Y. Ye and H. Zhou (Eds.)

doi:10.3233/FAIA250110

The Effects of Generative Artificial Intelligence Technologies on Writing Tasks in Foreign Language Learning

Anna P. Avramenko[a,1], Anna A. Nasonova[b], Alexey A. Tarasov[c] and Vladimir V. Ternovski[d]

[a] *Faculty of Foreign Languages and Area Studies, Lomonosov Moscow State University, Russia*

[b] *Faculty of Foreign Languages and Regional Studies, Lomonosov Moscow State University, Russia*

[c] *Higher School of Translation and Interpretation, Lomonosov Moscow State University, Russia*

[d] *Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Russia*

Abstract. The proliferation of generative artificial intelligence (AI) technologies has introduced challenges related to the use of automatically generated texts within foreign language learning settings. This study is aimed at developing clear-cut principles of incorporating large language models (LLM) into the English teachers' workflow. Our approach employed examining research papers and conducting a small-scale study designed to obtain and evaluate texts produced by state-of-the-art LLMs in response to the General English course writing assignments. The experiment revealed peculiarities and limitations of the chosen LLMs in generating texts for study purposes. The educational potential of this technology as well as suggested conditions for its effective integration into instructional practices were also presented. We concluded that the transformation of writing assignments is necessary, with a focus on fostering critical thinking skills. Furthermore, fostering students' information and communication technology (ICT) skills while engaging with an LLM chatbot is considered of paramount importance.

Keywords. Innovations in education, generative artificial intelligence, self-regulated learning, lifelong learning

1. Introduction

The UNESCO Institute for Lifelong Learning highlights the transition from the information society to the learning society. The main features of the latter are defined as the combination of self-education and productive activities. In a learning society, the role of the lifelong learner, or simply the student, is becoming increasingly important. Adaptive learning takes into account personal traits. In the digital environment, adaptability is understood as the ability of modern technology to tailor content to the user through machine learning, neural networks or AI technologies.

[1] Corresponding Author: Anna P. Avramenko, avram4ik@gmail.com.

Generative artificial intelligence (AI) technologies, namely large language models (LLMs), can act as a learning partner or digital tutor for foreign language learners. Compared to earlier model architectures, ChatGPT (Generative Pre-trained Transformer), launched by OpenAI on November 30, 2022, is more sophisticated in generating texts, especially for lengthy essays and 'creative' works, thanks to the integration of various natural language processing (NLP) methods within the "transformer" architecture (the first transformer language model was Google's Bert in 2019). However, despite the daily growth of technological progress in the field of creating large language models, there are limitations associated with this technology.

The UNESCO document "Artificial Intelligence Technologies in Education: Prospects and Consequences" aims at shaping a shared understanding of the opportunities and challenges that AI opens for education, as well as its implications for the essential competencies required in the age of AI [1]. It is largely the rapid development of AI technologies, including speech recognition (speech-to-text) and text-to-speech synthesis, natural language processing, and generative neural networks, that has provided insights for the challenges associated with creating favourable conditions for language practice. It is obvious that these technologies collectively can potentially compensate for the lack of adequate speech practice, for the vast number of people learning foreign languages in self-study contexts. At the same time, there is a discernible scepticism toward methodological tools based on the aforementioned technologies, which are only now fully entering the foreign language teaching. The earlier iterations with AI-tools were marred by failures and a mismatch between expectations and reality.

Although ChatGPT4-omni is perceived as the most sophisticated AI tool (Chiang et. Al [2]) for text generation in comparison to other LLMs, we propose a broader outlook on AI application to the language learning domain. We believe that comparative approach is to serve as a reliable foundation in terms of analyzing both LLMs' capabilities and limitations for language learners' advancement. However, we have to refrain from broad generalizations of recent AI studies in the language acquisition domain which focus on trends and research issues (Huang [3]), literature reviews (Law [4]), opportunities and challenges (Creely [5]). These and other papers cover the scope of AI application into language learning without taking into account real-life use of sophisticated tools within classroom settings or self-regulated learning mode. There are a few works that attempt to compensate for the lack of practical-oriented studies. These papers target a writing sub-domain within Computer assisted language learning (CALL) paying particular attention to AI evaluation tools including Grammarly, Pigai, Criterion (Ding [6]), ChatGPT3.5 implementation into classroom settings for writing training (Athanassopoulos [7]), scientific writing improvement through ChatGPT4, Claude, Bard (Lozic [8]). Although the latter study implies a comparative approach, most of the chosen tools appear to be outdated as of June 2024, according to the AI chatbot leaderboard presented at https://chat.lmsys.org/. Therefore, there is an evident gap between the publicly available LLMs and their research coverage within CALL, particularly in the area of teaching writing skills.

For research purposes, it is essential to investigate cutting-edge tools in order to see the technology's frontier. Thus, we have selected ChatGPT-omni (ChatGPT-o for short) as a benchmark, which could help in researching an array of other advanced AI tools. ChatGPT and other chosen tools for this study (elaborated in the succeeding section) are LLMs that may foster an individual study path that occurs through a set of user requests (prompts) rather than a predetermined scenario. In other words, LLM's users are engaged in a dialogue system, basically in the chatbot environment within a web or mobile app settings, which give immense opportunities for learning collaboration with AI through input-output natural language exchanges.

We have analyzed studies [Ahsan [9]; Barrot [10]; Cingillioglu [11]; Cotton [12]; Dawson [13]; Gimpel [14]; Hopp [15]; Imran [16]; Kohnke [17]; Ramalingam [18]; Stokel-Walker [19]; Yan [20]; Zawacki-Richter [21]] published during the first years of ChatGPT's commercial usage. In our view, understanding the technical features and functional limitations of dialogue systems is to identify necessary steps for adjusting these systems to address urgent methodological challenges, for example, enhancing the efficiency of chatbots as simulators of real-life foreign language communication as well as designing new learning opportunities, particularly in the area of foreign language writing.

Technically speaking, modern sophisticated dialogue systems can be divided into closed (task-based systems), aimed at solving a specific task, including educational ones, and open (open-ended), which are not embedded in a specific context. Although such a simplified classification does not fully reflect the entire spectrum of educational dialogue systems, it presents a huge leap of open systems in comparison to their predecessors:

- Closed systems (goal-oriented or task-based systems) - the user's freedom of expression is artificially limited with a restricted context for the sake of successfully performing a communicative task.

- Open systems (reactive or open-ended systems) - contextual initiative is given to the user, while the system's task is to select a semantically appropriate output to a user’s input.

The fundamental difference between these two systems is the amount of input and output data, which is seriously and intentionally limited in pragmatic systems. The presence of this limit affects the entire functionality of the system. Thus, a typical pragmatic system originally has a high structure, low flexibility, while having a clear objective. As for open systems, they are, in fact, the opposite to pragmatic systems concerning the above mentioned criteria: the possibility to input unstructured queries (low structure), high degree of flexibility (understanding a wide input context), implicit goal-setting (undefined objective), and an exploratory communication strategy (exploratory approach). All the above properties of these systems are shown in Figure 1.

Summarizing the above, it is obvious to assume that the capabilities of open systems, such as ChatGPT and similar ones, allow for working with a much larger amount of input data and obtaining much more creative output. Practically speaking, a student can easily formulate a quite advanced writing request, for instance a wording of a writing and receive a response that may be eligible for submitting, possibly with minor modifications, within an English language course. Obviously, this is not an outcome a teacher would likely envision within a writing course, even in a self-study mode. Bearing this in mind, it seems important to find out current limitations and possibilities of LLMs for dealing with the most complex ‘English as a foreign language’ (EFL) task that require the integration of a large number of linguistic and semantic patterns. One such task is undoubtedly to identify the capabilities and limitations of ChatGPT-like tools when dealing with different formats of writing tasks (essay writing, creative writing, etc).

The goal of our research is to develop a clear-cut set of principles easing a wise incorporation of LLMs into the teaching writing framework, taking into account the learning opportunities and threats posed by the use of GPT-like tools in foreign language learning, including written texts’ production. To achieve this goal, the following steps were undertaken:

- written products, generated by the chosen LLMs in response to prompts for nine types of writing tasks, were thoroughly studied;
- the main conditions for the successful integration of generative AI tools into teaching writing practice were identified;
- the current peculiarities and limitations of the chosen LLMs were discovered.

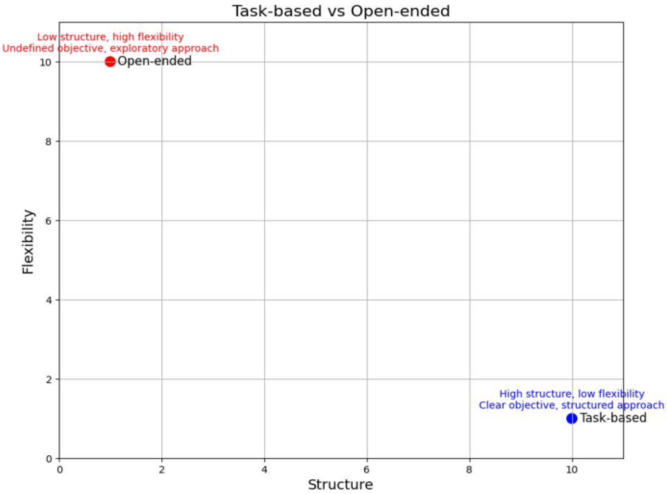


Figure 1. Basic properties of open and pragmatic dialogue systems.

2. Research methodology

The research bank for our study includes some of the most in-demand English language textbooks from the leading publishers: Oxford University Press and Pearson Education. The selection of writing exercises (assignments or tasks) was made from the following textbooks: English File, Headway, Solutions, and SpeakOut. Below one can see a short description for each textbook featuring their didactic peculiarities:

1. English File (by Oxford University Press): English File is a popular textbook series offering a comprehensive approach to learning English. Each level contains reading and listening lessons, grammar, vocabulary, writing, and speaking activities. Various texts and authentic materials are to aid students in developing various aspects of the English language.

2. Headway (by Oxford University Press): The Headway textbook series provides a systematic approach to learning English with a range of levels from the beginner to the advanced. Grammar, vocabulary, reading, listening, and writing are included in every lesson while gradually increasing in complexity. Authentic material is also utilized to assist students in developing practical skills and enhancing their English proficiency.

3. Solutions (by Oxford University Press): Solutions is a textbook series designed for English language learning tailored to students' needs. This BTM aids students in developing reading, listening, writing, and speaking skills. Solutions textbooks also contain engaging texts and dialogues as well as relevant exercises to reinforce the studied material.

4. SpeakOut by Longman Pearson: SpeakOut is an interactive textbook series prioritizing the development of students' spoken English skills. The BTM offers a range of engaging texts and audio materials for study, focusing on communicative skills and the functional use of language. SpeakOut also includes various tasks that help students confidently use English in daily life and the professional realm.

Undoubtedly, all writing assignment formats are important for developing students' EFL skills, but the value of these tasks may vary significantly. The key rationale for assignment selection in our research was to focus on developing practical skills and real-world communication within students' population. Such writing tasks like 'an email', 'a description', 'an article', and 'a proposal' can be used for communication at work or during studies. Therefore, all the chosen textbooks (B2 - upper-intermediate and C1 - Advanced levels) were thoroughly analysed for detecting such practical-oriented writing assignments. For instance, such tasks like anecdote writing and storytelling were dismissed due to lack of their daily-life application in real-world settings. As a result, out of at least 26 writing formats (some might be also attributed to the grammar and vocabulary clusters) a set of 14 tasks were chosen for final manipulation. One quite important consideration was to dwell solely on the writing assignments that do not require taking notes or any other specific preparation. In other words, we were focusing on the tasks descriptions that do not request any background information, which might be necessary for completing a task. Thus, some quite practical-oriented assignments were rejected as they contain these elaborations. For instance, a proposal format requires using a student's notes: 'Now write your proposal (220-260 words). Use your notes from exercise 5'.

The tasks' description has served as prompts which were fed to the chosen generative AI tools. No changes were made to the in-text descriptions apart from incorporating additional elements like a paragraph plan, the word limit, as well as specific recommendations for extending prompts (stated below with the prompts).

For our study, we focused on writing tasks fitting the above mentioned criteria and conventionally utilized within the 'General English' course tailored for students mastering B2-C1 CEFR level. We have selected the following task formats: e-mail, essay, article, cover letter, and description. Within the range there were a few variations showcasing slight modifications within the chosen formats and input (prompt) deviations. The e-mail type incorporates an e-mail (prompt - Write an email to the human resources department of a company, concerning a job you applied for) and an e-mail with a detailed instruction (the same prompt + stressing adding two apologizing parts alongside a thank-you part). The description type has also two variable structures. The first description promotes a part of town description (prompt - 'Write a description of your favourite part of town (about 250 words). Use the given plan to help you'). Eventually, the following writing tasks have been considered for generating AI texts:

- A formal e-mail
- A formal e-mail with detailed instructions
- A complaint letter
- An informal e-mail
- A cover letter
- An article describing a web-site
- A for-and-against essay
- A description of a town
- A description of a person or a place

The next stage was choosing the advanced AI tools that could perform similarly to ChatGPT4-o. To complete the following goal, the LMSYS Chatbot Leaderboard was thoroughly studied. The main rationale behind the choice was considering chatbots developed by different IT teams. Thus, Gemini-pro-1.5 (Google) was chosen alongside Claude 3 Opus (Antropic). Another consideration was implementing an open model, unlike the two chosen earlier, which are proprietary licensed by Google and Antropic respectively. Therefore, Llama3-70b-instruct, featuring the 14th LMSYS chatbot Leaderboard, was picked up. There was a final consideration aiming at investigating another newly introduced type of LLM, which is a locally-run AI tool. For this sake, a small Microsoft's Phi-3-medium-instruct was included into the study set. As a result, two proprietary and two open LLMs were analysed in comparison to ChatGPT4-o performance.

The workflow of the experiment presupposed the following stages. First, the writing exercises from textbooks were given to GPT4-o and its rivals to have generated the texts for fulfilling the exercises' goals (done through Arena side-by-side mode at https://chat.lmsys.org/). Second, the AI-generated texts were checked for plagiarism. Third, the texts were assessed following the suggested criteria. The present study incorporates some qualitative analysis of the generated texts, but its primary focus is obviously on the quantitative features defining the potential use of the GPT-like tools within learning writing settings. Here they are:

- Word count - number of words generated by AI tools
- Vocabulary range - percentage of words belonging to B2-C1 CEFR level (parsed through Oxford 5000 at https://www.oxfordlearnersdictionaries.com/text-checker/)
- Suggestions - LLM's comments for adding personal information to the generated template
- Additional structure elements - LLM's generated parts of the text including titles, section names, etc.
- Uniqueness - percentage of unique text which is not presented in the web (parsed through https://plagiarismchecker.ai/)

The suggested framework appears to be a reliable foundation for initiating a preliminary quantitative AI tools' analysis, which is presented in the following study. The given criteria stem from the teaching practice of the research's authors and truly represent practical considerations of teachers involved in English language writing classes.

3. Research results

The most notable challenge for foreign language teachers is in detecting plagiarism in creative writing assignments, which presupposes a transformation of both assessment criteria and procedures; and more broadly, it requires transforming the formats of written assignments themselves, to minimize harm from AI implementation.

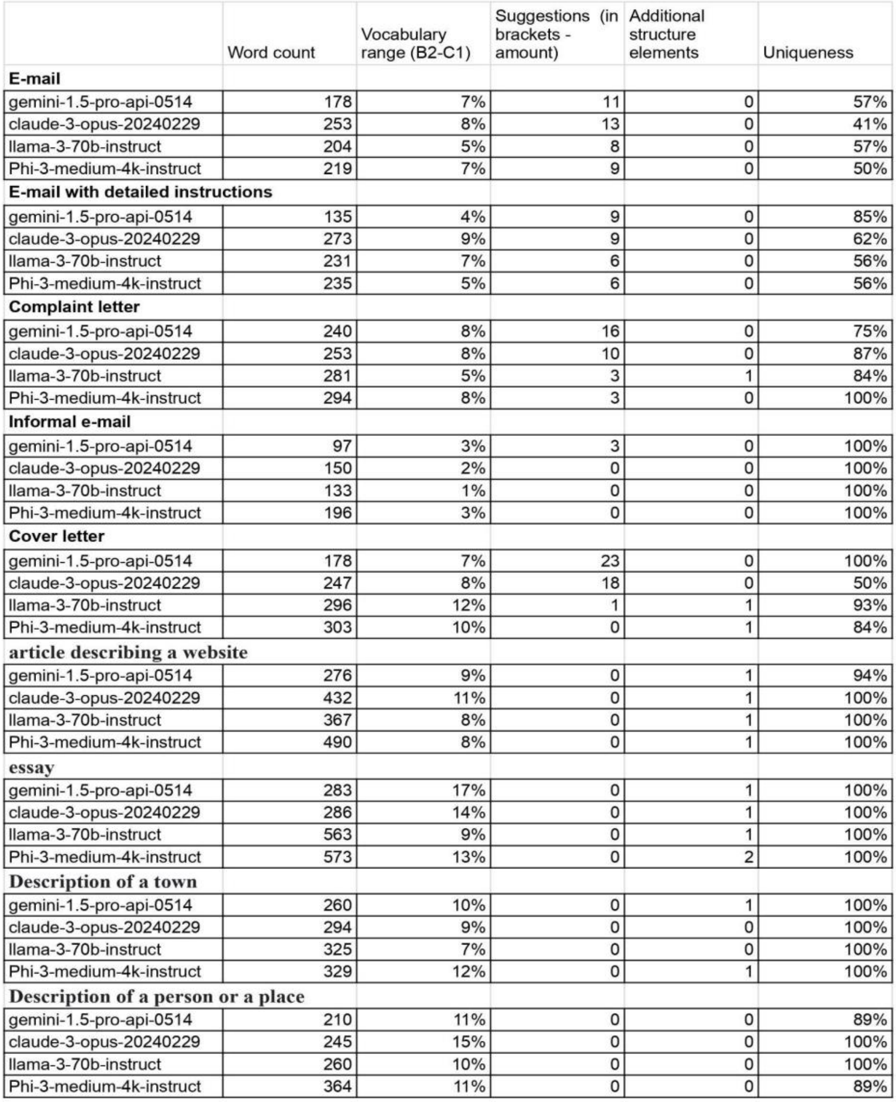
Turning back to the issue of plagiarism, it is notable that the evolution of generative AI technologies from GPT-3.5 to GPT-4 has practically rendered plagiarism checks, as understood today. Nevertheless, a year after the spread of generative AI technologies, professional communities in various fields agree that ChatGPT will never be considered the author of an artwork or scientific text. This is because large language models are trained on huge corpuses of texts, which then enables LLMs to generate text sequences literally replicating some existing chunks of speech. Since ChatGPT and other AI tools are by no means authors, they cannot be cited as a source (was initially behind the paragraph below).

All the generated texts from ChatGPT4-o and competing LLMs were parsed through an online AI detector at https://www.scribbr.com/ai-detector/. Not surprisingly, the AI detection has confirmed the well-known fact: AI texts are tracked as such in most occasions. According to the chosen AI detector, 32 out of the 36 generated texts (Gemini, Claude, Llama, Phi) were reasonably stated as 100% AI-generated texts. The rest has shown a slightly lower AI-generated probability hitting the following marks: 64% for Claude's informal e-mail, 46% for Llama's informal e-mail, 78% for Gemini's article describing a website, and 87% for Claude's article for the same task. The ChatGPT4-o performance in this respect was not anyhow different: out of 9 generated texts only one (an informal email) scored lower than 100% probability standing at 61% mark. Bearing in mind the experiment manipulation with the text data, it is reasonable to assume that the chosen AI detector works nearly perfectly in tracking the origin of an uploaded text.

The above given considerations arise an important question - how to treat plagiarism within AI-generated text? We believe that the 'Uniqueness' criterion from the present study might be instrumental in addressing the issue, at least within the scope of teaching writing. Given the fact that AI-tools generate text rather than extract them from the existing sources, it is still important to mention that in some writing formats, primarily with stricter language structure, some language patterns (word combinations, idioms, typical sentence patterns) do replicate the ones available on the web. Thus, non-detected patterns could be treated as unique ones and could potentially be AI-generated, especially when the overall grammar accuracy is exemplary. The given table below incorporates the 'Uniqueness' criterion as well as the other ones for the pack of 4 AI tools. By analysing the tools as compared to ChatGPT4-o, we will be able to generalize on the possible study effects of their implementation (Figure 2).

The first two writing types are intrinsically similar. Thus we may roll out any e-mail task by closely studying the criteria. The chosen benchmark, which is the ChatGPT4-o version, shows a performance similar to the Gemini generation. For the first e-mail (without detailed description), the GPT4-o word count is 172 words making it nearly identical to 178 words as that of Gemini. The vocabulary range (B2-C1) of GPT4-o boasts its high mark hitting a 9% level, which is still quite comparable with the Claude's and Gemini's figures. The GPT4-o number of suggestions does not significantly deviate from other LLM's performances as its mark of 8 levels the Llama's mark. Although the figures look similar, it is also notable that GPT4-o uniqueness rate, like the word count, show off its deviation: 60% originality, the top performance within the pack, with the least used words.

The complaint letter type represents an interesting case of partial deviation of the benchmark as compared to the studied LLMs. For instance, GPT4-o generated text hits the highest mark in terms of word count and vocabulary range, which are 336 words and 11% respectively. The latter parameter appears to be intrinsically connected to the former one as a higher volume may incorporate more advanced vocabulary items. However, the word count mark for GPT4-o naturally stands out as the closest rival's score is 294 words. Nonetheless, the other GPT4-0 parameters of a complaint letter fall into the mode scope within the pack of generated texts: 3 suggestions correlate with those of the latter two LLMs; 80% uniqueness put it above Gemini's and just behind Llama's and Claude's.



	Word count	Vocabulary range (B2-C1)	Suggestions (in brackets - amount)	Additional structure elements	Uniqueness
E-mail					
gemini-1.5-pro-api-0514	178	7%	11	0	57%
claude-3-opus-20240229	253	8%	13	0	41%
llama-3-70b-instruct	204	5%	8	0	57%
Phi-3-medium-4k-instruct	219	7%	9	0	50%
E-mail with detailed instructions					
gemini-1.5-pro-api-0514	135	4%	9	0	85%
claude-3-opus-20240229	273	9%	9	0	62%
llama-3-70b-instruct	231	7%	6	0	56%
Phi-3-medium-4k-instruct	235	5%	6	0	56%
Complaint letter					
gemini-1.5-pro-api-0514	240	8%	16	0	75%
claude-3-opus-20240229	253	8%	10	0	87%
llama-3-70b-instruct	281	5%	3	1	84%
Phi-3-medium-4k-instruct	294	8%	3	0	100%
Informal e-mail					
gemini-1.5-pro-api-0514	97	3%	3	0	100%
claude-3-opus-20240229	150	2%	0	0	100%
llama-3-70b-instruct	133	1%	0	0	100%
Phi-3-medium-4k-instruct	196	3%	0	0	100%
Cover letter					
gemini-1.5-pro-api-0514	178	7%	23	0	100%
claude-3-opus-20240229	247	8%	18	0	50%
llama-3-70b-instruct	296	12%	1	1	93%
Phi-3-medium-4k-instruct	303	10%	0	1	84%
article describing a website					
gemini-1.5-pro-api-0514	276	9%	0	1	94%
claude-3-opus-20240229	432	11%	0	1	100%
llama-3-70b-instruct	367	8%	0	1	100%
Phi-3-medium-4k-instruct	490	8%	0	1	100%
essay					
gemini-1.5-pro-api-0514	283	17%	0	1	100%
claude-3-opus-20240229	286	14%	0	1	100%
llama-3-70b-instruct	563	9%	0	1	100%
Phi-3-medium-4k-instruct	573	13%	0	2	100%
Description of a town					
gemini-1.5-pro-api-0514	260	10%	0	1	100%
claude-3-opus-20240229	294	9%	0	0	100%
llama-3-70b-instruct	325	7%	0	0	100%
Phi-3-medium-4k-instruct	329	12%	0	1	100%
Description of a person or a place					
gemini-1.5-pro-api-0514	210	11%	0	0	89%
claude-3-opus-20240229	245	15%	0	0	100%
llama-3-70b-instruct	260	10%	0	0	100%
Phi-3-medium-4k-instruct	364	11%	0	0	89%

Figure 2. The performance of Gemini, Claude, LLama, and Phi in the chosen tasks.

The next type of writing, namely informal e-mail, does not show unique performance of GPT4-o, yet there is one important observation that has to be perceived accordingly. Although the mode and median scores were detected among the pack of LLMs as compared to the benchmark GPT4-o (131 words for word count, 2% for vocabulary range, no suggestions and additional structures), the uniqueness rate might be mistakenly treated as a sort of deviation as it vividly exercises lower score (85%) in comparison to 100% consistent performance of the other LLMs. However, a more thorough analysis of the generated text using the suggested plagiarism tool shows an interesting coincidence: one name and one phrase of the generated text were traced in the open online sources, thus lowering the overall score. Virtually, we might assume that the uniqueness rate for GPT4-o text is also 100% as a frequently used phrase pattern and a typical name happened to be generated by GPT4-o. All in all, the GPT4-o text represents a unique informal letter, which is similar to the other generated texts. Below is the GPT4-o informal letter with marked parts as the plagiarized ones.

To: chris@email.com
Subject: Starting Work

Hi Chris,

Thanks for your email—it was great to hear from you! Congrats on passing all your exams! I know exactly how you feel about starting work; I've been having similar thoughts myself. Waking up in the middle of the night, wondering how the first day will go, is all too familiar.

Preparations are a lot, right? I've spent a small fortune on new clothes to look the part—new suit, shirts, tie, and some smart shoes. I'm also trying to think of good questions to ask my new colleagues to make a positive first impression and avoid any awkward moments.

Well, that's all for now. Enjoy your last week of freedom and try not to stress too much. I'm sure everything will go smoothly.

Best wishes,
Mike

The article type of writing generated by GPT4-o has similar features as compared to the other articles by the suggested LLMs. Alongside most of the pack GPT's text has no suggestions in brackets, only one additional element (in this case it is a title for the article), and 100% uniqueness rate. However, there is an obvious difference in terms of the size of the GPT's text (245 words), making it the shortest generation of all. The advanced vocabulary range parameter for GPT4-o is also the smallest of all. But in such a case one has to take into account not just plain figures, but the ratio between the size of the text and B2-C1 vocabulary distribution. There is no direct correlation, yet the trend within the chosen LLMs is visible: the longer the text, the likelihood of having more advanced vocabulary is higher. Thus, it is reasonable to assume that GPT4-o text could have been on par with the rest of suggested LLMs in case it had been around 300 word size.

If we pass on to the cover letter analysis, one may recognize a noticeable performance of GPT4-o. For instance, the word count (269 words) nearly represents the mean for the rest of LLM's pack which vary significantly. Similarly, GPT4-o shows a solid in-the-middle-performance in terms of suggestions, 15 ones versus 23, 18, 1, and 0. Therefore, GPT4-o's generated text, like the Gemini's and Claude's, does represent a kind of sample text that has to be filled with the relevant information. The least originality rate (47%) may suggest the GPT4-o replication of the standardised format of a cover letter, which under no means is not a failure and may be treated as a benefit of the generated text. The 10% vocabulary mark also suggests the GPT4-o 'professional' expertise as this figure is at the top two of all the analysed texts.

A closer look at the essay distribution criteria proves a variable performance of the chosen AI tools which are drawn into a detailed cluster (Gemini and Claude) and a short essay pattern (Llama and Phi). The GPT4-o as benchmark aligns with the detailed cluster as it shows off a 523 word count figure. Its 10% rate of Vocabulary range does not put it atop of the rest of the pack as it is the second worst score. Nonetheless, this may hardly be considered a drawback as a higher essay vocabulary range does not generally contribute to its overall linguistic quality. Moreover, 4% of the analysed vocabulary were tracked as ‘unclassified’ meaning that the words are not included within the Oxford 5000 list. Even a quick analysis of these vocabulary outliers, containing such words as *procrastination, disparities, conducive*, attribute the words to more narrowly used word lists, which adds a ‘proficiency’ touch to the generated text. Llama and Phi have identical 4% unclassified words. As for the uniqueness rate, all the LLMs from the pack constitute the very same top level of 100%, which is quite predictable due to the creative nature of essay writing.

The last two types of writing in the list appear to be quite specific due to the wording of the task in question, which served as a prompt for text generation. Unlike the previous writing tasks, the suggested descriptions include a word limit. Thus, for instance, the description of the place, person or an animal is to fall into the 220-280 limit: “*Write a description of a person, place or animal using the theme ‘Looks can be deceiving’. Your audience is the readership of a university creative writing magazine (220-280 words)”*. As one may notice in the table presented earlier, 2 LLMs (Gemini, Phi) do not align their texts to the suggested limit. So does GPT4-o containing 309 vocabulary items. To the naked eye, this might be regarded as a relatively big violation of the alignment to the human actor. Also, in a way it could ‘cast shadow’ on GPT4-o benchmark status. However, if we take into the consideration other parameters, then it is easily noticed that GPT4-o performs quite consistently showing 100% uniqueness rate alongside the leaders. Also, 11% of advanced vocabulary proves that GPT4-o has a tendency to perform ‘reasonably’ without complicating the vocabulary range. Therefore, we may assume that GPT4-o shows a predictable and consistent performance for the user as, in such a case and the similar text generation tasks, it has less outliers as compared to other LLMs.

It is commonly pointed out that ChatGPT, like other LLMs, often generates non-existing facts, including quotes. Some of the generated texts within the present study also have incorporated a creative touch of 3 LLMs (Claude, LLama, Phi) stating the non-existing names in cover letters. Yet on balance we have to assume that 3 LLMs, like the GPT4-o benchmark, have generated articles containing existing websites. Both scenarios contribute to the fact that students are always in the position of fact-checkers when it comes to dealing with texts presupposing including some facts.

All the generated texts demonstrate a high level of pragmatic and language proficiency as no evident grammar and vocabulary errors have been detected. The style of writing was successfully matched to the given formats, whether formal, commanding, friendly, or polite. Structural integrity was meticulously maintained. For instance, an email with a detailed instruction, created with the specific intent to request rescheduling of a job interview, follows the structure of requests for this format. The letter begins with a polite expression of gratitude, then passes on to explaining the grounds for the request, suggests alternative dates, and concludes with an optimistic expression of hope for understanding.

Likewise, a formal letter of complaint addressed to customer support in response to a received faulty item maintains an assertive tone used to express dissatisfaction and seeks a suitable resolution for the problem. The structure fully complies to the typical framework of formal complaints. The email starts with a formal greeting, followed by comprehensive order details suggestions (have to be filled by the user). The only exception is a Llama's generated text which does not constitute a sample-like letter of complaint for further writing development. Instead, it is a complaint letter which could be perceived as a real one. Yet still some missing parts, for instance no order number given, may catch a student off guard with its format (XXX standing for order numbers).

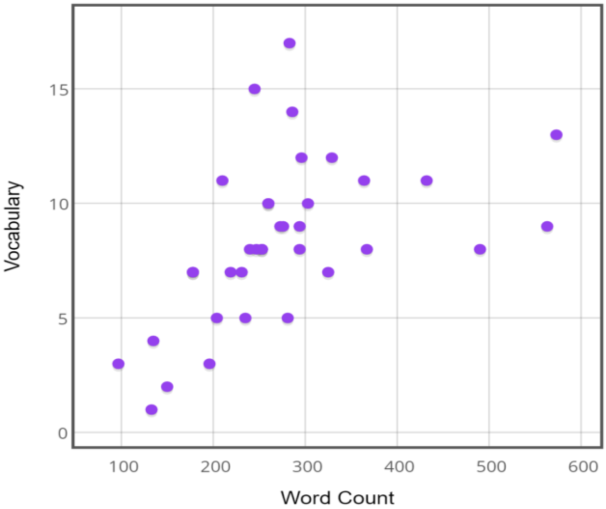


Figure 3. The vocabulary range and word count correlation for the pack of 4 LLMs.

The above pictured scenario exemplifies linguistic outliers of LLMs which may occur for some of them even when there is no hint for such occurrences while feeding other AI tools with identical inputs. Nevertheless, it is possible to highlight the statistically significant similarities for the whole pack of 4 analysed LLMs. Thus, for instance the scatter plot above (Figure 3) shows a medium positive word count-vocabulary range (B2-C1) correlation. With a greater size of the generated text it is reasonable to expect a higher rate of advanced vocabulary presence. Both Pearson (0.5) and Spearman correlation coefficients (0.64) suggest a moderate correlation between the suggested criteria.

Yet the presented in-criteria dependency cannot be aligned with other positive correlations. Thus, we may conclude that the word count rate does not significantly affect the uniqueness rate for the bank of 36 generated texts. In the scatter plot given below it is possible to trace only a weak positive correlation between the above mentioned criteria. Pearson (0.27) and Spearman correlation coefficients (0.3) confirm the assumption statistically (Figure 4).

The depicted weak correlation case appears to be a vivid illustration for the variability of the 36 generated texts with 4 distinctive LLMs. Literally speaking, a generated text of either a tiny size (about 100 words) may have the same 100% uniqueness rate as that of a huge generation (about 600 words). This immense dispersion of the size of the generated texts might be misleading unless one takes the prompts' guidelines into account. Informal e-mail (140-190 words), cover letter (250 words), description of a town (250 words), description of a place (220-280 words) do have either a target word count or a word limit. However, browsing through the table with the LLMs' data one may easily spot evident variability. While some LLMs in specific tasks perform within or close to the suggested limit, others violate the target substantially. Thus, the mean figures for the LLMs in question may be quite intact with the suggested limit (144 for Informal letter, 256 for cover letter, 302 for description of a town, 269 for description of a place), although there are obvious outliers representing considerable deviations from the 'compliant' LLMs within particular text types. All in all, it can be assumed that the degree of variability is significant within the analyzed pack of LLMs.

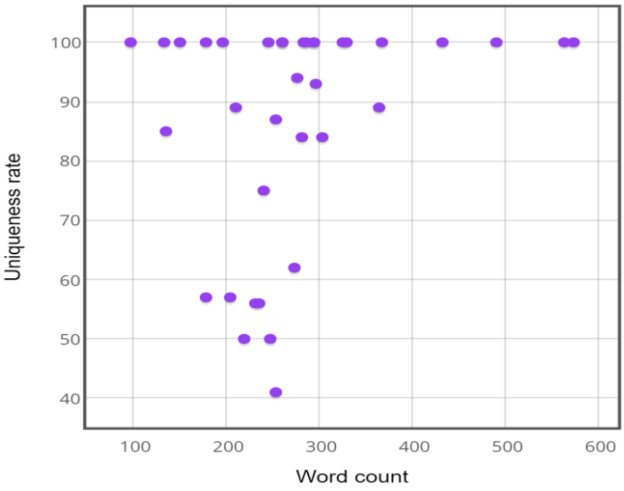


Figure 4. The uniqueness rate and word count correlation for the pack of 4 LLMs.

Before turning to certain didactic principles of the studied LLMs, it is important to highlight its technical features:

- All the studied LLMs including ChatGPT4-o might be available for free using the source from this article, but there might be quite discouraging generation limits, especially for ChatGPT4-o and Claude.
- All the generated texts fulfil the tasks' goal showing considerable degree of variability within 2 suggested criteria: word count and suggestions.
- ChatGPT4-o has demonstrated similar performance as compared to 4 other LLMs within the range of writing tasks. Its top position in the mentioned rating could be interpreted through the main peculiarity of its performance: ChatGPT4-o normally does not generate a considerable number of outliers showcasing consistent execution of writing tasks and its predictable high quality.

Bearing in mind the aforementioned, let us summarize the didactic principles of incorporating LLMs for teaching foreign language writing. They include the following:

Adaptability. Students may be advised to use a few LLMs at a time to generate texts for further analysis, which may lead to editing the chosen text or compiling a new version integrating initially generated texts. Also, comparing their own editions (self-written or written in assistance with LLMs) with the generated texts can help identify areas where they may need additional help or guidance.

- Interactivity. Real-time feedback from LLMs might include not only requested texts, but also some LLMs excuses (about some missing parts that were expected to be generated) as well as recommendations (suggestions) for adding some meaningful information to the generated texts. Such occasions trigger follow-up discussion of a student with a LLM. It might be encouraged by a teacher narrowing down (requesting specific words, word-combinations, idioms) or widening (requesting guidelines and tips on writing) the focus of discussion.

- Accessibility. LLMs including ChatGPT4-o can be accessed through various platforms, including websites, smartphone applications, or messengers with integrated chatbots. This contributes to various possibilities within self-regulated learning mode, especially in preparation for writing parts of standardized exams. Teachers may outline a basic preparation plan with suggested modes of learning giving way to LLMs to feed students with relevant and tailored content.
- Flexibility. Varying prompts prepared with the help of teachers may allow students to learn about capabilities and limitations of LLMs in connection to their own foreign language proficiency level. This practice may encourage students to initiate self-assessment of their own prompts and other self-written texts.
- Openness. Hallucinations and real-life facts, generated during interaction with LLMs, can be exploited by the teacher willing to exercise exploratory approach for fact-checking and elaborating on the generated parts of texts. This practice should be coordinated by the teacher and may contribute to acquiring critical thinking skills as well as would-be linguistic benefits in case of encountering unknown speech patterns or writing styles.

The above-mentioned principles, if applied to classroom settings, can definitely boost students' writing proficiency in foreign languages, although a broader scope is opened up by the introduction of LLMs into learning practices. LLMs in their current iterations can serve as intelligent tutoring systems enabling students to be receiving tailor-made (personalized) feedback in the form of sample texts, explanations to language phenomena as well as corrective feedback to students' input (a written text). In case of preparation for a high-stakes language proficiency exam, for which a student might employ a self-regulated mode (Tarasov, 2023), this cooperation with a LLM appears to be of great value. However, other possible self-regulated practices appear to be less beneficial for students as their intrinsic motivation can be deeply affected by ambiguity of LLM capabilities, especially in writing studies for which students need a lot of exposure to text samples.

4. Discussions

Analyzing ways to minimize the threats of LLMs and harness its didactic potential, we concur with the authors of the aforementioned studies, highlighting the following conditions for implementing this technology in teaching:

1. Developing ICT competence within students' population as a component of their cognitive activity appears to be the only correct path as opposed to attempts to ban the technology.

2. In creative writing assignments, a phased execution of tasks can allow teachers to reduce the likelihood of students submitting an automatically generated text.

3. In the assessment criteria, on the one hand, the evaluation of the work process should be prioritized; on the other hand, particular attention should be paid to assessing critical thinking skills.

Although the main potential threat (automatic generation of well-structured text on any topic) of using GPT-like tools appear to be obvious, some other LLMs' functions present challenges for foreign language teachers:

- Idea Generation. LLMs can assist students in generating new ideas for written assignments by suggesting themes, plots, and perspectives.
- Machine Translation. ChatGPT can translate text from one language to another, while also performing additional text analysis and transformations (e.g., sentence and paragraph structuring).
- Text Editing. Similar to other language correction tools like Grammarly, Paperpal, and QuillBot, LLMs can identify grammatical and syntactical errors, improving the structure and coherence of the text.

Eventually, students' independent work becomes more akin to individual sessions with a tutor, in our case a digital AI-tutor. LLMs can also be utilized by instructors to create personalized content tailored to their students' individual needs and abilities. Materials and formats of work that align with students' own interests and goals not only transform homework assignments but also may better engage learners into subsequent classroom activities. This cooperative approach to using LLMs has to be thoroughly investigated in further studies as well as implementation of LLMs within the scope of scientific and ESP texts.

5. Conclusions

The presented research work has revealed that the texts generated by state-of-the-art LLMs (the chosen iterations of Gemini, Claude, LLama, and Phi) in response to writing tasks prompts are characterized by clear structure and exemplary language proficiency as well as a considerably high degree of variability, but sometimes these LLMs' written products contain unexpected entities including non-existing facts that are conventionally called 'hallucinations' in AI terms. It should be noted that at times LLMs outputs may not just contain requested texts in response to task-oriented prompts, but also additional LLM's comments drawing students' attention to the specifics of the generated texts (including textual suggestions and meta information for the text). As a result, students are charged with huge loads of information that has to be grasped properly as there might be unexpected data. Therefore, it can be assumed that at the current stage of generative AI technology development, foreign language educators are required to transform writing tasks with a focus on developing and assessing critical thinking skills.

Acknowledgement

This work was done with the support of MSU Program of Development, Project No 23-SCH02-26.

References

[1] Miao F., Holmes W., Ronghuai Huang, Hui Zhang. AI and education: guidance for policy-makers. United Nations Educational, Scientific and Cultural Organisation, 2021. – 45 p.

[2] Chiang, W. L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., ... & Stoica, I. Chatbot arena: An open platform for evaluating llms by human preference. 2024, arXiv preprint arXiv:2403.04132.

[3] Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. Trends, research issues and applications of artificial intelligence in language education. Educational Technology & Society, 2023, 26(1), 112-131.

[4] Law, L. Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. Computers and Education Open, 2024, 100174.

[5] Creely, E. Exploring the Role of Generative AI in Enhancing Language Learning: Opportunities and Challenges. International Journal of Changes in Education. 2024.

[6] Ding, L., & Zou, D. Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. Education and Information Technologies, 2024, 1-53.

[7] Athanassopoulos, S., Manoli, P., Gouvi, M., Lavidas, K., & Komis, V. The use of ChatGPT as a learning tool to improve foreign language writing in a multilingual and multicultural classroom. Advances in Mobile Learning Educational Research, 2023, 3(2), 818-824.

[8] Lozić, E., & Štular, B. ChatGPT v Bard v Bing v Claude 2 v Aria v human-expert. How good are AI chatbots at scientific writing, 2023.

[9] Ahsan, K., Akbar, S., & Kam, B. Contract cheating in higher education: A systematic literature review and future research agenda. Assessment & Evaluation in Higher Education, 2022, 47(4), 523–539, https://doi.org/10.1080/02602938.2021.1931660.

[10] Barrot, J. S. Using ChatGPT for second language writing: Pitfalls and potentials. Assessing Writing, 2023, 57, 100745, https://doi.org/10.1016/j.asw.2023.100745.

[11] Cingillioglu, I. Detecting AI-generated essays: The ChatGPT challenge. The International Journal of Information and Learning Technology, 2023, 40(3), 259-268, https://doi.org/10.1108/IJILT-03-2023-0043.

[12] Cotton, D. R., Cotton, P. A., & Shipway, J. R. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International. 2023, https://doi.org/10.1080/14703297.2023.2190148.

[13] Dawson, P., & Sutherland-Smith, W. Can markers detect contract cheating? Results from a pilot study. Assessment & Evaluation in Higher Education, 2018, 43(2), 286–293, https://doi.org/10.1080/02602938.2017.1336746.

[14] Gimpel H. et al. Unlocking the power of generative AI models and systems such as GPT-4 and ChatGPT for higher education: a guide for students and lecturers. Universität Hohenheim, 2023, – 44 p.

[15] Hopp, C., & Speil, A. How prevalent is plagiarism among college students? Anonymity preserving evidence from Austrian undergraduates. Accountability in Research, 2021, 28(3), 133–148, https://doi.org/10.1080/08989621.2020.1804880.

[16] Imran, M., & Almusharraf, N. Review of teaching innovation in university education: Case studies and main practices. The Social Science Journal. 2023, https://doi.org/10.1080/03623319.2023.2201973.

[17] Kohnke, L., Moorhouse, B. L., & Zou, D. ChatGPT for language teaching and learning. RELC Journal, 00336882231162868. 2023, https://doi.org/10.1177/00336882231162868.

[18] Ramalingam, S., Yunus, M. M., & Hashim, H. Blended learning strategies for sustainable English as a second language education: A systematic review. Sustainability, 2022, 14(13), 8051, https://doi.org/10.3390/su14138051.

[19] Stokel-Walker C. AI bot ChatGPT writes smart essays–should professors worry? Nature. 2022, https://doi.org/10.1038/d41586-022-04397-7.

[20] Yan, D. Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. Education and Information Technologies. 2023, https://doi.org/10.1007/s10639-023-11742-4.

[21] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. Systematic review of research on artificial intelligence applications in higher education – Where are the educators? International Journal of Educational Technology in Higher Education, 2019, 16(1), 1–27, https://doi.org/10.1186/s41239-019-0171-0.