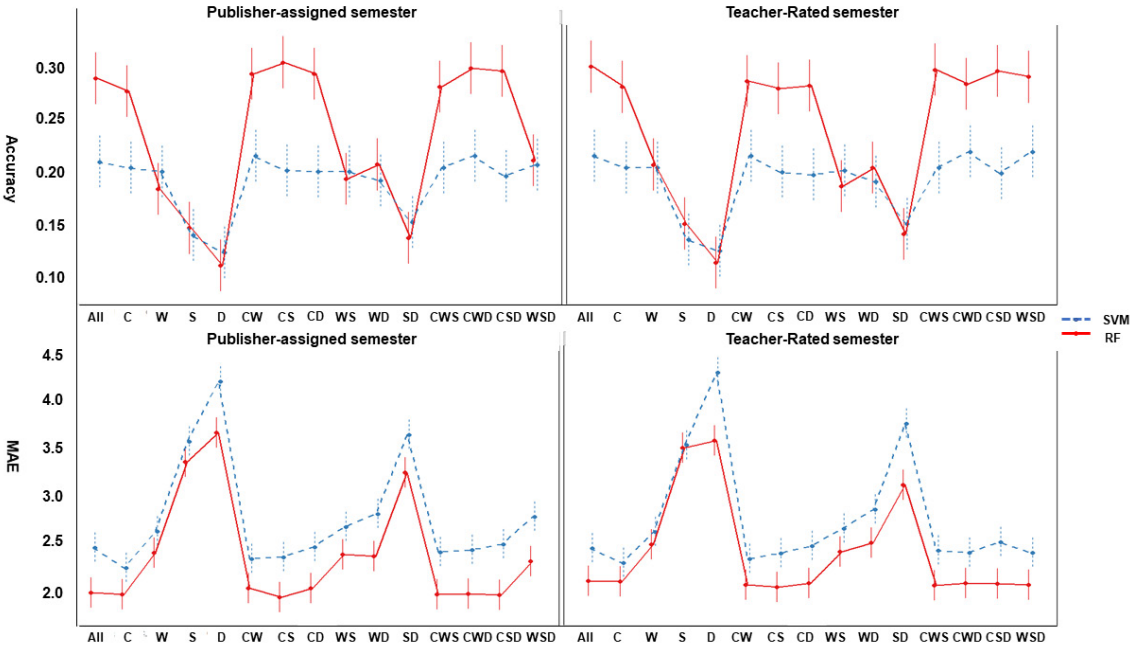
	Y1 (Publisher-assigned semester)			Y2 (Teacher-rated semester)	
	X	ACC	MAE	ACC	MAE
SVM	All	0.21 (0.02)	2.45 (0.18)	0.21 (0.02)	2.45 (0.14)
	C	0.20 (0.02)	2.24 (0.09)	0.20 (0.02)	2.29 (0.14)
	W	0.20 (0.03)	2.63 (0.23)	0.20 (0.02)	2.62 (0.20)
	S	0.14 (0.03)	3.59 (0.33)	0.13 (0.03)	3.55 (0.27)
	D	0.12 (0.02)	4.23 (0.53)	0.12 (0.02)	4.32 (0.54)
	C_W	0.21 (0.04)	2.34 (0.27)	0.21 (0.03)	2.33 (0.22)
	C_S	0.20 (0.02)	2.35 (0.08)	0.20 (0.02)	2.40 (0.11)
	C_D	0.20 (0.03)	2.46 (0.21)	0.20 (0.03)	2.47 (0.20)
	W_S	0.20 (0.03)	2.68 (0.22)	0.20 (0.03)	2.66 (0.22)
	W_D	0.19 (0.04)	2.81 (0.18)	0.19 (0.03)	2.86 (0.12)
	S_D	0.15 (0.02)	3.66 (0.37)	0.15 (0.01)	3.78 (0.36)
	C_W_S	0.20 (0.02)	2.41 (0.18)	0.20 (0.02)	2.43 (0.15)
	C_W_D	0.21 (0.02)	2.43 (0.20)	0.22 (0.02)	2.40 (0.12)
	C_S_D	0.20 (0.03)	2.49 (0.10)	0.20 (0.03)	2.52 (0.07)
	W S D	0.21 (0.03)	2.78 (0.18)	0.22 (0.02)	2.40 (0.12)

	Y1 (Publisher-assigned semester)			Y2 (Teacher-rated semester)	
	X	ACC	MAE	ACC	MAE
RF	All	0.29 (0.04)	1.97 (0.13)	0.30 (0.02)	2.10 (0.19)
	C	0.28 (0.05)	1.96 (0.22)	0.28 (0.04)	2.10 (0.11)
	W	0.18 (0.04)	2.40 (0.04)	0.21 (0.02)	2.49 (0.08)
	S	0.15 (0.03)	3.37 (0.18)	0.15 (0.03)	3.52 (0.25)
	D	0.11 (0.02)	3.69 (0.20)	0.11 (0.02)	3.60 (0.11)
	C_W	0.29 (0.04)	2.02 (0.20)	0.29 (0.03)	2.06 (0.18)
	C_S	0.30 (0.04)	1.93 (0.15)	0.28 (0.02)	2.03 (0.10)
	C_D	0.29 (0.04)	2.02 (0.19)	0.28 (0.03)	2.08 (0.21)
	W_S	0.19 (0.03)	2.38 (0.13)	0.19 (0.04)	2.41 (0.19)
	W_D	0.21 (0.04)	2.37 (0.07)	0.20 (0.05)	2.51 (0.17)
	S_D	0.14 (0.02)	3.26 (0.09)	0.14 (0.03)	3.12 (0.28)
	C_W_S	0.28 (0.04)	1.96 (0.17)	0.30 (0.03)	2.05 (0.10)
	C_W_D	0.30 (0.04)	1.96 (0.15)	0.28 (0.04)	2.08 (0.14)
	C_S_D	0.30 (0.05)	1.95 (0.18)	0.30 (0.05)	2.07 (0.22)
	W_S_D	0.21 (0.04)	2.31 (0.14)	0.29 (0.03)	2.06 (0.13)

Note. SVM = support vector machine; RF = random forest; C = Character features; W = Word features; S = Syntax features; D = Discourse features; All = features at all levels; Letter combinations represent the combination of features at different levels, e.g., C_W = features at Character and Word levels.

Figure 1

Results of Prediction Accuracy and Mean Absolute Error (MAE) in Readability Models Using Different Linguistic Features



Note. SVM = support vector machine; RF = random forest; C = Character; W = Word; S = Syntax; D = Discourse; All = Character_Word_Syntax_Discourse; Letter combinations represent the combination of features at different levels, e.g., C_W = features at Character and Word levels.

The interactions between the machine learning algorithm and the contrasts between Xall and Xs without character-level features, except for W_S_D, were significant in the accuracy models (Estimates = 1.03 – 1.63, *SEs* = 0.33 – 0.35, *ts* = 2.91 – 4.88, *ps* < .01). Regarding MAE, the interactions between the machine learning algorithm and the contrasts of Xall with S (Estimate = -0.52, *SE* = 0.26, *t* = -2.01, *p* = .047) and D (Estimate = 0.61, *SE* = 0.26, *t* = 2.39, *p* = .019) were significant, while the other interactions were not significant (*ps* > .05). Post-hoc analyses indicated that among the RF models, the differences between C and Xall in terms of both accuracy and MAE, were not significant (*ps* > .05). Moreover, C exhibited higher accuracy than the Xs that did not include the character features (Y1: Estimates = 1.12 – 2.66, *SEs* = 0.28, *ts* = 3.99 – 9.48, *ps* < .05; Y2: Estimates = 1.20 – 2.72, *SEs* = 0.26, *ts* = 4.56 – 10.32, *ps* < .01). Additionally, C had lower MAE than those with Xs without the character and word features (i.e., S, D, and S_D, Y1: Estimates = -2.74 – -2.06, *SE* = 0.20, *ps* < .001; Y2: Estimates = -2.43 – -1.66, *SE* = 0.20, *ts* = -8.20 – -11.98, *ps* < .01).

Feature Importance Analyses in the Best-Fitting Models

The LMMs showed that the RF models with all linguistic features (all-RF) and character-level (character-RF) features were optimal for predicting text readability. Feature importance analyses showed that both Y1 and Y2 features were similar, indicating that the character-level features are superior to other features, especially those from the syntax and discourse levels. More importantly, all models highlighted the importance of psycholinguistic information for characters. Specifically, the semester and grade when a character is first introduced in the textbook, measured either in token or type counts, played the most important roles in all optimal models. In Y1, the character-RF models revealed that the reaction times and error rates of character naming by Chinese adults (Liu et al., 2007) were highly important. Age of acquisition was important in three of the four accuracy models, but not in the all-RF model of Y2. The availability of pronunciation cues in a character, the ease of describing the meaning of a character, and the semantic radical transparency of characters were also ranked in the top 10 features in all models, except for the two character-RF models for Y2.

In addition, character frequency was also highlighted across the optimal models. The summed and averaged character frequency was highlighted in all models, although the results that were calculated according to different corpora were selected in different models. The ratio of type count to token count was only selected in the character-RF model of Y2 in terms of MAE. Differing from the models for Y1, three of the four models for Y2 showed that the character structural complexity, i.e., the number of strokes, was important.

At the same time, some word-level features demonstrated high importance. The summed frequency of one-character words and the numbers of different kinds of words (including multiple-character words, one-character words, and low-frequent words) played critical roles in the all-RF models. Two indicators of part-of-speech syntactic categories, i.e., the raw numbers of adverbs and quantifiers, were only selected in the all-RF model of Y1 in terms of MAE. Only one model—the all-RF model of Y2 in terms of ACC—highlighted a discourse-level feature about referential cohesion, i.e., the proportion of all possible paragraph pairs that share the same content words.

DISCUSSION

Using machine learning techniques, research has demonstrated the importance of linguistic features from diverse levels, such as word, syntax, and discourse levels, on text readability (Fitzgerald et al., 2015). In Chinese, studies in Mainland China and Taiwan consistently found that using features from multiple levels outperformed those using features from a single level (e.g., Liu et al., 2024; Sung et al., 2015; Wu et al., 2020). This study was one of the first to investigate text readability in Hong Kong. It extracted 274 linguistic features from 723 digitized Chinese language-arts textbooks commonly used in Hong Kong primary schools, representing character, word, syntax, and discourse levels. Two machine learning algorithms, namely SVM and RF, were utilized to examine the predictive capacity of these features in assessing text readability. The present study extended previous studies in Chinese (e.g., Liu et al., 2024; Sung et al., 2015; Wu et al., 2020) by focusing on a finer, semester-level scale for text readability and introducing a subjective index: the teacher-rated semester level, along with the publisher-assigned semester level. Meanwhile, the important linguistic features for predicting text readability in the context of Hong Kong were identified. The current findings showed that the models with single character-level features and those with multilevel features incorporating character-level features performed similarly to the models with all 274 features. The findings demonstrate the central role of character features in predicting text readability in Chinese. The results of feature importance indicated similarities between the perspectives of publishers and teachers. Results from both perspectives showed that the character-level features, i.e., the semester and grade when a character is first introduced in the textbook, were crucial. Meanwhile, findings from these two perspectives had differences. Models with teacher-rated semesters underscored the importance of the number of strokes, while those with publisher-assigned semesters highlighted the influence from research results, i.e., adults' response times and error rates in lexical decision tasks (Liu et al., 2007).

The Central Role of Character Features in Predicting Text Readability in Chinese

Consistent with previous studies conducted in Mainland China (e.g., Wu et al., 2020) and Taiwan (e.g., Sung et al., 2015), the current findings illustrated that lexical features (i.e., character and word levels) were more advantageous than syntax-level and discourse-level features in determining text readability. More specifically, the models for both publisher-assigned semester level and teacher-rated semester level were similar and demonstrated that models with single character-level features and models with all 274 features performed best in terms of accuracy and MAE in predicting text readability. The RF models further demonstrated that models with character-level features outperformed those without, showing higher accuracy and lower MAE. These similar results suggest that the lexical features had greater effects than the syntax-level and discourse-level features on text readability across Chinese communities using different written and spoken languages. Although it has been found that character-level and word-level features are more predictive of text readability, this doesn't mean syntax-level and discourse-level features have no influence. Prior research highlights the influence of syntactic and discourse skills on reading comprehension (Chik et al., 2012). For instance, a study involving Hong Kong fourth graders (Yeung et al., 2013) revealed that after controlling for word reading, syntactic skills (word-order knowledge, morphosyntactic knowledge) and discourse skills (sentence-order knowledge) uniquely contributed to reading comprehension. Thus, even though character- and word-level features may significantly impact text readability, syntax- and discourse-level features also play a vital role.

On the other hand, the current finding was inconsistent with previous studies in Taiwan (i.e., Chen et al., 2013; Sung et al., 2015), which did not find that models with single-level features could perform as well as those with multiple-level features in predicting text readability. Such a difference might be due to differences in the models' design in the current study compared to previous studies. The present study incorporated both RF and SVM algorithms and used a more granular indicator (semester level) than the grade level used in previous studies. Also, we used more linguistic features (274) compared to previous studies and distinguished between character-level and word-level features, which were not separated in the previous studies.

Differing from the current study, where character-level features outperformed other features in predicting text readability, a study conducted in Mainland China (i.e., Wu et al., 2020), where simplified Chinese is used, found word-level features had an advantage over character-level features. The strong performance of character-level features in our study might be attributable to the use of traditional Chinese in Hong Kong (McBride-Chang et al., 2005). Traditional Chinese characters, known for their ideographic origins,

have a close connection between form and meaning. The simplification process used in Mainland China often weakens this connection, which could make simplified characters more difficult to recognize and read, especially for beginning readers. For example, the simplified character 爱 (love) was developed by removing an element associated with the whole character's meaning, i.e., 心/sam1/ (heart), from its traditional counterpart 愛/oi3/. Studies have shown that children learning simplified characters perform better in visual skill tasks compared to those learning traditional characters (e.g., McBride-Chang et al., 2005). This suggests that some character-level features make traditional characters relatively easy to recognize and read, thereby influencing text comprehension. These aforementioned differences between our study and those conducted in Taiwan and Mainland China may be due to language-related differences between Hong Kong and Taiwan, underscoring the need for specific text readability formulas for different Chinese communities.

Similar and Dissimilar Roles of Features across Perspectives of Publishers and Teachers

The current study advanced previous research (e.g., Liu et al., 2024; Sung et al., 2015; Wu et al., 2020) by incorporating both the publisher-assigned semester level and the teacher-rated semester level as indicators of text readability. The present study revealed consistent findings from both publishers and teachers in terms of feature importance. Specifically, the semester and grade at which a character is first introduced in the textbook significantly impacted text readability. This influence remained notable even when all features across the four aspects were considered. Additionally, the age of acquisition, which correlates with the semester and grade of a character's introduction, was also found to significantly influence text readability. These features indicate the learning sequence of characters, which is not arbitrary. In Hong Kong, publishers are required to refer to *Lexical Lists for Chinese Learning in Hong Kong* (Education Bureau, 2007) when editing textbooks. According to this document, characters that appear at the early stages of learning commonly have lower visual complexity and higher frequencies, e.g., 一/jat1/ (one), 我/ngo5/ (me), and 你/nei5/ (you), than those that are usually taught later, e.g., 勢/sai3/ (power), 滲/sam3/ (seep), and 癒/jyu6/ (heal). This suggests that characters taught at initial stages are designed to be simpler than those introduced later. Moreover, characters taught earlier may also have more exposure, enabling children to better understand them through contextual reading (Brent & Siskind, 2001). Consequently, children might master early-introduced characters, which could enhance text readability.

In line with Liu et al. (2024), the feature-importance analyses highlighted the significance of character frequency and word frequency in text readability. This aligns with previous research showing a strong frequency effect where high-frequency words are read more accurately and faster across

multiple languages (e.g., Cai & Brysbaert, 2010). It was suggested that higher frequencies could facilitate character and word comprehension.

Meanwhile, there were a few differences between the present findings regarding the feature importance for the publisher-assigned semesters and teacher-rated semesters models. Specifically, the features about the number of strokes, which correlate with the visual complexity of characters, featured prominently in the top 10 features of the optimal models for teacher-rated semesters, but not in the models for publisher-assigned semesters. Our feature-importance analysis reported fewer complexity-related features compared to frequency-related ones, suggesting a relatively minor influence of complexity on readability. Consistently, a study on Chinese children (Su & Samuels, 2010) found a diminishing effect of visual complexity on word processing as children's reading skills matured.

On the other hand, the features linked with adults' response times and error rates in lexical decision tasks (Liu et al., 2007) were only observed in the publisher-assigned semesters' analysis. This discrepancy could be attributed to the relative readability for teachers in directly grasping information about adults' response times and error rates in lexical decision tasks, compared to the number of strokes. Consequently, while publishers might rely on diverse information sources to assign semesters, teachers likely prefer to utilize indexes that can be directly derived from the texts themselves to gauge readability levels.

Limitations and Future Directions

As one of the first studies to investigate text readability in Hong Kong, our corpus only covered textbooks from three publishers. Future research could include a wider variety of texts, such as storybooks. Although we engaged experienced teachers to rate the texts considering an average reader at a certain grade, it remains challenging to directly reflect children's readability. Future studies could involve children's ratings and their reading comprehension performances, which are closely related to text readability (Mesmer & Hiebert, 2015). Furthermore, future research could employ additional machine learning algorithms suitable for classification, such as the K-nearest neighbor and decision-tree classifier (Rodriguez-Galiano et al., 2015).

Initially, the accuracy rates of our models were not particularly high, but they improved when the grade level (Grades 1-6) was used as the readability level. Specifically, the SVM and RF models performed similarly. Both models, whether using the single level of character features (SVM: mean accuracy = 66.08%, SD = 0.04; RF: mean accuracy = 70.94%, SD = 0.04) or all features (SVM: mean accuracy = 65.57%, SD = 0.04; RF: mean accuracy = 68.34%, SD = 0.04), performed equally well and outperformed other models that did not include character-level features. Our models achieved accu-

cy rates for grade levels comparable to previous studies conducted in Mainland China (e.g., Wu et al., 2020) and Taiwan (e.g., Sung et al., 2015), which contributes to the existing research on readability across Chinese communities. However, this also emphasizes the need for further exploration of models with finer scales that can achieve higher accuracy in predicting readability. Future studies should focus on investigating such models.

CONCLUSION

The primary aim of this study was to investigate the predictive power of linguistic features at the character, word, syntax, and discourse levels in assigning texts to primary school semester levels. By employing robust machine learning techniques, the study demonstrated the significant predictive power of linguistic features, particularly at the character level. In addition, as a secondary objective, the study analyzed two optimal RF models based on all features and character-level features, which achieved high accuracy and low MAE in predicting semester levels. The feature importance analyses specifically revealed that character learning sequences, character frequency, and word frequency are crucial in predicting text readability. These findings directly address our research questions by identifying the key linguistic features that influence readability assessments from the perspectives of both publishers and teachers.

Practically, these findings offer valuable insights for teaching. Teachers can concentrate on lexical-level features, especially when teaching new characters. Furthermore, future studies could develop an automated text readability analyzer centered on character-level features using the two optimal RF models identified in the current study. Such an analyzer could streamline the semester assignment of textbooks and identify readability levels of texts from other resources, like storybooks. Consequently, children, parents, and teachers could more easily select formal and informal reading materials that align with children's reading abilities.

ACKNOWLEDGEMENTS

This work was partially supported by the Research Seed Fund of the Department of Special Education and Counseling, The Education University of Hong Kong (Ref. No. 04670) to Dr. Duo Liu. This work was also partially supported by a fellowship award to Dr. Zhengye Xu, from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. EdUHK PDFS2122-8H09).

DATA AVAILABILITY STATEMENTS

The data supporting this study's findings are available from <https://osf.io/h9ew4/>

DECLARATION OF COMPETING INTEREST

None declared.

tion; Software; Supervision; Visualization; Writing – original draft; Writing – review & editing.

Yixun Li: Data curation; Investigation; Methodology; Resources; Writing – original draft.

AUTHOR CONTRIBUTIONS

Zhengye Xu: Conceptualization; Data curation; Formal analysis; Funding acquisition; Methodology; Project administration;

Duo Liu: Conceptualization; Data curation; Funding acquisition; Investigation; Methodology; Project administration; Software; Writing – original draft; Writing – review & editing.

REFERENCES

- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390-412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2), B33-B44. [https://doi.org/10.1016/S0010-0277\(01\)00122-6](https://doi.org/10.1016/S0010-0277(01)00122-6)
- Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles [Data set]. *PloS One*, 5(6), Article e10729. <https://doi.org/10.1371/journal.pone.0010729>
- Chen, Y., Chen, Y., & Cheng, Y. (2013). Assessing Chinese readability using term frequency and lexical chain. *Journal of Computational Linguistics & Chinese Language Processing*, 8(2), 1-18.
- Chen, Y., Tsai, Y., & Chen, Y. (2011). Chinese readability assessment using TF-IDF and SVM. *2011 International Conference on Machine Learning and Cybernetics* (pp. 705-710). IEEE. <https://doi.org/10.1109/ICMLC.2011.6016783>
- Cheng, Y., Xu, D., & Dong, J. (2020). 基于语文教材语料库的文本阅读难度分级关键因素分析与易读性公式研究 [A study on the analysis of key factors of text reading difficulty grading and the readability formula based on a corpus of language teaching materials]. *语言文字应用*, 1, 132-143. <https://doi.org/10.16499/j.cnki.1003-5397.2020.01.014>
- Chik, P. P., Ho, C. S., Yeung, P., Chan, D. W., Chung, K. K., Luan, H., Lo, L., & Lau, W. S. (2012). Syntactic skills in sentence reading comprehension among Chinese elementary school children. *Reading and Writing*, 25, 679-699. <https://doi.org/10.1007/s11145-010-9293-4>
- Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, 42(3-4), 541-561. <https://doi.org/10.1111/1467-9817.12283>
- Crossley, S., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. (2023). A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55(2), 491-507. <https://doi.org/10.3758/s13428-022-01802-x>
- Da, J. (2004). A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction [Data set]. *Proceedings of the Fourth International Conference on New Technologies in Teaching and Learning Chinese*, 501-511.
- Education Bureau. (2007). *Lexical Lists for Chinese Learning in Hong Kong*.
- Fitzgerald, J., Elmore, J., Koons, H., Hiebert, E. H., Bowen, K., Sanford-Moore, E. E., & Stenner, A. J. (2015). Important text characteristics for early-grades text complexity. *Journal of Educational Psychology*, 107(1), 4-29. <https://doi.org/10.1037/a0037289>
- Givón, T. (1995). Coherence in text vs. coherence in mind. *Coherence in Spontaneous Text*, 1995, 59-116.
- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-matrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223-234. <https://doi.org/10.3102/0013189X11413260>
- Huang, C. (2006). Automatic acquisition of linguistic knowledge: From sinica corpus to gigaword corpus [Conference presentation] [Data set]. *The 13th National Institute of Japanese Language International Symposium Language Corpora: Their Compilation and Application*. Tokyo.
- Jin, G. J., Xiao, H., Fu, L., & Zhang, Y. F. (2005). 现代汉语语料库建设及深加工 [Construction and further processing of Chinese National Corpus] [Data set]. *语言文字应用*, 2, 111-120. <https://doi.org/10.16499/j.cnki.1003-5397.2005.02.017>
- Jing, X. (1995). 中文国文教材的适读性研究：适读年级值的推估 [A study on the readability of Chinese national language teaching materials: Estimation of readability values of grade levels]. *教育研究资讯*, 5, 113-127.

- Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Flesch-kincaid grade level*. United States Navy.
- Korobov, M., & Lopuhin, K. (2019). *Permutation importance*.
- Liu, M., Li, Y., Su, Y., & Li, H. (2024). Text complexity of Chinese elementary school textbooks: Analysis of text linguistic features using machine learning algorithms. *Scientific Studies of Reading*, 28(3), 235-255. <https://doi.org/10.1080/10888438.2023.2244620>
- Liu, M., Li, Y., Wang, X., Gan, L., & Li, H. (2021). 分级阅读初探: 基于小学教材的汉语可读性公式研究 [Leveled reading for primary students: Construction and evaluation of Chinese reeadability formulas based on textbooks]. *语言文字应用*, 2, 116-126. <https://doi.org/10.16499/j.cnki.1003-5397.2021.02.010>
- Liu, Q., Zhang, H. P., Yu, H. K., & Cheng, X. Q. (2004). 基于层叠隐马模型的汉语词法分析 [Chinese lexical analysis using cascaded hidden Markov model]. *计算机研究与发展*, 41(8), 1421-1429.
- Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese [Data set]. *Behavior Research Methods*, 39(2), 192-198. <https://doi.org/10.3758/BF03193147>
- Ma, W., & Chen, K. (2005). Design of CKIP Chinese word segmentation system. *Chinese and Oriental Languages Information Processing Society*, 14(3), 235-249.
- McBride-Chang, C., Chow, B. W., Zhong, Y., Burgess, S., & Hayward, W. G. (2005). Chinese character acquisition and visual skills in two Chinese scripts. *Reading and Writing*, 18, 99-128. <https://doi.org/10.1007/s11145-004-7343-5>
- Mesmer, H. A. E. (2005). Text decodability and the first-grade reader. *Reading & Writing Quarterly*, 21(1), 61-86. <https://doi.org/10.1080/10573560590523667>
- Mesmer, H. A., & Hiebert, E. H. (2015). Third graders' reading proficiency reading texts varying in complexity and length: Responses of students in an urban, high-needs school. *Journal of Literacy Research*, 47(4), 473-504. <https://doi.org/10.1177/1086296X16631923>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12, 2825-2830.
- Pinney, C., Kennington, C., Pera, M. S., Wright, K. L., & Fails, J. A. (2024). Incorporating word-level phonemic decoding into readability assessment. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, Italia (pp. 8998-9009)*. ELRA and ICCL.
- Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71, 804-818. <https://doi.org/10.1016/j.oregeorev.2015.01.001>
- Solnyshkina, M., Zamaletdinov, R., Gorodetskaya, L., & Gabitov, A. (2017). Evaluating text complexity and Flesch-Kincaid grade level. *Journal of Social Studies Education Research*, 8(3), 238-248.
- Stutz, F., Schaffner, E., & Schiefele, U. (2016). Relations among reading motivation, reading amount, and reading comprehension in the early elementary grades. *Learning and Individual Differences*, 45, 101-113. <https://doi.org/10.1016/j.lindif.2015.11.022>
- Su, I., Yum, Y. N., & Lau, D. K. (2023). Hong Kong Chinese character psycholinguistic norms: Ratings of 4376 single Chinese characters on semantic radical transparency, age-of-acquisition, familiarity, imageability, and concreteness [Data set]. *Behavior Research Methods*, 55(6), 2989-3008. <https://doi.org/10.3758/s13428-022-01928-y>
- Su, Y., & Samuels, S. J. (2010). Developmental changes in character-complexity and word-length effects when reading Chinese script. *Reading and Writing*, 23, 1085-1108. <https://doi.org/10.1007/s11145-009-9197-3>
- Sung, Y., Chen, J., Cha, J., Tseng, H., Chang, T., & Chang, K. (2015). Constructing and validating readability models: The method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods*, 47, 340-354. <https://doi.org/10.3758/s13428-014-0459-x>
- Tsang, Y., Huang, J., Lui, M., Xue, M., Chan, Y. F., Wang, S., & Chen, H. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese [Data set]. *Behavior Research Methods*, 50, 1763-1777. <https://doi.org/10.3758/s13428-017-0944-0>
- Tse, C., Yap, M. J., Chan, Y., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese lexicon project: A megastudy of lexical decision performance for 25,000 traditional Chinese two-character compound words. *Behavior Research Methods*, 49, 1503-1519. <https://doi.org/10.3758/s13428-016-0810-5>
- Wang, F., & Wu, F. (2020). Postnominal relative clauses in Chinese. *Linguistics*, 58(6), 1501-1542. <https://doi.org/10.1515/ling-2020-0226>

-
- Wu S., Yu D., & Jiang X. (2020). 汉语文本可读性特征体系构建和效度验证 [Development of linguistic features system for Chinese text readability assessment and its validity verification]. *世界汉语教学*, 34(1), 81-97.
- Yeung, S. S., Siegel, L. S., & Chan, C. K. (2013). Effects of a phonological awareness program on English reading and spelling among Hong Kong Chinese ESL children. *Reading and Writing*, 26, 681-704. <https://doi.org/10.1007/s11145-012-9383-6>
- Zhang, J., McBride-Chang, C., Wong, A. M., Tardif, T., Shu, H., & Zhang, Y. (2014). Longitudinal correlates of reading comprehension difficulties in Chinese children. *Reading and Writing*, 27, 481-501. <https://doi.org/10.1007/s11145-013-9453-4>

Thanking our Reviewers, 2024

The editors of *Journal of Language and Education* would like to express their gratitude to all the reviewers who helped us work on each issue.

You helped us take JLE from an idea 10 years ago to a full-fledged quality publication indexed in Scopus, Web of Science, DOAJ, EBSCO, etc.

The reviewers' expertise and benevolence contributed greatly to this success. We value and truly appreciate the time and effort of our reviewers who provided insightful analysis of the manuscripts and helped us to assure quality of the journal.

Your reviews became the foundation of every decision the editorial board made regarding the submissions. At the same time, the reviewers' input proved incredibly helpful to the authors who took their manuscripts to the next level of publication quality by responding to your comments and suggestions.

Our reviewers are an essential part of the JLE family, so we would like to take this opportunity to thank you and to express our sincere appreciation of your contribution!

Abdulaziz Ali Al-Qahtani	Abdullah Al Fraidan	Abolfazl Khodamoradi
Afsar Rouhi	Ahmad Ansarifar	Ahmadreza Mohebbi
Ahmed Mohammed Moneus	Aisah Apridayani	Akram Faraj Abdul-Hussein Al-Rubaie
Ali Dabbagh	Ali H. Al-Hoorie	Alina-Andreea Dragoescu Urlica
Aliv Faizal Muhammad	Alpino Susanto	Ambra Fastelli
Amna Brdarević-Čeljo	Andrew Obermeier	Andrey Danilov
Andrés Canga Alonso	Angélica Ribeiro	Anna Gabets
Anna Ivleva	Anna Shevlyakova	Anssi Roiha
Anthony Anggrawan	Anton Afanasyev	Antonina Laposhina
Aretousa Giannakou	Asama Tasanameelarp	Aspasia Chatzidaki
Astrid Tiara Murti	Attila M. Wind	Audi Yundayani
Beatrice Benelli Retired	Bahman Gorjian	Balbina Moncada-Comas
Becky S.C. Kwan	Beyza Uçar	Biljana Radić-Bojanić
Bill Cope	Britt Cole	Budi Waluyo
Candido Jose Ingles Saura	Carmen Luján-García	Carrie Yea-Huey Chang
Christoph A. Hafner	Daniela Coelho	Daniela Wagner-Loera
Dedi Turmudi	Dedy Subandowo	Deniz Gokcora
Dennis Duncan	Dev Raj Paneru	Dr. Iskandar Abdul Samad
Eirene Katsarou	Ekaterini Nikolarea	Elizabeth Hitches
Elke Schneider	Ellianawati M.Si.	Eren Bilen
Esma Senel	Evgenia Theodotou	Fahriye Altınay
Fan Fang	Farzaneh Shakki	Faycal Farhi
Ge Shigang	Guangxiang Leon Liu	Gulten Genc
Hamid Marashi	Hassan Jalali	Heidi Byrnes
Henrietta Lempert	Hessameddin Ghanbar	Hilary Nesi
Hin Tat Cheung	Hossein Salarian	Hossein Bozorgian
Hussein Taha Assaggaf	Ian Moodie	Iryna Lenchuk

Ismail Xodabande	Iulia Dumitraşcu-Băldău	Javad Gholami
Jaya Nur Iman	Jean Parkinson	Jian Li
Jianmin Gao	John Paul C. Vallente	Joshua Matthews
Juhi Kim	Kadriye Dimici	Kang-Young Lee
Karolis Anaktototy	Katsunori Kotani	Kin Chung Jacky Chan
Kittiphong Praphan	Kübra Şik Keser	Lee Huan Yik
Lili Tian	Lilia Raitskaya	Ling Gan
Lirak Karjagdiu	Luan Nguyen	Luis J. Martín-Antón
Lynette Pretorius	Maria Victoria Lopez Perez	Marc Brysbaert
Maria Andreeva	Mariam Orkodashvili	Marina A. Kosycheva
Marina Solnyshkina	Mark Feng Teng	Marta Lacková
Marwa Mekni Toujani	Maryam Kazemi	María Jesús Sánchez
Matthew A. Vetter	Matthew Nall	Maya Lambovska
Meng Ge	Merve Savaşçı	Miguel Fernández Álvarez
Milica Vuković-Stamatović	Mimoza Hasani-Pllana	Mingzhe Wang
Mohamad Yahya Abdullah	Mohamed Ridha Ben Maad	Mohammad Al-Khresheh
Mohammed Abdulkareem A. Alkamel	Mohd Aidil Riduan Awang Kader	Moses Segbenya
Muhammad Sabri Sahrir	Muhammad Shaban Rafi	Muhammad Mran
Natalya Viktorovna Matveeva	Nawal Fadhil Abbas	Nesrine Triki
Nida Boonma	Nurul Aini	Obed Kwame Adzaku Penu
Olga Boginskaya	Omid Noroozi	Pedro Antonio Férez Mora
Pilar Mur-Deñás	Professor Yusri	Qian Lei
Qinqing Zhang	Quang Nam Pham	Quanguo Liu
Rebecca Knoph	Renat Bakirov	Renata Povolná
Reza Pishghadam	Reza Pishghadam	Rizgar Mahmood Qasim
Robert Macintyre	Rong Luo	Ryan Spring
S. Yahya Hejazi	Sabina Valente	Sarah J. Mccarthey
Saraswathy Thurairaj	Sayeed Naqibullah Orfan	Sharif Alghazo
Shashank Parmar	Sheena Gardner	Shizhou Yang
Simpson W.L. Wong	Songyun Chen	Sophia Skoufaki
Syarifudin Syarifudin	Tahmineh Khalili	Takayuki Nakanishi
Talizaro Tafonao	Tang Youjun	Taoufik Boulhrir
Thomas Hon Tung Chan	Tibor Vígh	Tim Stoeckel
Tomoko Sawaki	Tosin Makinde	Tzu Yu Tai
Ushba Rasool	Valentin Malyhk	Valery Solovyev
Vassilios Zorbas	Venera Bayrasheva	Verbra Frances Pfeiffer
Victor K Lai	Victoria Kim	Virginia H.Y. Kwok
Vladimir Ivanov	Wai Mar Phyoo	Weiwei Zhang
Xian Zhao	Xiaoli Su	Xiuwen Chen

Xixin Qiu
Yaqian Shi
Yiran Xu
Yun Jiang
Zahra Akbari

Xuesong Gao
Yayu Shi
Yiyi López Gándara
Yun Wang
Lseyar Alimova

Yang Yang
Yi-Chien Wang
Yongliang Wang
Yutaka Fujieda
Özgür Çelik

CONTENT

EDITORIAL

Lilia Raitskaya, Elena Tikhonova
Appliances of Generative AI-Powered Language Tools in Academic Writing: A Scoping Review5-30

RESEARCH PAPERS

Muhammad Ahmad, Usman Sardar, Farid Humaira, Ameer Iqra , Muhammad Muzzamil, Ameer Hmaza, Grigori Sidorov, Ildar Batyrshin
Hope Speech Detection Using Social Media Discourse (Posi-Vox-2024): A Transfer Learning Approach31-43

Vladimir Bochkarev, Anna Shevlyakova, Andrey Achkeev
Synchronic and Diachronic Predictors of Socialness Ratings of Words.....44-55

Nikita Login
Wrong Answers Only: Distractor Generation for Russian Reading Comprehension Questions Using a Translated Dataset.....56-70

Dmitry Morozov, Timur Garipov, Olga Lyashevskaya, Svetlana Savchuk, Boris Iomdin, Anna Glazkova
Automatic Morpheme Segmentation for Russian: Can an Algorithm Replace Experts?.....71-84

Sergey Pletenev
Probing the Pitfalls: Understanding SVD’s Shortcomings in Language Model Compression..... 85-97

Valery Solovyev, Marina Solnyshkina, Andrey Ten, Nikolai Prokopyev
A BERT-Based Classification Model: The Case of Russian Fairy Tales 98-111

Vladimir Starchenko, Darya Kharlamova, Elizaveta Klykova, Anastasia Shavrina, Aleksey Starchenko, Olga Vinogradova, Olga Lyashevskaya
Fighting Evaluation Inflation: Concentrated Datasets for Grammatical Error Correction Task 112-129

Mikhail Tikhomirov, Daniil Chernyshov
Facilitating Large Language Model Russian Adaptation with Learned Embedding Propagation..... 130-145

Zhengye Xu, Yixun Li, Duo Liu
Predictions of Multilevel Linguistic Features to Readability of Hong Kong Primary School Textbooks: A Machine Learning Based Exploration 146-157

THANKING OUR REVIEWERS

Our Reviewers 159-161