ISSN: 2716-9189
Journal of Studies in Language, Culture, and Society (JSLCS).
Received: 15/02/2025 Accepted: 16/04/2025

E-ISSN: 2676-1750
(08) 01, 2025 (PP. 188-205)
Published: 30/04/2025

ARTIFICIAL INTELLIGENCE IN LANGUAGE EDUCATION: EXPLORING PROMPTING STRATEGIES TO FOSTER ARGUMENTATIVE WRITING SKILLS

Marco Mezzadri [1] **Mariapaola Paita** [2]
[1] **University of Parma (Italy)**
marco.mezzadri@unipr.it
[2] **University of Parma (Italy)**
mariapaola.paita@unipr.it

Abstract: The rapid advancement of Artificial Intelligence (AI), particularly Large Language Models (LLMs), calls for a thorough examination of not only of the opportunities for innovation but also the conditions necessary to foster a productive and informed human-machine relationship in education. This article explores the integration of prompt engineering as a critical transversal skill for effectively implementing AI-based technologies in language education while promoting the development of digital competencies among educators and learners. The study observes variations in interactions between learners and ChatGPT during educational activities designed to enhance argumentative writing skills. Specifically, it examines the reliability and feasibility of ChatGPT in providing meaningful and relevant feedback on argumentative writing through the analysis of task-specific interactions between secondary school students and the language model. Additionally, it explores how the iterative process of prompt construction and refinement adopted by participants shapes ChatGPT's responses when evaluating learners' argumentative texts. By analysing the impact of different prompt strategies on the chatbot's outputs, the study offers practical guidelines for leveraging AI to foster language acquisition, AI literacy, and critical thinking through the evaluation, validation, and optimization of learners' interactions with ChatGPT.

***Keywords**:* ChatGPT, human-AI mediation, language education, Large Language Models, prompt engineering

How to cite the article :
Mezzadri, M., & Mariapaola, P. (2025). Artificial Intelligence in Language Education: Exploring Prompting Strategies to Foster Argumentative Writing Skills. *Journal of Studies in Language, Culture, and Society (JSLCS)8*(1), 188-205.

---

[1] **Corresponding author:** Marco Mezzadri **ORCID ID : https://orcid.org/0000-0002-4043-6330**

1. **Introduction**

The nature of the discipline, known in the Italian academic landscape as Glottodidactics and increasingly referred to as Educational Linguistics, plays a significant role in shaping decisions influenced by the growing impact of Artificial Intelligence (AI). These decisions, we argue, should be grounded in epistemological reflections. However, only a limited number of studies have addressed these themes, with foundational contributions from the past two decades including works by De Mauro and Ferreri (2005), Balboni (2011), Chini and Bosisio (2014). The challenges posed by language education have therefore driven the discipline to adopt an increasingly interdisciplinary framework. While remaining faithful to its theoretical and practical origins, it has fostered connections not only with language sciences and other soft sciences but also with fields such as neuroscience and cognitive sciences.

This theatrical-practical nature, however, presents challenges. As Balboni (2011) points out, "theoretical-practical sciences such as medicine, economics, or glotto-didactics, which aim simultaneously at problem-solving and knowledge-building, are more difficult to model because they draw from various scientific disciplines (both theoretical and practical), requiring the homogenization of different parameters and paradigms" (p. 22). This interdisciplinary relationship should not be approached from an "applicative" perspective but rather from an "implicative" one. In other words, it is not about applying principles and methods from one domain to another but about identifying implications from other domains that can contribute to expanding knowledge within one's own field. This approach is particularly relevant when collaborating with developers of AI-based technologies and, even more critically, when addressing the everyday needs of language education using tools such as Large Language Models (LLMs), including ChatGPT, which constitute the foundation of this study. Therefore, we argue that an "implicative" approach is essential for mitigating the risks associated with uncritically adopting elements from another disciplines or tools, such as LLMs.

The overarching question guiding this research is how LLMs can be effectively approached, beginning with an understanding of their mechanisms, followed by an evaluation of their implications in language education, and ultimately harnessing their didactic potential. To develop an interpretative framework, this research begins by examining its own disciplinary identity to uncover the values and beliefs that shape its relationship with LLMs. An analysis of the CEFR (Council of Europe [COE], 2001) and its Companion Volume (COE, 2020), regarded as the foundational framework for language education, reveals that AI is not explicitly addressed within these resources. Nevertheless, even prior to the emergence of ChatGPT, the literature highlighted the use of Machine Translation for multilingual reading tasks, as well as its application in writing tasks (Nurminen & Papula, 2018) and oral comprehension tasks. This body of research emphasises that, at all levels of language competence, human abilities can already be integrated with AI to perform tasks such as text summarization, thereby extending its application beyond literal translation.

Furthermore, the CEFR (COE, 2001) and its Companion Volume (COE, 2020) provide a foundation for identifying points of intersection, or common denominators, that support an "implicative" approach to AI. Foremost among these is mediation, as the relationship with AI inherently requires competencies rooted in a clear understanding of its underlying mechanisms and the ability to articulate thoughts effectively. Underpinning this human-AI mediation is one of the most emphasised soft skills: critical thinking, framed within a problem-solving approach. This ability is fundamental for facilitating effective interactions with LLMs through textual inputs, referred to as prompts (Liu et al., 2023), which users provide to elicit relevant and coherent responses or to direct LLMs in performing specific tasks (White et al., 2023).

Mediation thus occurs through the prompt, which is instrumental in shaping the quality of the interactions with LLMs by tailoring them to meet users' needs and expectations (Mondal et al., 2024). By providing clear instructions, prompts play a crucial role in directing the model's actions and ensuring that the desired outcomes are achieved (M. Wang et al., 2024). However, due to the prompt sensitivity characterizing LLMs, even minor variations in prompt structure and content can significantly influence the generated responses (Zhao et al., 2021; Lu et al., 2022). As a results, the ability of LLMs to produce accurate and user-aligned outputs largely depends on crafting effective prompts (Cain, 2024).

In this study, human-AI mediation is addressed by observing variations in learners' interactions with ChatGPT to analyse how the iterative process of prompt construction and refinement shapes the model's responses during the evaluation of an argumentative text. Building on this analysis, the study aims to foster a critical reflection on prompting strategies that enhance the effectiveness of LLMs in developing argumentative writing skills, placing particular emphasis on learners' and educators' acquisition of the digital competencies and soft skills needed to optimize, mediate, and evaluate interactions with AI-based technologies. This objective is framed through the following research questions:

Q1. Which aspects of feedback does ChatGPT prioritize when evaluating an argumentative text?

Q2. How, and to what extent, prompt characteristics influence the quality of feedback provided by ChatGPT?

Q3. Can prompt engineering be recognized as a soft skill necessary for the effective implementation of LLMs in language education?

1. Literature Review

Prompt engineering is an emerging field of study focused on designing and refining prompts to enhance the quality of interactions with LLMs, thereby fostering their effective and responsible use (Chen et al., 2024). Research in this field aims to identify the most effective prompting techniques by assessing the accuracy of responses in relation to the prompts provided and analysing interactions with LLMs to gain deeper insights into their operating principles and internal mechanisms (Linardatos et al., 2021). For instance, Li et al. (2024) proposed a classification of prompting techniques based on their role in facilitating tasks that require the model to perform a series of intermediate steps. Additionally, several studies investigated how linguistic and structural modifications to prompts can influence the quality of the model's responses (O'Connor & Andreas, 2021; Liu et al., 2024).

From a didactic perspective, prompt engineering is central for optimizing interactions with LLMs and developing the digital competencies and soft skills required to comprehend, analyse, and evaluate prompts' quality in relation to the outputs they generate (Bozkurt, 2023; Walter, 2024). To this end, research has primarily focused on three key areas. The first involves the development of theoretical frameworks and guidelines for the effective formulation of prompts. For example, the CLEAR framework (Lo, 2023) outlines essential characteristics for prompt construction, while the AIRPROMPT framework (Korzynski et al., 2023) provides design guidelines and highlights key components of effective prompts. Moreover, Anderson et al. (2023) investigate the internal mechanisms of LLMs, identifying the main factors for optimizing their performance through prompting. Finally, several studies concentrated on the iterative process of prompt construction, testing, and validation, detailing the fundamental steps required to improve the model's performance (Eager & Brunton, 2023; Cain, 2024).

Another field of inquiry explores LLMs' capability to create didactic materials and activities aimed at strengthening learners' linguistic knowledge and skills. Studies in this field investigate how different prompt techniques influence the validity and accuracy of responses generated by LLMs. Lee et al. (2023) proposed and validated a prompt engineering method for Automatic Question Generation (AQG) using ChatGPT, employing various prompt formats to create reading comprehension tasks. Through the implementation of this method, researchers identified the most effective prompting strategies for AQG. However, the results also highlight a decline in performance when generating multiple-choice questions or cloze activities. A similar study examined ChatGPT's capability to generate factual knowledge on the topic of gamification in response to different types of prompts (Liu, 2023). The analysis of interactions revealed improved accuracy and reliability in responses when detailed and interactive prompts were employed.

Furthermore, this area of research underscores the importance of prompt engineering in developing writing skills, particularly in the production and evaluation of argumentative texts. Mo Wang et al. (2024) conducted a study on the accuracy of ChatGPT's feedback in response to learners' argumentative writing skills to identify the input elements that most significantly impact the quality of the model's responses. The findings revealed that the absence of connectives, the complexity of the argumentations, and the length of the text can hinder the model's ability to effectively evaluate textual coherence and cohesion. Additionally, Su et al. (2023) discuss the potential applications of ChatGPT during the planning, production, and revision phases of argumentative writing, providing guidance on prompt construction to improve feedback quality. However, despite highlighting the fundamental role of prompt engineering, these studies primarily focus on analysing LLM outputs without providing detailed insights into the specific techniques employed for prompt construction.

The third area of investigation focuses on the development of AI literacy skills to implement AI-based technologies within the educational context (see Ranieri et al., 2023). More specifically, research focuses on identifying the key competencies required for effective and responsible interaction with LLMs (Cain, 2024; Walter, 2024; see also Theophilou et al., 2023). Recent studies, however, indicate that inexperienced users often adopt an opportunistic, trial-and-error approach (Dang et al., 2022), experimenting with prompt strategies in a random and disorganized manner (Zamfirescu-Pereira et al., 2023). For instance, Woo et al. (2023) analysed learners' interactions with four conversational chatbots during a writing task and observed that their prompts frequently included generic instructions, questions, and incomplete phrases. In contrast, prompt refinement not only enhances the performance of LLMs but also has a positive impact on participants' engagement and problem-solving skills (Sawalha et al., 2024). Finally, Knoth et al. (2024) attribute the challenges in crafting effective prompts to a lack of users' prompting skills and limited understanding of LLM functionality, which often leads to "humanise" interactions with these tools.

Despite growing interest, research on prompt engineering in the field of education remains in its early stages (Sawalha et al., 2024). To date, most contributions focus on proposing theoretical frameworks for prompt design, while empirical studies exploring learners' and educators' prompting skills are still scarce (Knoth et al., 2024). Most participants in these studies are from STEM disciplines, particularly computer science and engineering (Sawalha et al., 2024), with research primarily focusing on analysing interaction during programming language generation tasks (see Sheese et al., 2024; Kazemitabaar et al., 2023). This trend is also evident in the field of educational linguistics, particularly in relation to developing writing skills, where studies tend to focus on examining the quality of feedback generated by LLMs while neglecting an in-depth and systematic analysis of prompt effectiveness.

2. Methodology

2.1 Context

This research is part of an empirical study conducted at a high school in Parma, Italy, aimed at exploring the impact of AI, specifically ChatGPT, on the acquisition and enhancement of students' argumentative writing skills. The study was carried out from October to December 2023 and involved four classes attending the penultimate year of high school and two classes attending the second year. Participants were divided into an experimental group of 58 students and a control group of 38 students. Throughout the study, both groups engaged in the same instructional phases, which consisted of activities designed to produce and revise argumentative texts aimed at fostering their writing skills. However, while the experimental group utilized ChatGPT to assist with these activities, the control group completed each task without the support of the LLM.

2.2 Participants

For this study, 58 participants from the experimental group were selected. These students interacted individually with ChatGPT to receive feedback on the quality of the argumentative texts they had produced, following instructions provided by the researchers and refraining from using the tool during the writing process. Moreover, all participants provided informed consent for the anonymous processing of personal data for research purposes.

As a result, 58 interactions with ChatGPT were collected and analysed, categorized as follows: 36 interactions focused on evaluating an argumentative text about the impact of social media on contemporary society. Among these, 25 interactions were from students in the penultimate year of high school, and 11 were from students in the second year. The remaining 22 interactions involved the evaluation of an argumentative text addressing the theme of technology and automation in the future workplace, with 14 interactions conducted by students in the penultimate year of high school and 8 students in the second year.

Each participant used ChatGPT to evaluate both argumentative texts; however, it was not possible to collect and analyse all interactions with the LLM. The primary reason was that the interactions were gathered at the end of the study, and some were no longer available for external sharing. Nevertheless, we consider the quality of the collected interactions valid for our research purposes as the type of writing task and the linguistic and argumentative aspects of the texts remained consistent across both cases despite the different topics addressed.

2.3 Procedures

This study employs a qualitative approach to analyse interactions between students and ChatGPT to understand how variations in prompt construction used during the evaluation of an argumentative text may affect the accuracy of the feedback generated by the LLM. The analysis was conducted in two phases. In the first phase, ChatGPT's outputs were automatically analysed using an assessment rubric specifically designed to evaluate both the linguistic and argumentative aspects of the text. The linguistic evaluation focused on clarity and coherence, register, grammatical accuracy, punctuation, spelling, and syntactic clarity and complexity, with particular emphasis on the use of connectives. The argumentative evaluation examined the introduction and thesis statement, the development of arguments and counterarguments, the appropriate use of evidence, and the conclusion. This rubric served as a reference point for determining the recall rate of key elements in argumentative writing within the feedback generated by ChatGPT.

In the second phase, a qualitative analysis was conducted on the prompting strategies employed by students during their interactions with ChatGPT. This analysis aimed to investigate how, and to what extent, linguistic and structural variations in prompts influence the accuracy of the feedback provided. Three key factors in prompt formulation were identified by drawing on the contributions of Korzynski et al. (2023), Eager and Brunton (2023), Lo (2023), O'Connor and Andreas (2021), and Anderson et al. (2023), providing a comprehensive understanding of effective strategies for prompt crafting. The first factor concerns the components of the prompt, consisting of:

1. Clear and detailed instructions outlining the task to be performed.
2. Input data, or information related to the content on which the task is based.
3. Contextual information providing guidance to the LLM on how to execute the task.

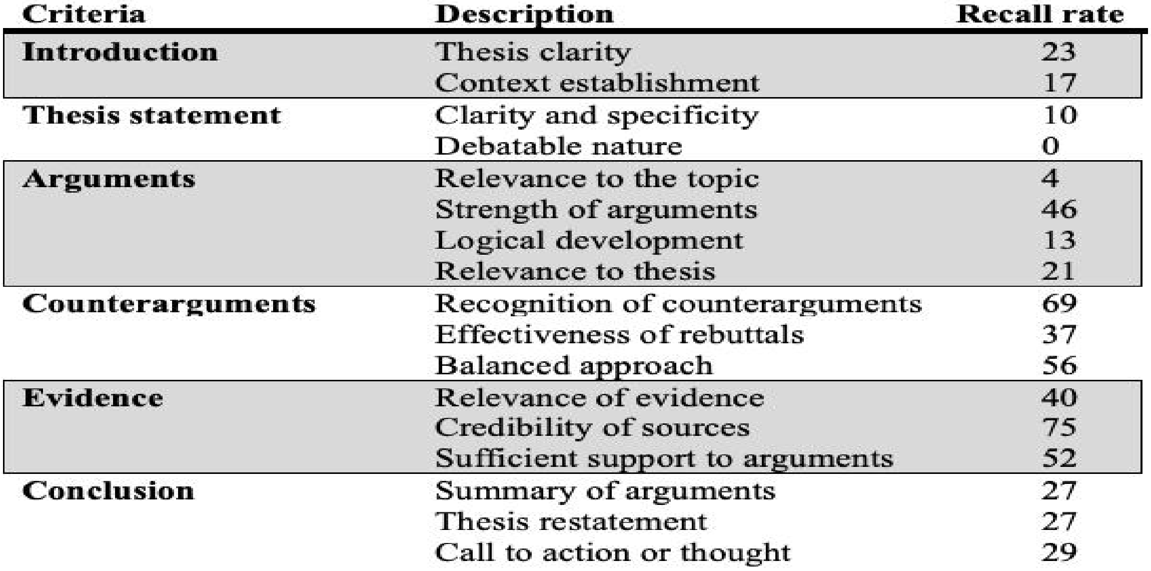
The second factor pertains to the language, which must be direct and exhaustive in defining the task objectives, using relevant keywords and employing concise, unambiguous phrases to ensure accurate and relevant responses. The final factor involves the structure of the prompt, which encompasses the organization of its components, and the complexity of the task. Each interaction was analysed according to these factors to identify variations in the quality of ChatGPT's feedback in relation to the prompts used.

3. Results

The analysis of the interactions revealed a varied distribution of linguistic and argumentative elements in ChatGPT's feedback. As shown in Table 1, ChatGPT demonstrates a recall rate exceeding 50% for aspects related to the argumentative structure of the text, including the inclusion of counterarguments, the balance between arguments and counterarguments, the validity of evidence, and its relevance to the thesis. In contrast, aspects related to the introduction, the logical development of arguments, and the conclusion are addressed less frequently, with recall rates of 30% or lower. This decline is particularly pronounced for elements requiring a clearer distinction between the thesis and the antithesis, such as the clarity of the thesis in the introduction, the relevance of arguments with respect to the thesis, and the summary of arguments in the conclusion. ChatGPT's difficult in differentiating between aspects related to either the thesis or the antithesis is even more evident when evaluating elements concerning the thesis statement, where recall rates are 10% or lower.

Table 1.

Recall Rate (%) of Argumentative Elements in ChatGPT's Feedback

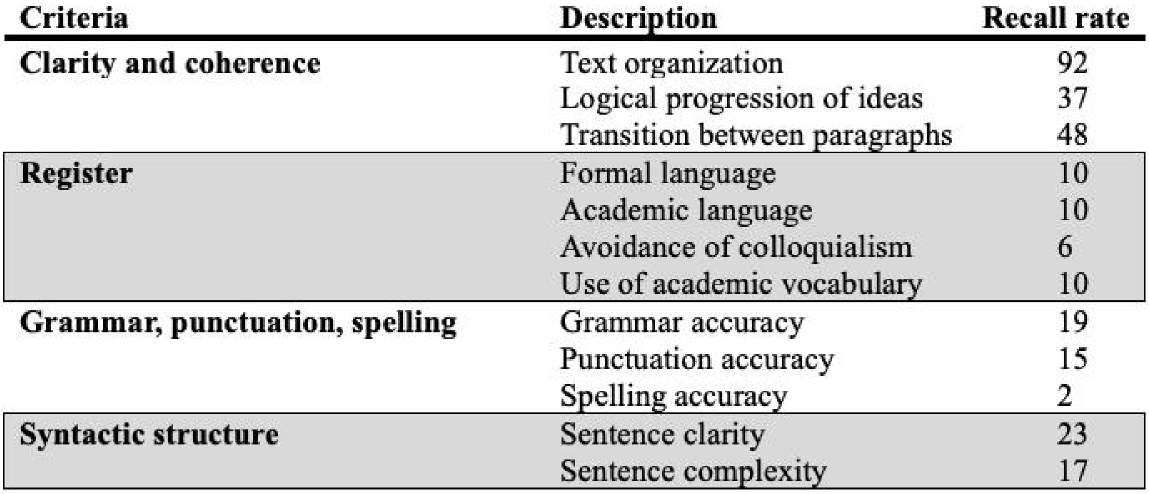


Criteria	Description	Recall rate
Introduction	Thesis clarity	23
	Context establishment	17
Thesis statement	Clarity and specificity	10
	Debatable nature	0
Arguments	Relevance to the topic	4
	Strength of arguments	46
	Logical development	13
	Relevance to thesis	21
Counterarguments	Recognition of counterarguments	69
	Effectiveness of rebuttals	37
	Balanced approach	56
Evidence	Relevance of evidence	40
	Credibility of sources	75
	Sufficient support to arguments	52
Conclusion	Summary of arguments	27
	Thesis restatement	27
	Call to action or thought	29

Similar results were observed in the analysis of linguistic elements included in the feedback. As illustrated in Table 2, ChatGPT predominantly concentrated on evaluating aspects related to clarity and coherence. Notably, textual organization was the most frequently mentioned criterion, with a recall rate of 92%. Moreover, elements such as the logical progression of ideas and transitions between paragraphs, although less frequent, were still identified in 37% and 48% of the responses, respectively. However, a decrease in performance was observed in areas such as grammatical accuracy, punctuation, and syntactic clarity and complexity, with recall rates below 30%. This decline is more evident for aspects such as register and spelling, where recall rates drop to 10% or less.

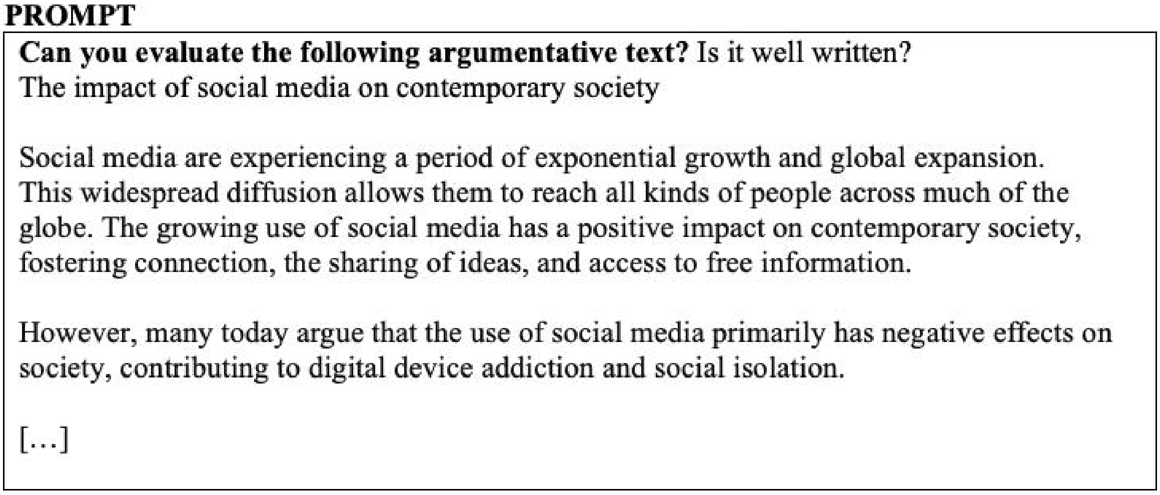
Table 2.

Recall Rate (%) of Linguistic Elements in ChatGPT's Feedback



Criteria	Description	Recall rate
Clarity and coherence	Text organization	92
	Logical progression of ideas	37
	Transition between paragraphs	48
Register	Formal language	10
	Academic language	10
	Avoidance of colloquialism	6
	Use of academic vocabulary	10
Grammar, punctuation, spelling	Grammar accuracy	19
	Punctuation accuracy	15
	Spelling accuracy	2
Syntactic structure	Sentence clarity	23
	Sentence complexity	17

Building on the initial findings regarding the distribution of feedback elements, a detailed analysis was conducted to explore how variations in the linguistic and structural features of prompts influence the quality of ChatGPT's responses. In particular, the presence or absence of key components, together with their position within the prompt and the language used for its formulation played a significant role in determining the relevance, clarity, and accuracy of the feedback provided by the LLM. As shown in Figure 1, prompts that included explicit and precise task instructions with relevant input data and contextual information were more likely to elicit responses that addressed the main elements of the argumentative writing. In this example, ChatGPT provided relevant feedback on both the linguistic and argumentative elements of the text. On the one hand, it addressed aspects such as clarity and coherence, grammatical accuracy, and punctuation. On the other hand, it addressed elements related to the introduction and conclusion, the use of concrete examples, and effective development of arguments and counterarguments.



PROMPT

Can you evaluate the following argumentative text? Is it well written?
The impact of social media on contemporary society

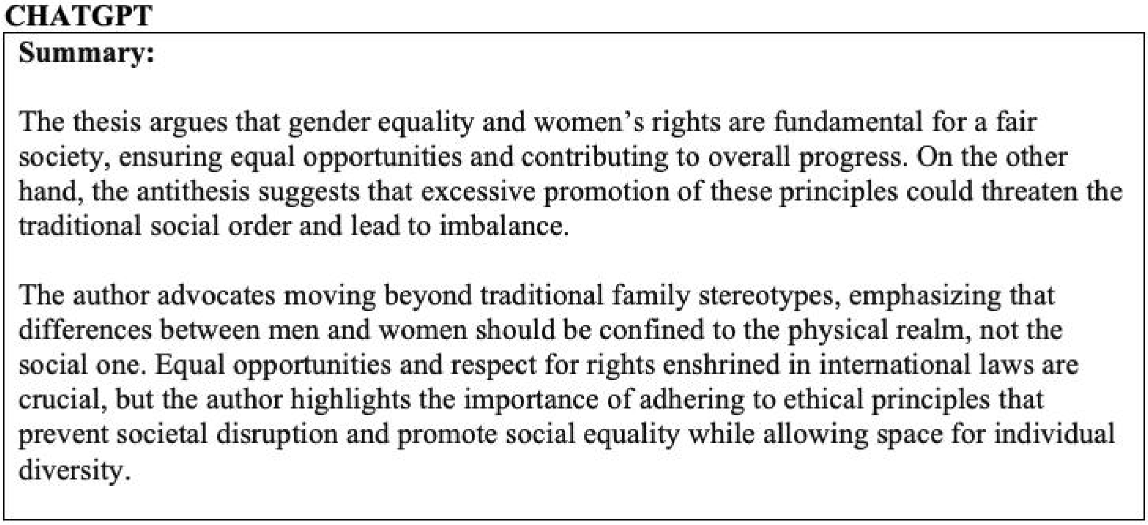
Social media are experiencing a period of exponential growth and global expansion. This widespread diffusion allows them to reach all kinds of people across much of the globe. The growing use of social media has a positive impact on contemporary society, fostering connection, the sharing of ideas, and access to free information.

However, many today argue that the use of social media primarily has negative effects on society, contributing to digital device addiction and social isolation.

[…]

Figure 1

Conversely, prompts lacking clear and detailed instructions significantly hindered ChatGPT's ability to effectively evaluate the quality of an argumentative text. In particular, the prompting strategies adopted by participants had a significant impact on how ChatGPT interpreted and performed the task. For instance, when only the argumentative text is provided without specific instructions regarding the task, ChatGPT often summarizes its content or offers generic observations about its topics, as illustrated in Figure 2 and Figure 3. Another interesting example involved a student who divided their text into paragraphs labelled as "thesis" and "antithesis" without giving task-specific instructions. Therefore, ChatGPT misinterpreted this prompt as a directive to generate an argumentative text. When subsequently prompted to evaluate the text, ChatGPT provided feedback on the argumentative structure; however, its response referred to the text it had generated rather than the original text submitted by the student.



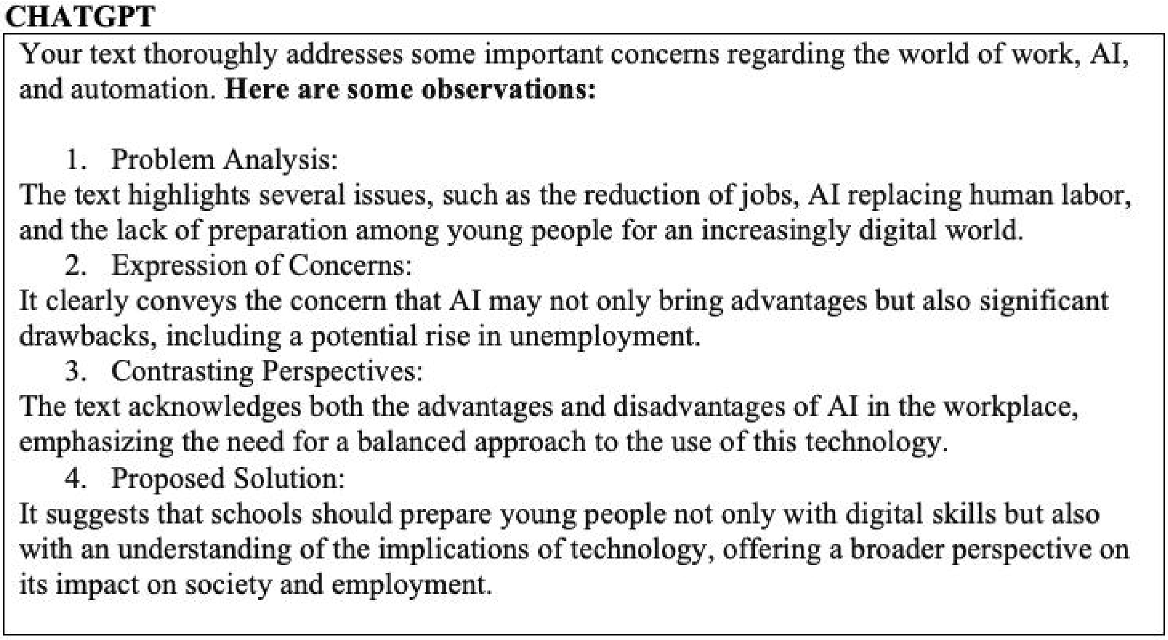
CHATGPT

Summary:

The thesis argues that gender equality and women's rights are fundamental for a fair society, ensuring equal opportunities and contributing to overall progress. On the other hand, the antithesis suggests that excessive promotion of these principles could threaten the traditional social order and lead to imbalance.

The author advocates moving beyond traditional family stereotypes, emphasizing that differences between men and women should be confined to the physical realm, not the social one. Equal opportunities and respect for rights enshrined in international laws are crucial, but the author highlights the importance of adhering to ethical principles that prevent societal disruption and promote social equality while allowing space for individual diversity.

Figure 2



CHATGPT

Your text thoroughly addresses some important concerns regarding the world of work, AI, and automation. **Here are some observations:**

1. Problem Analysis:

The text highlights several issues, such as the reduction of jobs, AI replacing human labor, and the lack of preparation among young people for an increasingly digital world.

2. Expression of Concerns:

It clearly conveys the concern that AI may not only bring advantages but also significant drawbacks, including a potential rise in unemployment.

3. Contrasting Perspectives:

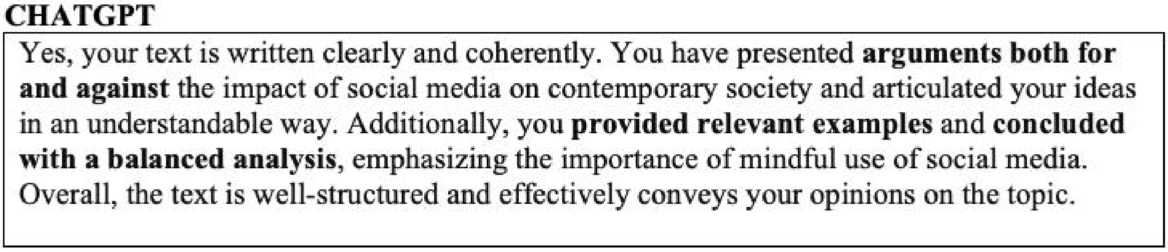
The text acknowledges both the advantages and disadvantages of AI in the workplace, emphasizing the need for a balanced approach to the use of this technology.

4. Proposed Solution:

It suggests that schools should prepare young people not only with digital skills but also with an understanding of the implications of technology, offering a broader perspective on its impact on society and employment.

Figure 3

A similar phenomenon was observed in prompts where the input data, i.e., the text type for which ChatGPT is expected to provide relevant and comprehensive feedback, was not explicitly stated. This issue was particularly evident in prompts where students phrased their instructions as "Is it well written?" without clarifying that the text was an "argumentative text". As a result, ChatGPT encountered difficulties in interpreting the required task. Instead of offering feedback on the argumentative structure, it evaluated the text's quality in terms of content, emphasising the role of evidence and the importance of adopting a balanced perspective when addressing both the positive and negative aspects of the topics, as illustrated in Figure 4.



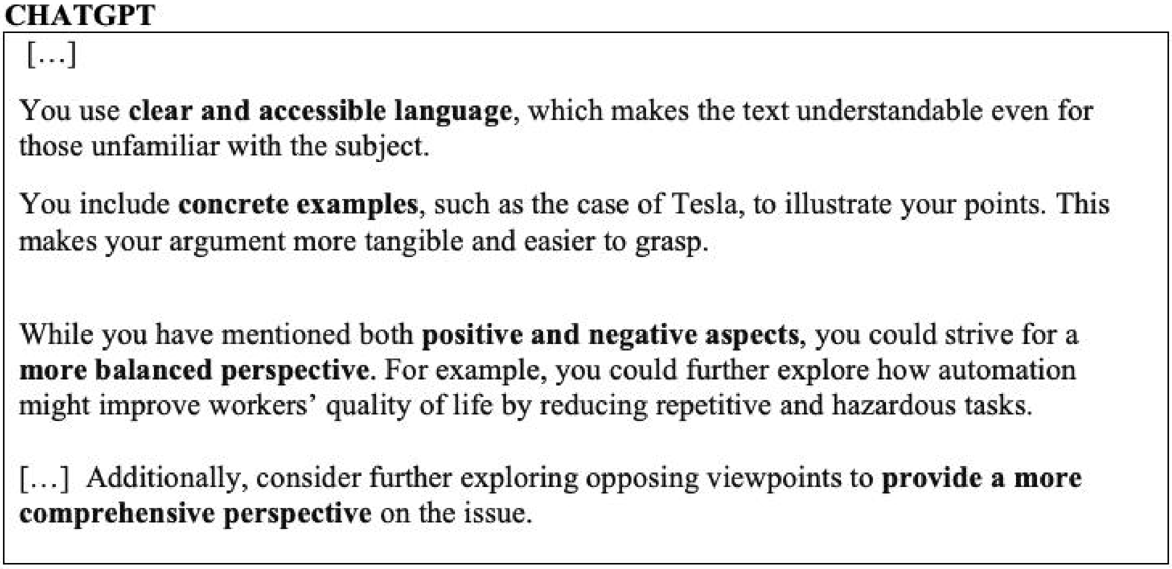
CHATGPT

Yes, your text is written clearly and coherently. You have presented **arguments both for and against** the impact of social media on contemporary society and articulated your ideas in an understandable way. Additionally, you **provided relevant examples** and **concluded with a balanced analysis**, emphasizing the importance of mindful use of social media. Overall, the text is well-structured and effectively conveys your opinions on the topic.

Figure 4

Prompts that lacked relevant contextual information about the task had a negative impact on the accuracy of ChatGPT’s feedback. Although all students included the argumentative text to be evaluated, nearly all omitted details about the instructions for the writing task. In particular, the instructions for the argumentative text on the impact of social media on contemporary society explicitly required students to develop arguments supporting a predefined thesis and refuting the corresponding antithesis. This omission significantly influenced ChatGPT’s feedback in relation to the argumentative structure. Aspects requiring a distinction between thesis and antithesis, such as the clarity of the thesis, the logical development of arguments, their strength, and their relevance to the thesis, were often overlooked. Instead, the feedback focused on achieving a “balanced perspective” or a “balanced presentation of positive and negative points”, as well as on emphasising the use of “concrete examples or case studies” to support the student’s opinions on the topic.

Similar instructions were given for the argumentative text on technology and automation in the future workplace, with the exception that participants were allowed to choose which thesis to support and were required to use connectives appropriately. However, even in this case, the absence of contextual instructions about the task resulted in feedback that did not explicitly reference the thesis and antithesis. Conversely, as shown in Figure 5, the feedback emphasised the use of a “clear and accessible language” and “concrete evidence”, while recommending a more balanced approach to “provide a more comprehensive perspective” on technological innovation. Moreover, ChatGPT’s difficulty to distinguish between the thesis and antithesis may have been stressed by the lack of a clearly defined thesis in the introduction and the thesis statement of the argumentative text. This resulted in the use of generic terms such as “point of view” and “negative/positive aspects” to refer to arguments and counterarguments. Moreover, the feedback lacked an in-depth analysis of register specific to argumentative writing and the effective use of connectives.



CHATGPT

[…]

You use **clear and accessible language**, which makes the text understandable even for those unfamiliar with the subject.

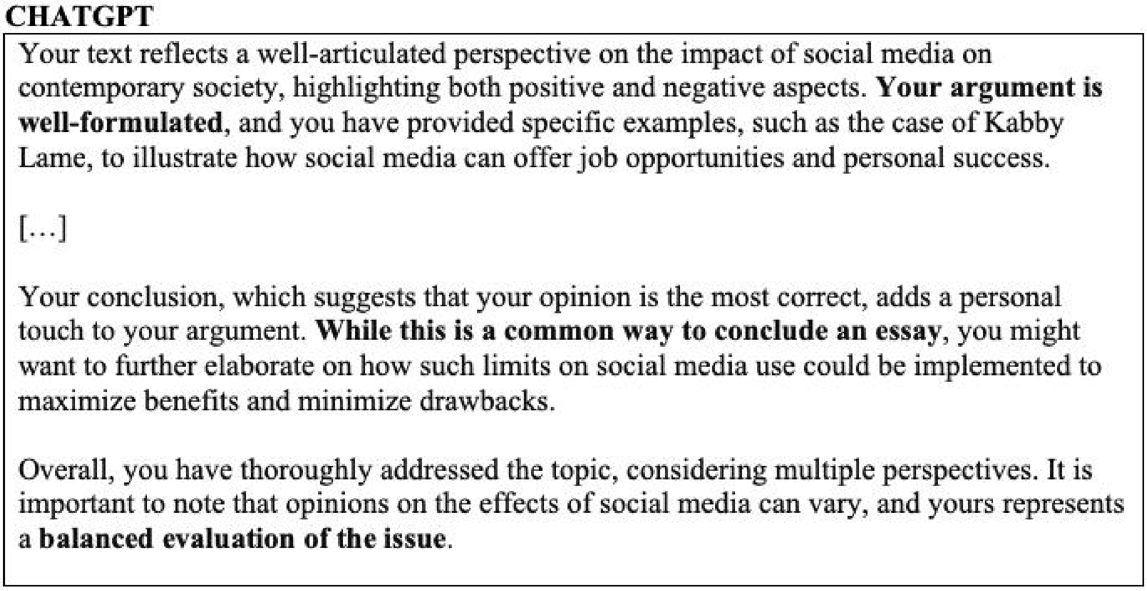
You include **concrete examples**, such as the case of Tesla, to illustrate your points. This makes your argument more tangible and easier to grasp.

While you have mentioned both **positive and negative aspects**, you could strive for a **more balanced perspective.** For example, you could further explore how automation might improve workers' quality of life by reducing repetitive and hazardous tasks.

[…] Additionally, consider further exploring opposing viewpoints to **provide a more comprehensive perspective** on the issue.

Figure 5

A decrease in the relevance and accuracy of ChatGPT's feedback was also observed in response to prompts characterized by imprecise or ambiguous language. This phenomenon was evident where instructions and input data such as "What do you think about this text?" or "Is this text well written?". On the contrary, the absence of precise keywords like "evaluate" or "argumentative text" prevented the LLM from clearly defining assessment criteria specific to argumentative writing. For instance, references to "adequate" or "comprehensive" language appear too generic and vague to be clearly attributed to linguistic elements such as register, grammatical accuracy, or syntactic structure. Moreover, arguments are often described as "well-formulated" opinions that provide a "balanced evaluation of the issue", as in Figure 6. This ambiguity could either focus on aspects related to the argumentative structure, such as their relevance to the thesis or their logical progression, as well as on elements concerning the syntactic clarity and complexity.



CHATGPT

Your text reflects a well-articulated perspective on the impact of social media on contemporary society, highlighting both positive and negative aspects. **Your argument is well-formulated**, and you have provided specific examples, such as the case of Kabby Lame, to illustrate how social media can offer job opportunities and personal success.

[…]

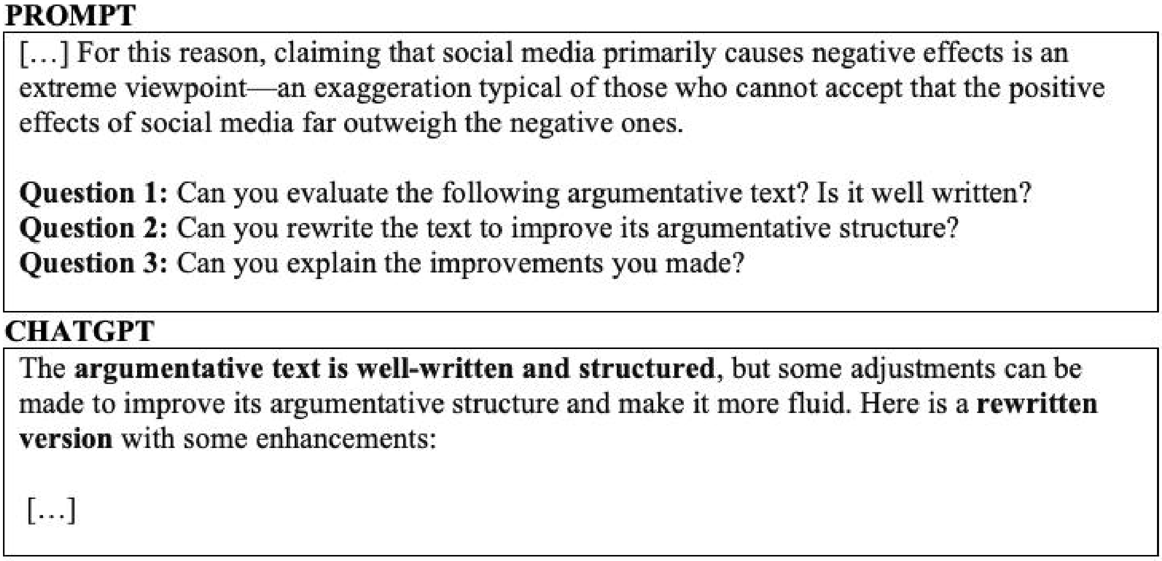
Your conclusion, which suggests that your opinion is the most correct, adds a personal touch to your argument. **While this is a common way to conclude an essay**, you might want to further elaborate on how such limits on social media use could be implemented to maximize benefits and minimize drawbacks.

Overall, you have thoroughly addressed the topic, considering multiple perspectives. It is important to note that opinions on the effects of social media can vary, and yours represents a **balanced evaluation of the issue**.

Figure 6

Furthermore, the use of imprecise language hindered ChatGPT's ability to understand the type of text being evaluated when input data were not clearly defined. As illustrated in Figure 6, ambiguous phrases such as "this text" led the model to interpret the submission as an essay rather than an argumentative text, thereby tailoring its feedback accordingly. Moreover, ChatGPT misinterpreted the task's objectives when instructions were phrased as questions, such as "Is it well-written?" that fail to explicitly inform the LLM about the evaluation of an argumentative text. Consequently, the feedback was often incomplete, overlooking aspects related to argumentative structure and focusing primarily on linguistic elements or general suggestions, such as emphasising the importance of equitably addressing both positive and negative perspectives.

The analysis of prompt structure revealed that including multiple instructions within a single prompt significantly impacts the quality of ChatGPT's feedback. This phenomenon was observed in interactions related to the argumentative text on the impact of social media. The activity was divided into a series of intermediate steps: initially, participants were required to prompt the model for an evaluation of their text, then request a rewritten version to improve its argumentative structure, and finally, ask for an explanation of the changes made. Due to the complexity of this task, which combines three distinct objectives into a single prompt, ChatGPT tended to generalize its responses, resulting in a decrease in the relevance and accuracy of the feedback (Figure 7). Additionally, this complexity often led to outcomes such as a partial revision of the argumentative text or a summary of its main points rather than a targeted evaluation of the argumentative structure.



PROMPT

[…] For this reason, claiming that social media primarily causes negative effects is an extreme viewpoint—an exaggeration typical of those who cannot accept that the positive effects of social media far outweigh the negative ones.

Question 1: Can you evaluate the following argumentative text? Is it well written?
Question 2: Can you rewrite the text to improve its argumentative structure?
Question 3: Can you explain the improvements you made?

CHATGPT

The **argumentative text is well-written and structured**, but some adjustments can be made to improve its argumentative structure and make it more fluid. Here is a **rewritten version** with some enhancements:

[…]

Figure 7

4. Discussion

This study confirms the interdependence between prompting strategies and the quality of LLM outputs (White et al., 2023). Indeed, the quality of responses from ChatGPT largely depends on prompt design, which shapes the interactions between the user and the model and "serves as the foundation for obtaining accurate, relevant, and meaningful responses" (Mondal et al., 2024, p. 1). As a result, effective prompts can significantly enhance the accuracy of feedback, whereas the absence of key components, the use of ambiguous or generic language, or structural variations in prompts can lead to inaccurate or irrelevant answers, ultimately decreasing ChatGPT's performance. In the context of language education, the acquisition of comprehensive knowledge and understanding of prompt engineering is therefore fundamental for developing the soft skills necessary to ensure an effective and responsible implementation of AI-based technologies such as LLMs.

The analysis of the feedback highlights ChatGPT's stronger ability to evaluate the quality of counterarguments and evidence relevant to the thesis. However, there is a noticeable decline in accuracy when addressing aspects that require a clear distinction between the thesis and the antithesis. This issue may be attributed to the probabilistic nature of LLMs, which relies on the textual data they are trained on to generate responses by predicting the statistically most likely sequence of words, without understanding their semantic meaning (Liu et al., 2023; Bender & Koller, 2020). Therefore, the quality of feedback decreases when tasks demand a differentiation of elements supporting the thesis or the antithesis, as ChatGPT is unable to fully grasp or analyse all the semantic nuances and the "intrinsic logic relationships among various pieces of information in the same manner as humans" (L. Wang et al., 2024, p. 9). This limitation emphasises the challenges of relying on LLMs for tasks requiring deeper semantic comprehension.

Regarding linguistic elements, it is observed that despite greater attention to grammatical accuracy, syntactic structure, and textual cohesion, ChatGPT struggles to provide an accurate evaluation of register. This may be attributed both to the task complexity (Wu et al., 2022) and to linguistic variations in the prompt (Korzynski et al., 2023). On one hand, evaluating both linguistic and argumentative aspects within a single prompt can create overlapping demands that the model finds challenging to manage simultaneously and in an effective way. On the other hand, the use of imprecise language in the instructions, which fails to explicitly specify the text type, may have limited the overall accuracy of the feedback. This lack of clarity likely hindered the model's ability to fully understand the distinctive linguistic features of the argumentative text, further impacting the quality of its responses.

Linguistic and structural variations in prompts significantly influenced the quality of feedback in terms of accuracy and relevance to the task and objectives. Lack of clarity in the instructions and input data, coupled with the absence of relevant contextual information, contributed to generic or ambiguous responses regarding both the argumentative structure and the formal aspects of the texts. Because of this lack of precision ChatGPT occasionally misinterprets the task or the type of text to evaluate (see Korzynski et al., 2023; Lo, 2023). Conversely, the inclusion of keywords such as "evaluate" and "argumentative text" can direct the LLM towards more relevant and accurate responses aligned with the task requirements (see O'Connor & Andreas, 2021). Moreover, prompt structure containing multiple instructions or tasks often result in a rewriting of the argumentative text rather than providing comprehensive feedback. This phenomenon can be attributed to LLM being "inherently limited for complex problems" (Wu et al., 2023, p. 1) that require multi-step reasoning tasks to be solved.

Our findings indicate that developing prompt engineering skills is a critical component of effective human-AI mediation. To this end, the ability to design prompts that elicit responses aligned with user expectations and desired outcomes requires the acquisition of specific transversal skills, which are essential for the effective and responsible integration of AI-based tools into language education. First, one must possess a basic understanding of the models' internal mechanisms to enhance their knowledge about the advantages and inherent limitations of mainstream LLMs (Gao et al., 2021; Cain, 2024). For instance, prompt sensitivity can significantly impact on the accuracy and reliability of LLM outputs (see Zhao et al., 2021; Lu et al., 2022), as even slight linguistic and structural variations in prompts, such as differences in word choices or information provided, can yield substantially different feedback on the argumentative texts. Therefore, by examining the working principles that characterise LLMs, learners and educators can critically evaluate their performance while making informed decisions on how to integrate these tools effectively into learning processes.

Furthermore, understanding how LLMs function plays a fundamental role in developing the digital competencies and critical thinking skills necessary to craft and refine prompts, thereby leveraging these models' potential across different tasks and operations (Cain, 2024). In particular, the process of prompt design requires users to provide precise and detailed instructions regarding the task and desired outcomes, along with all relevant information needed to guide ChatGPT throughout its performance (Korzynski et al., 2023). For instance, it is essential to describe the learning activity in a precise, directive, and explicit manner, using phrases such as "evaluate this argumentative text", to ensure that ChatGPT accurately recognizes learners' intended objectives, i.e., assessing their argumentative skills, and avoids misinterpreting the type of text to be evaluated. Moreover, as Su et al. (2023) point out, including comprehensive contextual information about the writing task, such as an evaluation rubric, can help prevent ChatGPT from generating inaccurate, overly generic, or incomplete feedback. In this case, our findings emphasise the importance of specifying the instructions students received for the writing task, such as the distinction between the thesis and antithesis, to further enhance the relevance and accuracy of the model's responses.

Finally, the ability to critically evaluate the effectiveness of a prompt in shaping the quality of the model's responses is a fundamental soft skill for ensuring the reliability and validity of interactions with ChatGPT (Lo, 2023; Cain, 2024). LLMs often experience a decline in the quality of interactions when generating long bodies of text. This decline, known as exposure bias, occurs when errors and inaccuracies from earlier outputs accumulate throughout the interaction (Tan et al., 2021). Furthermore, limitations in interpreting students' specific requirements due to lack of sufficient information about the task can generate inaccurate answers, which further diminish the quality of feedback (Anderson et al., 2023). That is the case of prompts that do not provide instructions about the evaluation of the argumentative text or include vague questions such as "Is it well written?" and "What do you think about this text?", leading ChatGPT to summarize or comment the content of the text, of even generate its own text. Hence, learners and educators must develop critical thinking skills to evaluate interactions with LLMs by identifying and understanding potential causes behind errors or misunderstandings to iteratively adapt and refine the prompt to align ChatGPT's feedback with expected outcomes (Anderson et al., 2023; Eager & Brunton, 2023).

5. Conclusion

Questions, doubts, concerns, and even fears surrounding the evolving relationship with AI are increasingly central to a significant body of research across various scientific disciplines. Similarly, the anticipation for the results of ongoing studies, aimed at deepening the discourse on the use of AI in educational contexts, continues to grow alongside its expanding presence in everyday life. This contribution has primarily aimed to present the findings of an empirical, evidence-based study within the field of educational linguistics, while also offering a reflection on how the relationship with AI can be shaped through the development of specific transversal skills. From the analysis of variations in interactions between students and ChatGPT during learning activities designed to enhance argumentative writing skills, prompt engineering emerges as a key educational objective necessary to facilitate effective communication between humans and AI-based technologies. Prompt design therefore becomes a direct manifestation of mediation that entails the ability to understand how AI functions, interpret effective modes of interaction with it, and foster a collaborative relationship in which AI's role is defined by an integrative approach. Such an approach enables humans to communicate effectively with AI and comprehend its outputs, ultimately positioning AI as a co-tutor in language learning and teaching processes.

References

Anderson, N., McGowan, A., Galway, L., Hanna, P., Collins, M., & Cutting, D. (2023). Implementing generative AI and large language models in education. In *Proceedings of the 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS 2023)* (pp. 1-6). IEEE. https://doi.org/10.1109/isas60782.2023.10391517

Balboni, P. E. (2011). *Conoscenza, verità, etica nell'educazione linguistica*. Guerra.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185-5198). Association for Computational Linguistics. 10.18653/v1/2020.acl-main.463

Bozkurt, A. (2023). Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift. *Asian Journal of Distance Education, 18*(1), 198–204. https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/718

Cain, W. (2024). Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends, 68*, 47–57. https://doi.org/10.1007/s11528-023-00896-0

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2024). Unleashing the potential of prompt engineering in large language models: A comprehensive review. *ArXiv*, *abs/2310.14735v5*. https://doi.org/10.48550/arXiv.2310.14735

Chini, M., & Bosisio, C. (2014). *Fondamenti di glottodidattica*. Carocci.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing.

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Busheck, D. (2022). How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models. In *GenAICHI: Generative AI and Computer Human Interaction, Workshop (CHI'22)* (pp. 1-7). Association for Computing Machinery.

De Mauro, T., & Ferreri, S. (2005). Glottodidattica come linguistica educativa. In Voghera, M., Basile, G., & Guerriero, A. R. (Eds.), *E.LI.C.A. Educazione linguistica e conoscenze per l'accesso* (pp. 17–28). Guerra.

Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-augmented teaching and learning practice. *Journal of University Teaching and Learning Practice, 20*(5), 1–19. https://doi.org/10.53761/1.20.5.02

Gao, T., Fisch, A., & Chen, D. (2021). Making pre-trained language models better few-shot learners. In C. Zong, F. Xia, W. Li & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long papers)* (pp. 3816–3830). Association for Computational Linguistics. 10.18653/v1/2021.acl-long.295

Kazemitabaar, M., Hou, X., Henley, A., Ericson, B. J., Weintrop, D., & Grossman, T. (2023). How novices use LLM-based code generators to solve CS1 coding tasks in a self-paced learning environment. In A. Mühling & I. Jormanainen (Eds.), *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research* (pp. 1-12). Association for Computing Machinery. https://doi.org/10.1145/3631802.3631806

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education, 6*, 1–14. https://doi.org/10.1016/j.caeai.2024.100225

Korzynski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review, 11*(3), 25–38. https://doi.org/10.15678/EBER.2023.110302

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies, 29*(9), 11483–11515. https://doi.org/10.1007/s10639-023-12249-8

Li, H., Leung, J., & Shen, Z. (2024). Towards goal-oriented prompt engineering for large language models: A survey. *ArXiv*, *abs/2401.14043v3*. https://doi.org/10.48550/arXiv.2401.14043

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*, 1–18. https://doi.org/10.3390/e23010018

Liu, L. (2023). Analyzing the text contents produced by ChatGPT: Prompts, feature-components in responses, and a predictive model. *Journal of Educational Technology Development and Exchange, 16*(1), 49–70. https://doi.org/10.18785/jetde.1601.03

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. In *Transactions of the Association for Computational Linguistics* (pp. 157–173). MIT Press. 10.1162/tacl_a_00638

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computer Surveys, 55*(9), 1–35. https://doi.org/10.1145/3560815

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship, 49*(4), 1–3. https://doi.org/10.1016/j.acalib.2023.102720

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In S. Muresan, P. Nakov & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8086–8098). Association for Computational Linguistics. 10.18653/v1/2022.acl-long.556

Mondal, S., Bappon, S. D., & Roy, C. K. (2024). Enhancing user interaction in ChatGPT: Characterizing and consolidating multiple prompts for issue resolution. In D. Spinellis, A. Bacchelli & E. Constantinou (Eds.), *Proceedings of the 21st International Conference on Mining Software Repositories (MSR '24)* (pp. 222-226). Association for Computing Machinery. https://doi.org/10.1145/3643991.3645085

Nurminen, M., & Papula, N. (2018). Gist MT users: A snapshot of the use and users of one online MT tool. In J. A. Pérez-Ortiz, F. Sánchez-Martínez, M. Esplà-Gomis, M. Popovic, C. Rico, A. Martins, J. Van den Bogaert & M. L. Forcada (Eds.), *Proceedings of the 21st Annual Conference of the European Association for Machine Translation* (pp. 199–208). European Association for Machine Translation.

O'Connor, J., & Andreas, J. (2021). What context features can transform language models use? In C. Zong, F. Xia, W. Li & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 851–864). Association for Computational Linguistics. 10.18653/v1/2021.acl-long.70

Ranieri, M., Cuomo, S., & Biagini, G. (2023). *Scuola e intelligenza artificiale: Percorsi di alfabetizzazione critica*. Carrocci.

Sawalha, G., Taj, I., & Shoufan, A. (2024). Analyzing student prompt and their effect on ChatGPT's performance. *Cogent Education, 11*(1), 1–20. https://doi.org/10.1080/2331186X.2024.2397200

Sheese, B., Liffiton, M., Savelka, J., & Denny, P. (2024). Patterns of student help-seeking when using a large language model-powered programming assistant. In N. Herbert & C. Seton (EDS.), *Proceedings of the 26th Australasian Computing Education Conference (ACE '24)* (pp. 49-57). Association for Computing Machinery. https://doi.org/10.1145/3636243.3636249

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, 1–13. https://doi.org/10.1016/j.asw.2023.100752

Tan, B., Yang, Z., Al-Shedivat, M., Xing, E. P., Hu, Z. (2021). Progressive generation of long text with pretrained language models. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, S. Bethard, R. Cotterell, T. Chakraborty & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4313-4324). Association for Computational Linguistics. 10.18653/v1/2021.naacl-main.341

Theophilou, E., Koyutürk, C., Yavari, M., Bursic, S., Donabauer, G., Telari, A., Testa, A., Boiano, R., Hernandez-Leo, D., Ruskov, M., Taibi, D., Gabbiadini, A., & Ognibene, D. (2023). Learning to prompt in the classroom to understand AI limits: A pilot study. In R. Basili, D. Lembo, C. Limongelli & A. Orlandini (Eds.), *Proceedings of the 22nd International Conference of the Italian Association for Artificial Intelligence* (pp. 481-496). Springer. *https://dx.doi.org/10.1007/978-3-031-47546-7_33*

Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education, 21*(1), 1–29. https://doi.org/10.1186/s41239-024-00448-3

Wang, M., Wang, M., Xu, X., Yang, L., Cai, D., & Yin, M. (2024). Unleashing ChatGPT's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering. *IEEE Transactions on Learning Technologies, 17*, 629–641. https://doi.org/10.1109/TLT.2023.3324714

Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity, 51*, 1–14. https://doi.org/10.1016/j.tsc.2023.101440

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., et al. (2023). A prompt pattern catalogue to enhance prompt engineering with ChatGPT. *ArXiv, abs/2302.11382*. https://doi.org/10.48550/arXiv.2302.11382

Woo, D. J., Guo, K., & Susanto, H. (2023). Case of EFL secondary students' prompt engineering pathways to complete a writing task with ChatGPT. *ArXiv, abs/2307.05493*. https://doi.org/10.48550/arXiv.2307.05493

Wu, T., Terry, M., & Cai, C. J. (2022). AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, A. Drucker, J. Williamson & K. Yatani (Eds.), *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)* (pp. 1-22). Association for Computing Machinery. https://doi.org/10.1145/3491102.3517582

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson & M. L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)* (pp. 1-22). Association for Computing Machinery. https://doi.org/10.1145/3544548.3581388

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning (ICML '21)* (pp. 12697–12706). PMLR.