Danijela Ljubojević
Institute for Educational Research
Serbia
danijela.ljubojevic@ipi.ac.rs

UDK [811.111'243:37.091.3]:[808.1:004.89]
DOI: 10.4312/vestnik.17.301-322
Izvirni znanstveni članek

EVALUATING CHATGPT IN RELATION TO ITS EFFECTIVENESS FOR ARGUMENTATIVE ESSAY WRITING

ABSTRACT

The rapid advances in artificial intelligence, particularly in language models like ChatGPT, have led educators to re-evaluate traditional teaching methods, especially regarding essay writing assignments. Teachers express growing concern that AI-generated essays could compromise the essence of student writing, potentially diminishing critical skills such as deeper analysis and personal expression. This case study explores an innovative approach to integrating ChatGPT within the classroom by making use of the flipped learning model and sentence skeletons technique. For the purpose of this research four argumentative essays were generated by ChatGPT. These essays were assessed by four independent teachers using a structured checklist. The checklist focused on essay structure, cohesion and coherence, argument presentation, vocabulary use, and grammatical accuracy, and the intellectual standards necessary for critical thinking, such as clarity, precision, accuracy, depth, breadth, logic, significance, relevance, and fairness. The results revealed that ChatGPT excels in essay writing in terms of essay structure, developing an appropriate introduction, body, and conclusion. It generates high-quality argumentative essays, rich in vocabulary (C1–C2 CEFR level), and text linking (numerous transitions signals and linkers). On the other hand, these essays did not follow word limit requirements, write a funnel introduction/paragraph, provide opposing arguments, and give personal opinions. When it comes to the element of critical thinking, the generated essays lacked depth, breadth, and fairness. Nevertheless, the findings indicate that ChatGPT holds potential as a valuable tool for improving student engagement in writing and fostering critical thinking skills. The case study suggests that ChatGPT can be more than just a text generator; it can serve as an interactive resource for students to analyse and learn from model arguments (sentence skeletons), increasing their understanding of cohesive essay structure and logical reasoning. Overall, these insights propose a dual role for ChatGPT in educational settings – not merely as a source of content, but as a catalyst for developing essential academic skills.

Keywords: ELT, ChatGPT, essay writing, argumentative essays, critical thinking

IZVLEČEK

OCENJEVANJE UČINKOVITOSTI CHATGPT-JA PRI PISANJU ARGUMENTATIVNIH ESEJEV

Hiter napredek na področju umetne inteligence, zlasti jezikovnih modelov, kot je ChatGPT, je učitelje spodbudil k ponovnemu razmisleku o tradicionalnih učnih metodah, še posebej pri nalogah pisanja esejev v angleščini. Učitelji izražajo vse večjo zaskrbljenost, da bi lahko eseji, ki jih ustvari umetna inteligenca, ogrozili bistvo avtonomnega pisanja učencev. To bi lahko zavrlo razvoj ključnih veščin, kot sta poglobljena analiza in osebno izražanje. V prispevku proučujemo inovativen pristop k uvajanju rabe ChatGPT-ja v razredu s pomočjo modela obrnjenega učenja in tehnike stavčnega ogrodja. Za namen te raziskave so štiri argumentativne eseje, ki jih je ustvaril ChatGPT, ocenili štirje neodvisni učitelji s pomočjo strukturiranega kontrolnega seznama. Slednji se je osredotočal na zgradbo eseja, kohezijo in koherenco, predstavitev argumentov, rabo besedišča in slovnično natančnost ter na intelektualne standarde, potrebne za kritično mišljenje, kot so jasnost, točnost, natančnost, globina, širina, logičnost, pomen, relevantnost in pravičnost. Rezultati so pokazali, da so eseji, ki jih ustvarja ChatGPT, odlični z vidika strukture ter primernosti uvoda, jedra in zaključka. Gre za visokokakovostne argumentativne eseje z bogatim besediščem (na ravni C1–C2) in številnimi povezovalnimi besedami (označevalci prehodov in vezniki). Po drugi strani v teh esejih ni upoštevana omejitev števila besed, v uvodu in posameznih odstavkih ni prehoda od splošnega h konkretnemu, poleg tega pa niso predstavljeni niti nasprotujoči si argumenti ali osebna mnenja. Z vidika kritičnega mišljenja tovrstni eseji kažejo na pomanjkanje globine, širine in pravičnosti. Kljub temu ugotavljamo, da ChatGPT lahko postane dragoceno orodje za izboljšanje zavzetosti učencev za pisanje in spodbujanje kritičnega mišljenja. Naša študija primera kaže, da je ChatGPT lahko več kot generator besedil: lahko je interaktivni vir, s pomočjo katerega učenci analizirajo zglede argumentov (stavčnih ogrodij) in se iz njih učijo, kar lahko izboljša njihovo razumevanje kohezivne strukture esejev in logično mišljenje. Na splošno izsledki raziskave kažejo na dvojno vlogo ChatGPT-ja v izobraževanju – je vir vsebine, pa tudi orodje za razvoj ključnih akademskih veščin.

Ključne besede: poučevanje angleščine, ChatGPT, pisanje esejev, argumentativni eseji, kritično mišljenje

1 INTRODUCTION

The most significant recent advances in the field of artificial intelligence involve the use of chatbots powered by large language models (LLMs) for educational purposes. In short, a chatbot is a computer programme that uses natural language to interact with users (Mavropoulou & Arvanitis, 2023). Conversational AI chatbots, such as ChatGPT, are examples of dialogue systems in which users can talk to an automated interlocutor, either orally or in writing. Bibauw et al. (2019) define dialogue-based computer assisted language learning (CALL) as "any system allowing a user to have a dialogic interaction with an automated agent as a language learning task". In light of these definitions, recent research has explored the pedagogical potential of AI-driven tools in the context of English Language Teaching (ELT). Huang et al. (2022) conducted a systematic review of empirical studies and identified five key pedagogical ways of using chatbots in language learning: as an interlocutor (a learning companion to assist student's language learning), as a simulation of an authentic learning environment, for the transmission of information, as a helpline, and as a source of recommendation for further learning materials. Furthermore, Chang et al. (2023) argue that the emergence of ChatGPT and generative AI technologies challenges educators to rethink traditional pedagogy by integrating AI into teaching through three key principles – prompting, self-assessment, and personalization in order to support students' self-regulatory processes and self-evaluation in the learning process. On the other hand, research has also shown that this tool cannot fully handle the complexity of articulating the educational process. While AI chatbots can support practice in grammar and vocabulary, they fall short in managing the complex, contextualized communicative competences and socio-cultural nuances essential in ELT. Godwin-Jones (2024) argues that AI systems lack the lived experience to be able to use language with the same social awareness as humans, leading to producing language that is pragmatically inappropriate. Moreover, they lack linguistic and cultural authenticity. Similarly, Retelj (2023) found out that when generating lesson plans, language models (such as ChatGPT) lack appropriate didactic content, fail to consider students' needs or cultural differences, and do not adhere to modern didactic-methodological principles of foreign language teaching. Additionally, there has been a growing concern for the misuse of this technology, especially when it comes to academic integrity and plagiarism. Language teachers are particularly disturbed by the numerous homework assignments being generated by chatbots and not done by their students. Of particular concern is the use of ChatGPT, which originally built on the GPT-3.5 architecture but now operates on more advanced GPT-4–based models. Many students enjoy the benefits of generating text for their homework assignments, but this behaviour can profoundly impact the writing process and development of critical thinking (CT) skills. On the one hand, some research has shown that ChatGPT does not enhance students' essay-writing performance: students who use

ChatGPT for essay writing did not deliver higher quality content (when the following elements were assessed: mechanics, style, content, and format), did not write faster, nor was there a greater degree of authentic text (Bašić, Banovac, Kružić, & Jerković, 2023). In contrast, Khampusaen (2024) reported significant improvements in students' argumentative writing, including better organization, coherence, and language use, although concerns about over-reliance and academic integrity remained. Furthermore, questions have been raised not only about academic integrity and plagiarism, but also about the lack of development of some essential skills for the 21st century, such as critical thinking. The rationale behind this research was the study by Nakrowi et al. (2023), who stated that

> future research must think of learning models that can build awareness of the importance of reading for students. By reading, students can find the concept of an excellent argumentative text. With adequate knowledge related to argumentative texts, students will find it easier to find ways to write good argumentative texts.

In the similar vein, Khampusaen (2024) emphasized that while ChatGPT can scaffold students' essay writing by improving content development and organization, it does not automatically foster critical thinking, which must be intentionally cultivated in classroom practice. With this paper, the authors would like to propose another perspective of how AI-based tools can be integrated into lesson plans to enhance, rather than hinder, students' critical thinking and academic writing skills. We explore the ways to find these good learning models and presume that ChatGPT has the potential to be used for developing essay writing skills with special emphasis on critical thinking skills. Our research set out to analyse argumentative essays generated through ChatGPT[1] and determine what this system can generate in terms of writing. The case study sought to answer the following specific research questions:

RQ1 To what extent can ChatGPT effectively generate argumentative essays based on the input in terms of essay organization?

RQ2 What are the linguistic devices that are characteristic of ChatGPT-generated content?

RQ3 To what extent does Chat GPT ChatGPT incorporate elements of critical thinking?

The findings of this research have implications for teaching writing and the integration of AI in English language classrooms.

[1] The study was conducted using ChatGPT based on OpenAI's GPT-4-turbo model, available through the ChatGPT Plus subscription plan at the time of writing (October 2023).

2 THEORETICAL FRAMEWORK

Writing argumentative essays is very important for developing critical thinking skills, and providing sufficient and sound arguments is essential to their success. It is not enough just to support an idea with enough details and examples, as there are other aspects of such writing that should also be included.

There are several models of argumentation that can be used in essay writing, such as the Classical Model of Argumentation (Aristotle, 2024), the Toulmin Model (Toulmin, 2003), and the Rogerian method of argumentation (Young, Becker, & Pike, 1970). The choice of model depends on the writer's goals, the task given, and the nature of the argument. Each model offers a unique approach to constructing and presenting persuasive arguments. Because of its simplicity and the objective of this study to tackle the notion of critical thinking skills – the need to address the opposing arguments as part of analysing and evaluating – the Classical Model of Argumentation has been chosen for this case study.

The Classical Model of Argumentation was first developed by the Greek philosopher Aristotle in his work *Rhetoric* (Aristotle, 2024, pp. 238-260). The classical argument consists of five main components, arranged in this order:

(1) Introduction: Introduces the topic, present the thesis and the main point.

(2) Narration: Presents relevant background information or context related to the issue.

(3) Confirmation: Presents the main arguments, evidence, and reasoning supporting the claim.

(4) Refutation: Addresses and counters opposing arguments or potential objections; explains the other side.

(5) Conclusion: Summarizes the key points, restates the thesis, and gives a personal opinion on the topic.

According to academic writing conventions in ELT, a well-developed argumentative essay should begin with the introductory paragraph containing a thesis statement and a controlling idea, followed by paragraphs with arguments and counterarguments, and ending with a conclusion that usually conveys the student's strong final thoughts on the topic (Oshima & Hogue, 2006). This structure not only reflects the principles of the Classical model but also aligns with contemporary academic writing practices, helping learners organize their ideas logically and persuasively. In addition, argumentation, as a core component of argumentative writing, supports the development of higher-order thinking skills and promotes the development of critical thinking by requiring students to evaluate evidence, consider multiple perspectives, and construct coherent reasoning.

Critical thinking is defined as "the intellectually disciplined process of actively and skilfully conceptualizing, applying, analysing, synthesizing, and/or evaluating information gathered from, or generated by, observation, experience, reflection, reasoning, or communication, as a guide to belief and action" (National Council for Excellence in

Critical Thinking, 1987). It is the process of arriving at a judgement about the value or impact of a text by examining its quality. Some fundamental facets of critical thinking are analysing arguments, claims or evidence; judging or evaluating based on evidence; making inferences using inductive or deductive reasoning; and making decisions and/or solving problems through reasoning.

To assess the quality of argumentative essays and critical thinking skills, i.e. the quality of reasoning, intellectual standards are used (Nakrowi et al, 2023). Good critical thinking requires having a command of these standards. According to Paul and Elder (2013, 2016), the ultimate goal is for the standards of reasoning to become infused in all thinking so as to become the guide to better reasoning. The intellectual standards proposed by the Paul-Elder model are clarity, accuracy, precision, relevance, depth, breadth, logic, significance, and fairness. **Clarity** is the gateway standard of critical thinking, reflected in essay writing through the use of clear and comprehensible language, clear explanations of key terms and concepts, unified paragraphs that communicate one clear idea, the use of examples or elaboration to clarify complex points, and statements that are free from vague or confusing wording. **Accuracy** refers to the quality of being free from errors or distortions and representing things as they truly are, e.g. the facts presented in the essay are correct, verifiable and supported by valid data rather than misconceptions or inaccuracies. **Precision** represents the degree of specificity and detail in an essay, supported with sufficient examples and details, to fully convey meaning and avoid ambiguity. It can be evaluated through the clarity of definitions, the consistent use of accurate terminology, and the careful choice of statements to avoid vagueness or overgeneralization. **Relevance** indicates how closely a statement or idea relates to the issue or question at hand, emphasizing the importance of staying focused on the topic and ensuring all ideas support the central argument. **Depth** refers to the extent to which the thinking addresses the underlying complexities and interrelated factors of a topic, moving beyond a superficial understanding. The indicators of depth in essay writing include not only the presence of ideas developed through sufficient analysis, revealing layers of meaning, causes, and implications, but also the exploration of various aspects of the issue, particularly how they are connected or influence one another. **Breadth** in an essay is characterized by multiple relevant perspectives and points of view, demonstrating a comprehensive and open-minded understanding of an issue. It requires the student to reason insightfully within more than one point of view or frame of reference. **Logic** refers to the internal consistency and coherence of thinking, where all parts support one another and make sense together without contradictions. **Significance** denotes the degree of importance or consequence an idea, question, or piece of information holds in relation to the issue at hand, emphasizing a focus on what truly matters rather than on trivial or superficial elements. **Fairness** reflects impartial and honest thinking that considers all relevant viewpoints without bias, favouritism, or self-interest, ensuring that the reasoning used is just and equitable in addressing complex issues.

Research has shown that teaching critical thinking based on the Paul-Elder critical thinking model leads to the promotion of critical thinking in students (Mozaffari et al., 2021), and thus these standards are used for the instrument in this case study.

3 METHODOLOGY

The aim of this case study was to determine the attainment level of GPT-4-turbo model generated essays and discover their potential as a useful tool in the classroom. To this end, a comprehensive checklist for assessing essay structure and intellectual standards was implemented. Four independent teachers (reviewers) from Georgia (1), Portugal (1) and Serbia (2) graded the essays using the proposed checklist. Two teachers teach at secondary schools while the other two are university English language lecturers. The reviewers were not aware of the fact that they were assessing computer-generated essays.

Two topics were given, at C1 / C2 level, and the requirement was 240 to 280 words per essay (C2 Proficiency requirement). The following instructions were given as prompts for ChatGPT:

Prompt for Essay 1

> *Higher education increases the chances of employment. Agree or disagree with this statement. Support your opinion with reasons and examples. Write an essay in around 240 - 280 words.*

Prompt for Essay 2

> *It is better to work at home than in an office. Agree or disagree with this statement. Support your opinion with reasons and examples. Write an essay in around 240 - 280 words.*

The author of this research generated four essays on two topics, each of which was generated on a different computer and account. The essays were then sent to the reviewers for assessment.

Essay organization was assessed using a 14-item instrument based on the previous research by Ljubojević et al. (2023) and the Cambridge English Qualifications scales at C1 level of the CEFR. Each item was scored with different marks (Appendix 1).

The attainment of critical thinking skills was examined using a nine-item instrument whose indicators were derived from the above-mentioned Paul-Elder CT Model (Elder & Paul, 2016) with clarification by Inoshita et al. (2019). Each item was scored from 0 (lowest) to 3 (highest) (Appendix 2).

Four reviewers sent back their assessments to the author and the average mark was considered for the statistical analysis. Statistical analysis and visualization were conducted using Excel (Microsoft 365).

4 RESULTS

4.1 Essay organization

The first set of items in the checklist aimed to determine the essay organization as generated by ChatGPT. As shown in Figure 1, when it comes to the essay format, all essays have appropriate organization: an introduction, body, and conclusion. What stands out in the graph is that the responses were not of appropriate length, i.e. ChatGPT ignored the "240 – 280-word limit" instruction.

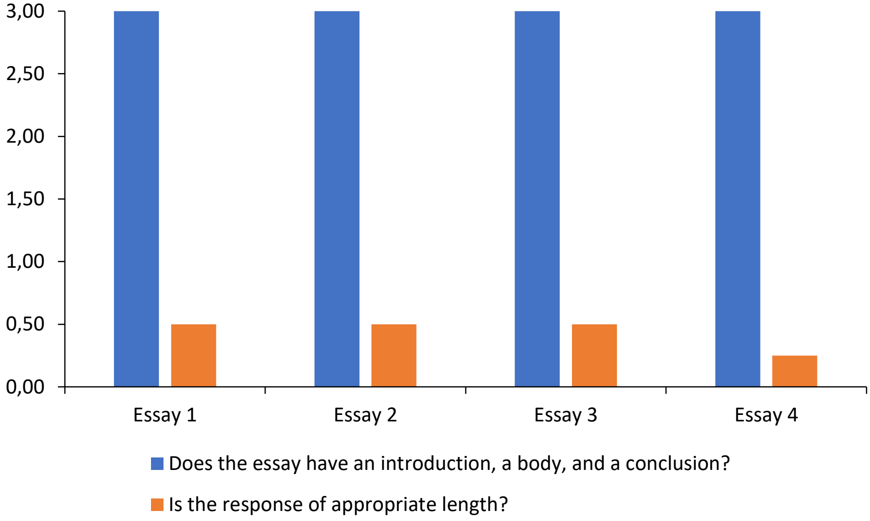

Figure 1: Average grade for the essay format (grading range is 0–3 points)

A more detailed analysis of word count per essay is shown in Table 1:

Table 1: Word count per essay

Essay	**Essay 1**	**Essay 2**	**Essay 3**	**Essay 4**
Word count	312	296	303	355

It should be noted that if the instructions were not given precisely to ChatGPT as a clear prompt for the number of words required, then the length of the essay ChatGPT generates exceeds the word limit by an average of 10.25%.

As can be seen from Figure 2, the introduction was written in an appropriate manner, except for the lack of a funnel introduction.

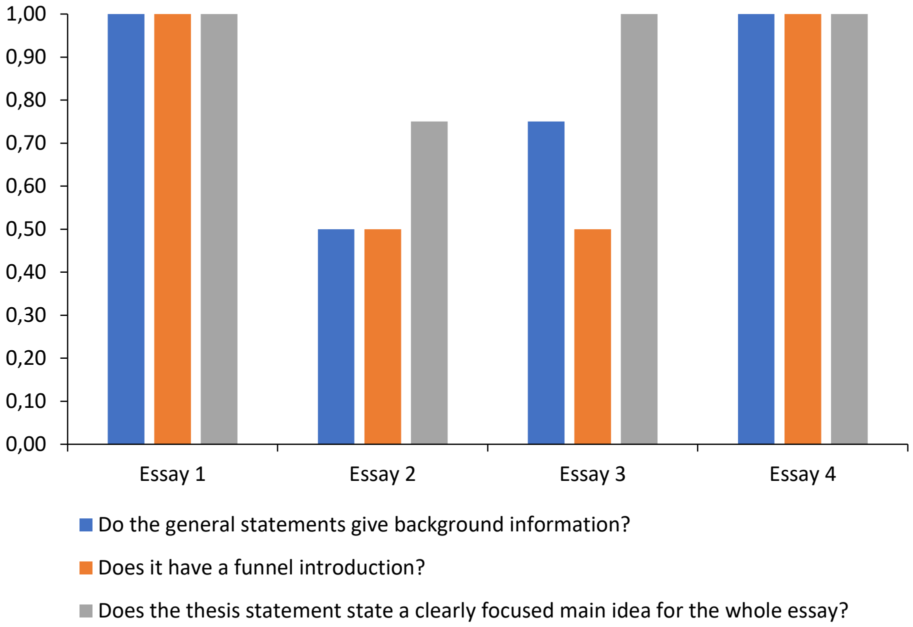

Figure 2: Average grade for the introduction (grading range is 0–1 point)

The third set of items examined the body paragraphs (Graph 3). From the figure below it can be seen that there were no arguments expressing the opposite points of view to that of the essay writers. Moreover, the ChatGPT-generated essays did not follow the rules for structuring paragraphs with a topic sentence, controlling idea, and supporting details. All other aspects were covered in an appropriate manner.

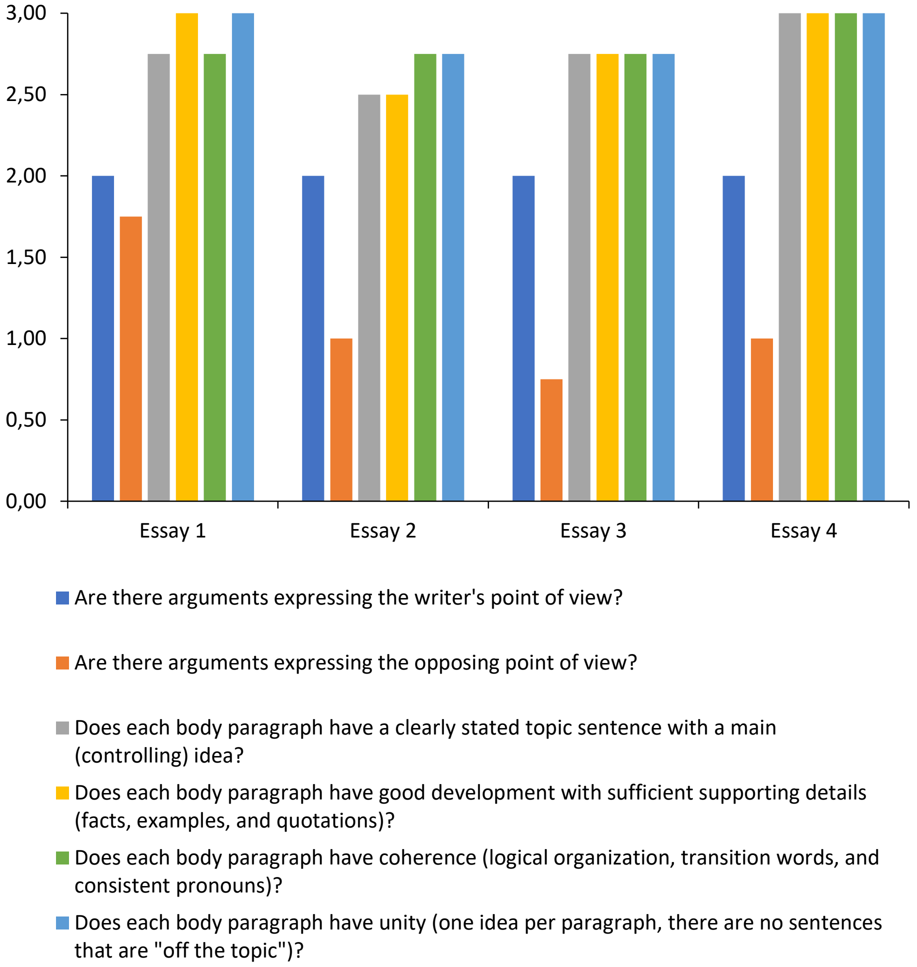

Figure 3: Average grade for the body (grading range is 0–2 points for the first two items, and 0–3 for the following four)

What can be seen from Figure 4 is that the conclusions written by ChatGPT lacked a personal opinion on the topic.

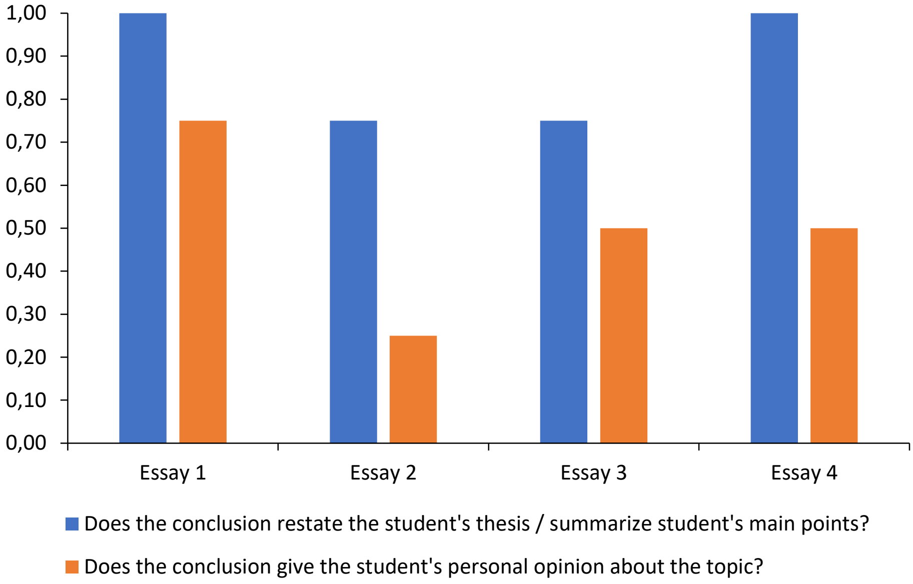

Figure 4: Average grade for the conclusion (grading range is 0–1 point)

When it comes to the language used in the essays, Table 2 shows the results obtained after the assessment. All four teachers stated that the vocabulary range was at a very high level (C1- C2 level: native like, CEFR), immaculate, containing no mistakes or faults. Examples vocabulary from the essays includes items such as, *contend, credentials, top-tier university, lucrative, yield (the same results), a valuable asset, (education) attainment, fosters (growth),* and *gain traction,* all of which belong to the C1-C2 level according to the *Oxford Learners Dictionary*.

Table 2: Average grade for language (grading range is 0–5 points)

Essay 1	Essay 2	Essay 3	Essay 4
4.50	4.00	4.00	4.75

Furthermore, it can be observed that all the essays shared the same linking phrases, such as “in this essay, i will argue”, “to begin with”, “moreover”, “however, it’s essential to”, “for example”, “in conclusion”, “therefore”, and “while some argue that … others contend that”, among others.

4.2 Intellectual standards

The next section of the survey was concerned with critical thinking skills, as measured through the presence of intellectual standards. There were nine items evaluated: clarity, precision, accuracy, depth, breadth, logic, significance, relevance, and fairness.

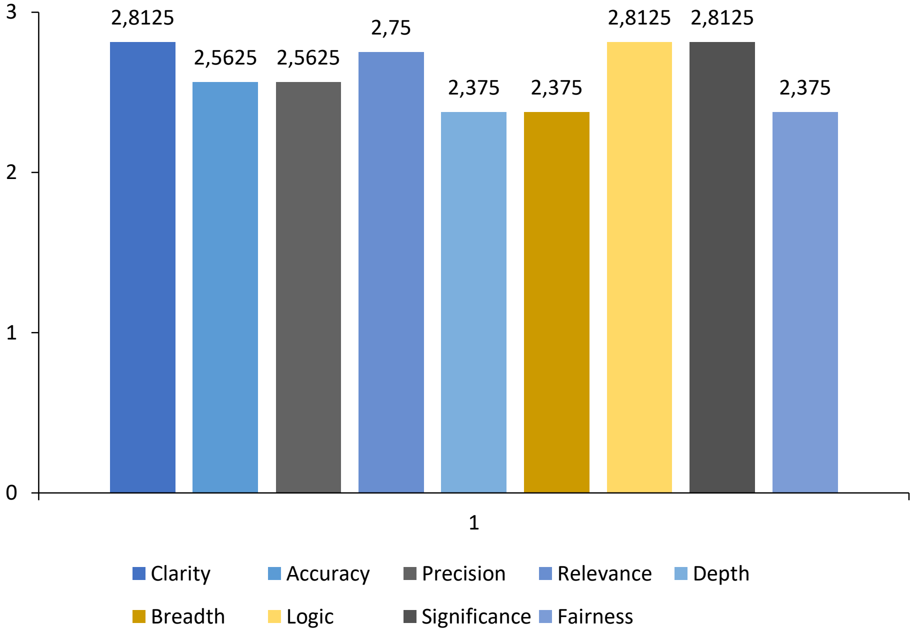

Figure 5: Average grade for intellectual standards (grading range is 0–3 points)

Looking at Figure 5, it is apparent that three items require further attention for deeper analysis and discussion: depth, breadth and fairness. The essays written by ChatGPT did not provide complex arguments in more detail, considering all the relevant views and perspectives, and also did not fairly assess the opposing viewpoints.

5 DISCUSSION

The present case study was designed to evaluate the performance of OpenAI's GPT-4-turbo model in generating argumentative essays and to explore its potential application in English Language Teaching. To this end, four independent teachers graded four different essays generated by ChatGPT based on the given instructions, using a checklist designed for this purpose. This checklist focused on three key areas: the organization of essays, the linguistic devices used in the generated content, and the extent to which elements of critical thinking are incorporated. These areas were addressed through three research questions aligned with the study's objectives.

RQ1 investigated the current level of performance when it comes to essay organization when writing argumentative essays. Prior studies in this area have shown that ChatGPT is outperforming humans in generating argumentative essays (Herbold et al., 2023). According to the findings, it was not surprising to see that ChatGPT excels in essay writing in terms of essay structure, developing an appropriate introduction, body, and conclusion. Our case study confirms the previous findings regarding these indicators (Herbold et al., 2023; Fitria, 2023; Jovic, 2023).

On the other hand, there are some aspects where ChatGPT did not perform well enough: following word limit requirement, writing a funnel introduction/paragraph, providing opposing arguments, and giving personal opinion. The tendency not to stay within the set word limit has already been noted in prior research (Herbold et al., 2023; Jovic, 2023; Ljubojević, 2024). In ELT contexts, word limits are intentionally set to ensure that learners produce focused, level-appropriate responses and that the writing task remains aligned with its communicative and assessment objectives. As such, many standardized tests and classroom assessments include strict word limits as part of their evaluation rubric. For example, Cambridge C1 Advanced writing tasks require approximately 220–260 words, and C2 Proficiency tasks require 240–280 words for Part 1 essays and 280–320 words for Part 2. These lengths are part of the task design to ensure sufficient depth and coherence. Exceeding the word count often means not meeting assessment requirements and leads to penalties, regardless of content quality. A short response suggests task failure, while exceeding the limit may indicate poor argument structure (e.g. main arguments may become weakened, transitions less coherent, conclusions less impactful), irrelevance, repetition, poor organization. A possible explanation for ChatGPT ignoring the word limit might be that the instructions given were not precise in terms of not specific prompting: *Write an essay in around 240 - 280 words.* Since ChatGPT is a chatbot, it was misled by the word "around", although this is a very commonly used by English language teachers. Better results would have been achieved if the instructions had simply required *280 words*:

> …when the model is prompted with a basic instruction, it may produce excessively general results because of insufficient contextual or supplemental details. [...] To address this issue, prompts should be unambiguous and specific, providing sufficient detail for narrowing the response space and aligning the output with the desired goals. (Chen et al, 2025, p. 4)

Clearer and more specific input would thus result in more accurate and structured responses.

Although the AI model provided a clear logical structure by dividing various arguments into paragraphs, no cohesive and coherent paragraph structure was recognizable either in the introduction nor the body paragraphs. This can also be seen in the conclusion.

> When a paragraph is coherent, the reader can move smoothly from sentence to sentence without becoming confused or losing the writer's train of thought. Coherence is achieved by arranging one's material in logical order and by providing signals that help the reader understand the relationships between the ideas in the paragraph. (Parks, Levernier, & Hollowell, 1986, p. 104)

The sentences and paragraphs should smoothly flow from one to another and the use of discourse markers should signal a proper logical structure, but in this study such coherence was somewhat missing in the essays generated by Chat GPT. While Fitria (2023) observed that ChatGPT adheres to conventional essay structure by incorporating topic and supporting sentences, and concluding with a summarizing paragraph, the closer analysis carried out in this case study reveals that such structural conformity does not necessarily ensure coherence. This discrepancy could be attributed to the fact that Fitria did not focus on argumentative essay writing, but on narrative essays that account a personal experience or a series of events (following only the logical sequence of events) (Fitria, 2023). Argumentative essays require students to produce arguments in a reasonable and logical manner, and to establish a position on the topic concisely. Furthermore, one must also discuss the other side's reasons and then rebut them. Students should support their opinions with good argumentation, solid reasons and firm evidence. As a machine learning model, ChatGPT cannot have personal preferences, personal experiences, beliefs, or opinions. Its responses are generated based on patterns and information present in the training data, and it may not always produce perfectly structured paragraphs, especially if the input is complex or there are multiple possible interpretations of it.

RQ2 focused on linguistic devices that are characteristic of ChatGPT-generated content. It generates high-quality argumentative essays, rich in vocabulary (C1–C2 CEFR level), and text linking (numerous transitions signals and linkers). This finding is consistent with that of Herbold et al. (2023), who also stated that "the AI-generated essays are highly structured, which for instance is reflected by the identical beginnings of the concluding sections of all ChatGPT essays" (Herbold, Hautli-Janisz, Heuer, & Trautsch, 2023). In a similar vein, Ljubojević (2024) stated that ChatGPT is especially useful in helping students structure their essays, generate cohesive paragraphs, and improve grammatical precision. It is interesting to note that the use of transition signals and linking devices is one of the biggest challenges in students' essay writing, and ChatGPT can therefore become a useful tool, particularly for learners who struggle with discourse organization and logical flow.

With respect to the third research question, it was found that ChatGPT-generated essays lack the depth, breadth, and fairness called for in Paul and Edler's model (Paul & Elder, 2013), as seen in the results for the essay organization items related to arguments expressing the opposing point of view. The findings of the current study in this respect are completely in line with those of Ljubojević (2024), which also revealed the limitations in ChatGPT's ability to help students with more complex and personalized tasks, such as thesis formulation, maintaining focus in arguments, and expressing personal viewpoints in conclusions (Ljubojević, 2024). Previous research has also shown that chatbots lack an understanding of the context, critical thinking skills, and the ability to make ethical decisions (Fitria, 2023). Ljujuć et al. compared essays generated by ChatGPT to those written by students, and also concluded that ChatGPT-generated essays were limited in terms of reflective insight and critical observations, while those written by students featured complex ideas and different critical attitudes with a more personal tone (Ljujić, Miljković, & Grozdić, 2023). ChatGPT does not fully understand the real-world context of a situation, making it important for learners to critically assess and interpret the information provided by such systems. Humans naturally think from a personal perspective, from a point of view that tends to privilege their own positions. In contrast, while language models have no personal opinions, they can reproduce the biases present in their training data. Learners should thus be taught to recognize and critically assess potential prejudices in ChatGPT-generated content based on this weakness. Moreover, ELT often benefits from personalized instruction, and since ChatGPT does not have any memory of past interactions, it cannot provide a truly personalized learning experience for individuals.

As it suggested by the results of this case study, there are a number of ways that teachers may be able to detect AI-written assignments. One of the red flags is the use of unusually advanced vocabulary (e.g., *yield (the same results), a valuable asset, (education) attainment, fosters (growth), gain traction*), which may not match the student's typical register or previous work. Next, AI-generated essays often avoid personal anecdotes, local references, or classroom-specific content, which students naturally refer to. Teachers who are familiar with a student's life or way thinking will quickly notice this. Moreover, while the grammar in AI-generated essays may be flawless, the ideas often lack depth, nuance, or logical development, as the AI tends to list reasons rather than build a sustained line of reasoning or effectively address counterarguments. In addition, AI-generated texts sometimes fail to adhere strictly to the assigned word limit, producing responses that are either too short brief or too long, with pacing that may feel rushed or unnaturally dense in an attempt to meet structural expectations.

Still, although this case study has demonstrated that ChatGPT cannot generate perfect argumentative essays, it has certain limitations in terms of the small sample of essays used and the number of teachers who participated as reviewers, which means that its findings need to be interpreted with caution.

To date, no similar studies have been undertaken longitudinally because ChatGPT is a new technology. Further research should thus aim to investigate the impact of using chatbots on both essay writing and developing critical thinking skills over time. It could also explore the different ways of using AI chatbots and different argumentative writing models (the Toulmin Model or Rogerian Argumentation Model) in order to better develop the critical thinking skills of students.

6 IMPLICATIONS FOR ELT

This research adds to the current understanding of how ChatGPT can be used in ELT, and provides insights for language teachers when it comes to using AI technology with their students. It has shown that numerous high-quality argumentative essays can be generated by ChatGPT, which has significant implications for language teaching and instruction. Herbold et al. (2023) have already called for "re-inventing homework" because "the way students do homework and teachers assess it needs to change in a world of generative AI models" .

The findings from this case study make several contributions to the current literature. Prior studies focused on AI as an interlocutor (a learning companion to assist student's language learning), as a simulation of an authentic learning environment, for the transmission of information, as a helpline, and as a source of recommendation for further learning materials (Huang, Hew, & Fryer, 2022) to support learning and teaching (Hong, 2023). In the same vein, this case study highlights the potential usefulness of a *flipped approach*, where students first learn/write about topics for themselves using AI-based technology (such as ChatGPT) and then discuss it later in class. Students could generate essays on a given topic on their own, then analyse the essays with the help from the teacher or checklists (underline the key phrases, identify and understand the topic sentence, supporting ideas, the power of arguments...). By being involved in the process of generating texts, choosing the appropriate version, analysing, and evaluating arguments, claims or evidence for their essays, making inferences using inductive or deductive reasoning and making final decisions, especially when they need to add a personal opinion or opposite arguments to the generated ChatGPT essays, students will be exposed to the process of developing critical thinking skills. As Herbold et al., (2023) put it: "AI models need to be seen as an integral part of education, but one which requires careful reflection and training of critical thinking skills."

Besides developing critical thinking skills, students can work on acquiring good language models through phrasal re-use from texts called *sentence skeletons* (Swales and Feak, 1994). By looking at the parts of sentences written by ChatGPT, students will be able to see patterns and practice essay writing by putting their own content into these skeletons. As can be seen from Table 3, there are sentence skeletons that ChatGPT uses generating the essays which can be very useful as models for learning.

Table 3: Introductions from the essays generated by ChatGPT and their skeletons

Essay 1	Skeleton
In today’s rapidly evolving job market, the role of higher education in increasing employment prospects is a topic of considerable debate. While some argue that a college degree significantly enhances one’s chances of securing a job, others contend that it is not always a guarantee of employment. In this essay, I will argue that higher education does indeed increase the chances of employment, but its effectiveness depends on various factors.	In today’s …, the role of … in … is a topic of considerable debate. While some argue that …, others contend that it is not always …. In this essay, I will argue that … does indeed …, but its effectiveness depends on various factors.
Essay 2	
I strongly agree that higher education increases the chances of employment. This belief is grounded in the fact that higher education equips individuals with valuable skills, knowledge, and qualifications that are highly sought after by employers.	I strongly agree that …. This belief is grounded in the fact that ….
Essay 3	
The debate over whether it is better to work from home or in an office has gained traction in recent years, driven by advances in technology that have made remote work more accessible. While both options have their advantages and disadvantages, this essay contends that working from home offers numerous benefits and is a preferable choice for many. These advantages include increased flexibility, improved work-life balance, reduced commuting stress, and enhanced productivity.	The debate over whether it is better to … has gained traction in recent years, driven by advances in technology that have made …. While both options have their advantages and disadvantages, this essay contends that … offers numerous benefits and is a preferable choice for many. These advantages include ….
Essay 4	
I strongly agree that working from home is often a superior option compared to working in a traditional office setting. This shift in perspective has been accelerated by the COVID-19 pandemic, which forced many businesses to adopt remote work practices. Several reasons and examples support this viewpoint.	I strongly agree that … is often a superior option compared to …. This shift in perspective has been accelerated by …, which forced …. Several reasons and examples support this viewpoint.

7 CLOSING REMARKS

This case study set out to gain a better understanding of using new technology in classrooms, specifically chatbots and ChatGPT. It has demonstrated that OpenAI’s GPT-4-turbo model effectively generates argumentative essays based on the input in terms of essay organization, uses high quality linguistic devices (metalanguage, transition signals, linkers) and can focus on the elements of critical thinking. In practice, students are not hesitant to use this new technology, and thus it is up to teachers to change the ways they approach traditional assignments, such as essay writing. Teachers could benefit from using a flipped approach and sentence skeletons technique, as explained in this paper. Students can then engage in critical reflection on the outputs generated by AI tools, particularly in terms of their reasoning quality and factual accuracy. More importantly, this integration also raises ethical considerations, including academic integrity, responsible use, and the need to foster students' own critical thinking skills rather than dependence on AI-generated content. Taken together, these findings suggest a role for ChatGPT in promoting not only writing but also critical skills, and it should be a tool that is not neglected but instead receives appropriate attention within educational settings.

Acknowledgements

This research was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-136/2025-03/ 200018).

Appendix 1 Essay Structure Checklist (Ljubojević et al., 2023)

Indicator number	**Indicator**	**Maximum score**
1	Does the essay have an introduction, a body, and a conclusion?	**3**
2	Is the response of appropriate length?	**1**
	Introduction	
3	Do the general statements give background information?	**1**
4	Does it have a funnel introduction?	**1**
5	Does the thesis statement state a clearly focused main idea for the whole essay?	**1**
	Body	
6	Are there arguments expressing the writer's point of view?	**2**
7	Are there arguments expressing the opposing point of view?	**2**
8	Does each body paragraph have a clearly stated topic sentence with a main (controlling) idea?	**3**
9	Does each body paragraph have good development with sufficient supporting details (facts, examples, and quotations)?	**3**
10	Does each body paragraph have unity (one idea per paragraph, there are no sentences that are "off the topic")?	**3**
11	Does each body paragraph have coherence (logical organization, transition words, and consistent pronouns)?	**3**

Indicator number	**Indicator**	**Maximum score**
	Conclusion	
12	Does the conclusion restate student's thesis / summarize student's main points?	**1**
13	Does the conclusion give student's personal opinion about the topic?	**1**
14	**Language (choose one from the list below)**	**5**
Performance level descriptors	Uses a (wide) range of vocabulary, including less common lexis, effectively and precisely. Uses a wide range of simple and complex grammatical forms with full control, flexibility and sophistication. Errors, if present, are related to less common words and structures, or occur as slips.	5
	Performance shares features of Bands 3 and 5.	4
	Uses a range of vocabulary, including less common lexis, appropriately. Uses a range of simple and complex grammatical forms with control and flexibility. Occasional errors may be present but do not impede communication.	3
	Performance shares features of Bands 1 and 3.	2
	Uses a range of everyday vocabulary appropriately, with occasional inappropriate use of less common lexis. Uses a range of simple and some complex grammatical forms with a good degree of control. Errors do not impede communication.	1
	Total	**30**

Appendix 2 Paul-Elder Critical Thinking Framework (Paul & Elder, 2013)

Indicator Number	**Intellectual Standards Indicators**	**Maximum score**
1	Clarity	**3**
2	Accuracy	**3**
3	Precision	**3**
4	Relevance	**3**
5	Depth	**3**
6	Breadth	**3**
7	Logic	**3**
8	Significance	**3**
9	Fairness	**3**
	Total	**27**

REFERENCES

Aristotle. (2024). Retorika. Beograd: Štampar Makarije.

Bašić, Z., Banovac, A., Kružić, I., & Jerković, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. Humanities and Social Sciences Communications, 10(750). doi:10.1057/s41599-023-02269-7

Bibauw, S., François, T., & Desmet, P. (2019). Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based CALL. Computer Assisted Language Learning, 827–877. doi:10.1080/09588221.2018.1535508

Chang, D. H., Lin, M. P.-C., Hajian, S., & Wang, Q. Q. (2023). Educational Design Principles of Using AI Chatbot That Supports Self-Regulated Learning in Education: Goal Setting, Feedback, and Personalization. Sustainability, 15(17), 12921. doi:10.3390/su151712921

Chen, B., Zhang, Z., Langrene, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. Patterns, 6(6). doi:10.1016/j.patter.2025.101260

Elder, L., & Paul, R. (2013). Critical Thinking: Intellectual Standards essential to Reasoning Well Within Every Domain of Thought. Journal of Developmental Education, 36(3), 34–35. https://files.eric.ed.gov/fulltext/EJ1067273.pdf

Elder, L., & Paul, R. (2016). The Thinker's Guide to Analytic Thinking: How to Take Thinking Apart and What to Look for When You Do (2nd ed.). Tomales, CA: Rowman & Littlefield Publishers / The Foundation for Critical Thinking.

Fitria, T. N. (2023). Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. ELT FORUM Journal of English Language Teaching, 12(1). doi:10.15294/elt.v12i1.64069

Godwin-Jones, R. (2024). Generative AI, Pragmatics, and Authenticity in Second Language Learning. doi:10.48550/arXiv.2410.14395

Herbold, S., Hautli-Janisz, A., Heuer, U. K., & Trautsch, A. (2023). AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays. Sci Rep. https://arxiv.org/pdf/2304.14276.pdf

Hong, W. C. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research. Journal of Educational Technology and Innovation (JETI), 5(1). https://jeti.thewsu.org/index.php/cieti/article/view/103

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. Journal of Computer Assisted Learning, 38(1), 237–257. doi:10.1111/jcal.12610

Inoshita, A., Garland, Sims, K. K., Keuma, T., J. K., & Williams, T. (2019). English Composition - Connect, Collaborate, Communicate. Honolulu: University of Hawaii, OER.

Jovic, M., & Mnasri, S. (2024). Evaluating AI-generated emails: A comparative efficiency analysis. World Journal of English Language, 14(2), 502–517. https://doi.org/10.5430/wjel.v14n2p502

Khampusaen, D. (2025). The impact of ChatGPT on academic writing skills and knowledge: An investigation of its use in argumentative essays. LEARN Journal: Language Education and Acquisition Research Network, 18(1), 963–988.

Ljubojević, D. (2024). Student Feedback on the Effectiveness of ChatGPT for Essay Writing. In M. Saqr, S. López-Pernas, M. Á. Conde, M. Raspopović Milić, & E. Kisić (eds.), Proceedings for the 15th International Conference on e-Learning 2024, 3938 (pp. 116–126). https://ceur-ws.org/Vol-3938/Paper_12.pdf

Ljubojević, D., Kadijevich, D. M., & Gutvajn, N. (2023). Towards a Checklist to Evaluate Argumentative Essays Composed by ChatGPT. In M. Saqr, S. López-Pernas, M. Á. Conde, & M. Raspopović Milić (eds.), Proceedings for the 14th International Conference on eLearning (eLearning-2023), 3696 (pp. 82–90). https://ceur-ws.org/Vol-3696/article_9.pdf

Ljujić, B., Miljković, J., & Grozdić, V. (2023). ChatGPT and Academic Writing in Higher Education. 29th International Scientific Conference "Educational Research and School Practice" (pp. 104–109). Belgrade: Institute for Educational Research.

Mavropoulou, E., & Arvanitis, P. (2023). Contribution/Implications of Chatbot integration on teaching and learning a foreign language. ICERI2023 Proceedings (pp. 7596–7605). doi:10.21125/iceri.2023.1899

Mozaffari, Z., Abdollahi, M. H., Farzad, V., & Ghayedi, Y. (2021). The effectiveness of critical thinking training based on the Paul-Elder model on students' critical thinking skills. Journal of Educational Phychology Studies, 18, 20–29. doi:10.22111/JEPS.2021.6536

Nakrowi, Z. S., Ansori, D. S., Mulyati, Y., & Setyaningsih, Y. (2023). The use of intellectual standards to assess the quality of students' argumentative writings. LITERA, 22(2), 200–212. doi:10.21831/ltr.v22i2.60465

National Council for Excellence in Critical Thinking. (1987). The Foundation for Critical Thinking. https://www.criticalthinking.org/pages/defining-critical-thinking/766

Oshima, A., & Hogue, A. (2006). Writing academic English (4th ed.). Pearson Longman.

Parks, A. F., Levernier, J. A., & Hollowell, I. M. (1986). Structuring Paragraphs: A Guide to Effective Writing, Second Edition. New York: St. Martin's Press.

Paul, R., & Elder, L. (2013). Critical Thinking: Intellectual Standards Essential to Reasoning Well Within Every Domain of Human Thought, Part Two. Journal of Developmental Education, 37(1), 32–36. https://files.eric.ed.gov/fulltext/EJ1067269.pdf

Retelj, A. (2023). Risks and Opportunities of Planning German Lessons Using the Language Model ChatGPT. Journal for Foreign Languages, 15(1), 259–275. doi:10.4312/vestnik.15.259-275

Swales, J., & Feak, C. (1994). Academic writing for graduate students. University of Michigan Press.

Toulmin, S. E. (2003). The Uses of Argument. New York: Cambridge University Press.

Young, R. E., Becker, A. L., & Pike, K. L. (1970). Rhetoric: Discovery and Change. New York: Harcourt, Brace & World.