IADIS International Journal on WWW/Internet
Vol. 22, No. 1, pp. 44-58
ISSN: 1645-7641

THE ADDED VALUE OF ACADEMIC WRITING INSTRUCTION IN THE AGE OF LARGE LANGUAGE MODELS: A CRITICAL ANALYSIS

Andrea Molinari[1] and Elena Molinari[2]
[1]*Dept. of Industrial Engineering University of Trento, Trento, 38122 Italy*
[2]*TUI Musement, Milan, 20139 Italy*

ABSTRACT

This study investigates the integration of Large Language Models (LLMs) into academic writing education, focusing on a case study involving bachelor's degree students. In an era where artificial intelligence (AI) significantly influences educational methodologies, our research explores the effective utilization of LLMs to enhance students' academic writing skills. The study spans three academic years and involves undergraduates enrolled in an academic writing course. The primary assignment required students to produce a minimum three-page paper, including a literature review on a self-selected topic with high-quality references. We employed an experimental approach by comparing each of the selected students' papers (n=24) with a parallel version generated by LLMs (OpenAI ChatGPT-4 and Jenni), starting from the same title and abstract of the student's papers. A blind review was conducted to evaluate the quality of the student-written and LLM-generated papers using identical criteria. Preliminary findings indicate that in the specific context of literature review and academic writing, (a) LLM-generated papers scored higher across most evaluation areas compared to their human-written counterparts in a twentieth of the time usually allowed to a student, and (b) LLM-generated papers were utterly indistinguishable from those written by students during the review process. The study also addresses challenges such as maintaining academic integrity and fostering critical thinking skills in an AI-assisted learning environment. Additionally, side effects on the role of educators in this context are considered, particularly in disciplines heavily impacted by generative AI regarding educational methodologies, content delivery, and evaluation processes.

KEYWORDS

Artificial Intelligence, Teaching, Academic Writing

1. INTRODUCTION

Academic writing is a university course typically offered in degree programs where high-quality written work is emphasized according to academic standards, particularly in programs where a thesis remains a crucial component of the final evaluation, requiring students to produce extensive written documents. Academic writing courses are often provided either by the university's language centers or through traditional courses. These courses cover the fundamental principles of academic language, the main structure of an article for both journals and conferences, the use of English, specifically academic English, as well as technical details such as formatting, bibliography structure, and the procedures for submitting articles to journals and conferences within primary academic contexts. Although this course often stands apart from the primary field of study, whether in the humanities or scientific disciplines, it receives positive feedback from students. They recognize the gap between their current writing skills and the expectations of academic writing as required by their instructors, starting from the bachelor's level and extending to master's theses and the doctoral research environment.

Two major approaches to structuring these programs exist within university academic writing courses. On the one hand, the approach interprets the academic writing course as heavily focused on the English language. These courses essentially center on the use of English at a much higher level than what non-native speakers typically possess. They are often taught by native English-speaking instructors who emphasize the use of academic English and reinforce the general use of the English language.

The second type of academic writing course, the article's focus, is oriented towards providing the technical foundations of academic writing. It is well-known that various models and techniques exist for crafting an engaging academic title, writing descriptive or research abstracts depending on the document type, and structuring a paper's introduction using models such as the C.A.R.S. model. Additionally, these courses cover rules for composing tables and figures, structuring paragraphs and paragraph titles, writing conclusions, and formatting bibliographies according to different standards. This type of course does not overlook the introduction of academic English. However, it primarily focuses on the technicalities necessary for writing a paper that possesses the characteristics of a scientific work, tailored to a specific audience accustomed to certain linguistic and conceptual constructs that align with scientific discourse.

The aim of this study is to analyze the impact of Large Language Models (LLMs) and Generative Pretrained Transformers (GPT) on the generation of academic essays and, consequently, on the structure and rationale of an "Academic Writing" course. We present the results of an experiment informed by specific insights into the teaching and evaluation methodologies of a traditional academic writing course within scientific and economic faculties, as well as the influence of artificial intelligence tools on these methodologies. Integrating advanced tools such as LLMs presents opportunities and challenges in the rapidly evolving educational technology landscape. Over the past two decades, research on implementing technology in the classroom has yielded mixed results. Some technology-based interventions have demonstrated significant positive effects (Englert et al., 2007; Warren et al., 2008), while others have reported minimal or adverse effects (Rowley and Meyer, 2003; Goldenberg et al., 2011). Several studies have focused on AI-assisted language learning tools to enhance the writing skills of English language learners (Little et al., 2018). This immediate application of these technologies is due to their superior performance, particularly in the English language, in

generating text that is often richer and of higher quality than that produced by average students (Liu et al., 2023; Wu et al., 2021; Fitria, 2021).

Many researchers have focused on the impact of AI on improving English language learning outcomes (Sun et al., 2021) and in technical disciplines (Huang et al., 2023). For instance, Huang et al. (2023) conducted a study comparing students' academic performance and engagement levels in an AI-enhanced course versus a traditional one. Results indicated that those in the AI course achieved higher academic success and engaged more actively in their learning tasks than their peers in the non-AI course. Other studies follow a similar line by examining the role of AI-assisted language learning tools in enhancing the writing skills of English language learners (Liu et al., 2023; Seufert et al., 2021; Hsiao and Chang, 2023; Yan, 2023). These studies explore how AI-supported language learning influences the writing abilities of English as a Foreign Language (EFL) learners, finding a substantial positive effect. Yan (2023) assessed the impact of ChatGPT, an AI-based language learning tool, on EFL learners' writing proficiency and observed notable enhancements in their writing due to AI-assisted learning. In today's digital era, characterized by the widespread accessibility of technology and online resources, learners can effortlessly practice their language skills, mainly writing, at their convenience and from any location (Yan, 2023). This practice is increasingly supported by sophisticated AI-enabled computer and mobile applications, which provide interactive and tailored resources for refining writing skills and improving motivation (Jiang, 2022; Meunier et al., 2022; Yan, 2023).

The sources highlight the potential benefits and challenges of integrating AI-powered writing tools in academic writing instruction. They provide insights into the positive impact of AI tools on learners' writing performance, such as improving efficiency and enhancing argumentative writing tasks. However, they also caution against overreliance on AI tools and stress the need for instruction and guidance to ensure students' skill development and prevent potential negative effects on their writing skills over time. Moreover, the sources emphasize the importance of a balanced approach to integrating AI tools in academic writing instruction. This approach involves incorporating AI tools alongside traditional teaching practices, such as discussions, drafting, revising, and providing formative feedback. By incorporating AI tools into these existing practices, instructors can support students in developing linked thinking and writing skills while leveraging the benefits of AI assistance in areas such as generating alternative versions of the text and assessing the AI output for relevance, accuracy, completeness, bias, timeliness, and writing quality (Stanford University, 2023). The sources also highlight the need for students to learn how to effectively use AI tools in academic writing tasks, including setting directions and prompts and assessing the AI output (Song and Song, 2023). The sources discuss integrating AI-powered writing tools in academic writing instruction. They acknowledge the positive impact these tools can have on improving learners' efficiency and performance in writing tasks. They also recognize the need for instruction and guidance to ensure responsible and effective use of AI tools and mitigate potential adverse effects on students' writing skills.

Integrating AI tools in teaching academic writing can enhance students' writing performance and efficiency. Instructors are expected to provide explicit instruction on the appropriate application of AI tools in academic writing tasks and continuously monitor their impact on students' writing skills (Chong, 2021). This case study highlights the importance of balancing AI assistance and students' independent thinking and engagement in the writing process. The findings are in line with other studies that suggest an integration of AI tools, specifically the use of the equipped AI-powered writing tool, to positively affect English second postgraduate

students' academic writing performance, thus contributing to their overall improvement in English writing proficiency (Song and Song, 2023). The study also emphasizes the importance of effectively addressing students' affective characteristics and attitudes toward technology to engage them in learning. Furthermore, instructors must provide ongoing support and instruction on the appropriate use of AI tools in academic writing tasks, ensuring that students understand how to leverage the AI tool effectively while critically evaluating its outputs and maintaining their writing voice.

However, it is essential to note that a complete reliance on AI for writing tasks may lead to a reduction in accuracy and in writing autonomous abilities. Therefore, instructors must balance AI assistance and students' critical thinking and engagement in the writing process. The studies mentioned emphasize the need for a structured learning environment incorporating formative feedback, drafts, and revisions to support students' writing development. By adopting these essential teaching practices, instructors can effectively navigate the integration of AI tools into their academic writing courses, promoting students' critical thinking skills while harnessing the benefits of AI technology (Stanford University, 2023) and enhancing their pedagogical strategies. Students must learn to set directions, prompt the AI effectively, assess the AI outputs, and make necessary revisions and adjustments to improve their writing.

This paper builds upon previously established research lines by addressing the research questions through a qualitative analysis of academic texts produced by bachelor's degree students for an academic writing course. It compares these student-generated texts with those produced by AI tools, using identical titles and abstracts. This study represents nearly a year of comparison between human-generated and AI-generated academic texts, aiming to explore the effectiveness of Large Language Models (LLMs) in a practical educational setting. Focusing on a case study involving undergraduate students enrolled in an academic writing course, we critically examine the role of LLMs, specifically OpenAI's ChatGPT-4 and Jenni, in enhancing academic writing skills. The research is situated in a contemporary educational environment where artificial intelligence (AI) is increasingly reshaping teaching and learning methodologies.

The study spans three academic years (2019-2020, 2020-2021, 2021-2022), providing a substantial temporal framework to assess the impact of LLMs on the academic writing capabilities of these undergraduates. Central to this research is an experimental approach where the outputs of top-performing students (n=24) are juxtaposed with those generated by LLMs. This comparison is grounded in a blind review process using uniform evaluation criteria to assess the quality and distinguishability of student-written and LLM-generated academic papers, particularly in literature reviews.

The paper is structured as follows: Chapter 2 details the methods used to collect and analyze academic texts produced by students and those generated by AI tools. Chapter 3 summarizes the most relevant results, considering the qualitative nature of this research. In Chapter 4, we present preliminary considerations on the collected data. Finally, Chapter 5 concludes and suggests directions for future research.

2. METHODS

The study included 24 students from different European countries enrolled in three courses related to academic writing: scientific (Computer Science/Engineering) and economic (Economy and Management). The bachelor's degree programs were offered at two different

Italian universities. The course encompasses instruction in the techniques for composing the main components of a scientific paper: abstract, introduction, research methods, results, discussion, and conclusions. The course introduces construction methods for each component of the paper without delving specifically into linguistic aspects, as it is a "technical" course rather than an English language course. For instance, the CARS model (Swales, 1990; Sánta, 2015). is employed for writing the introduction, dividing the writing process into various logical steps that culminate in the final product.

The final exam involved creating a literature review on a topic the student chose, composed of at least three pages according to a standard conference/journal template, with no figures, tables or any other graphical element. The evaluation focused not so much on the quality of the chosen scientific content but instead on the alignment between the academic writing techniques taught in the course and the work produced by the student. The students' final work was meticulously checked using the university's anti-plagiarism tools, which, until November 2022, were very effective in verifying the correctness and originality of the paper and the absence of plagiarism.

The evaluation criteria for the submitted paper are represented in Table 1, with the weights expressed in thirtieths. The student passed the exam with a score of 18/30. The evaluation was focused on two main elements: the techniques used to write the various sections (title, abstract, introduction, conclusions and references) and the use of academic language in the paper (vocabulary, grammar, cohesion, and coherence of text). The sum of weights is 31 to provide a possibility of the "cum Laude" score.

Table 1. Evaluation criteria for the papers

Title	Abstract	Intro	Conclusions	References	Vocabulary	Grammar	Cohesion and coherence	Total
1	3	8	3	2	4	4	6	31

The final results, obtained from the sum of the scores, are rounded up to the nearest whole number, and the grade is expressed in multiples of 0.5.

Because of the nature of the requested paper (a literature review) and in order to allow students to exert an effort proportional to the number of credits acquired (3 or 6 ECTS), sections like "methods", "results" and "discussions" were excluded from the final paper. Papers for the experiment were collected in the past three years before the disruptive advent of LLM on the market to prevent students from modifying their writing and to have reasonable assurance that the papers were not produced using ChatGPT, Jenni, or similar tools. The papers (n=24) have been selected in order to represent three classes of results, with eight papers for each class: a) fair (21 <= grade <=24), b) good (24 <= grade <=27), c) very good (28 <= grade <=31). Grades below 21 have not been considered, as these results are usually represented by students whose final essays are often of low quality and easily improvable. Many students of this type choose to retake the exam to enhance their score, so most of the type of paper is an "approach" to the exam rather than their final essay, and for this reason, they have not been included in the analysis.

For each grading category within the class, papers are distributed equally among the four different scores in the interval. This approach ensured a comprehensive analysis encompassing the full spectrum of student performance, from those achieving the least to those achieving the most and those representing the typical or average level of achievement within each grade category. An initial group of 80 students was selected and divided into the three mentioned

classes. Students' papers have been associated with the respective scores in the various sections assigned by the teacher. Authors' names were removed from their articles, and from these 80 anonymized articles, the three classes were randomly composed to prevent any association between a student and their paper.

After the composition of the sample of papers, we submitted the paper's title and abstract to the AI tools. Through a series of interactions with the AI tools using a standard set of prompts ("extend and rephrase the previous text", "provide academic references to the following text", "improve the academic tone of the following text"), we have generated a new paper of approximately the same length and content as the original one. Twenty-four new papers have been produced and blindly evaluated with the same criteria for the students' papers. Regarding time dedicated to recreating the original paper using LLMs, one-twentieth of the formally expected time was allocated for such an article as a maximum effort to recreate the text. The ECTS mechanisms stipulate a student effort of 25 hours for each ECTS, so the anticipated workload corresponds to approximately 100 hours. Deducting the hours spent on lectures from this workload, a student has roughly sixty hours left to dedicate to composing the paper. In creating the new paper using artificial intelligence tools, about three hours were allocated for generating each article's title, abstract, introduction, conclusions, and references, thereby limiting the number of interactions for subsequent refinements within this time frame. The result produced by the artificial intelligence engines was then evaluated, and we present the results of the comparison between the grade obtained by the student and the grade that would have been obtained had the student used an AI engine and one-twentieth of the anticipated time.

3. RESULTS

Table 2 is a summary of the most relevant data we collected. Data is presented for each student cohort concerning the various components of the required academic article, encompassing both the technical and linguistic aspects. A final column is included, showcasing the average outcomes achieved by the students in their articles alongside the results derived from the evaluation of articles generated by AI. Cells, where the AI yielded results inferior to student-generated texts, are highlighted with color.

It is observed that the texts generated by AI are of inferior quality compared to those produced by the students in the title section across all three student categories. The inferiority of the AI-generated texts is relatively consistent (albeit marginally) in the texts produced compared to the excellent students. At the same time, the AI generates superior texts in virtually all other sections for all other student categories.

The most relevant difference between human-generated and AI-generated texts in favor of humans is in the quality and appropriateness of the title and the references. Here, students' final scores are superior to or similar to AI-generated text. The techniques and tips taught in academic writing contexts appear to be sufficient to maintain a minimal difference. Due to the uniqueness of the results compared to the other sections, we manually analyzed these two sections of the papers. For titles, this is probably a matter of the evaluators' taste and the time devoted to the AI-generated text. Three hours were devoted to generating AI texts, and a relatively minimal portion of the time was devoted to the composition of the title, evidently in proportion to its impact on both the grade and the structure of the scientific essay. Conversely, students, recognizing the title's significance as a tool for the academic communication and promotion of their work, almost certainly allocated more attention to it. This heightened focus could account for the markedly superior outcome achieved by human effort compared to AI.

Table 2. Summary of results of student-/AI-generated texts

		C3 - Very good	C2 - Good	C1 - Fair
Title	Human	0,9	0,9	0,8
	AI	0,7	0,7	0,8
Abstract	Human	3	1,8	1,8
	AI	2,8	2,8	2,6
Intro	Human	7,4	6,3	4,1
	AI	7,9	7,9	7,8
Conclusions	Human	2,8	2,8	2,6
	AI	2,9	2,6	2,7
References	Human	2	1,7	1,3
	AI	1,9	1,7	1,6
Vocabulary	Human	4	3,6	3,3
	AI	3,8	3,8	3,8
Grammar	Human	3,9	3,4	3,3
	AI	3,8	3,9	3,8
Cohesion coherence	Human	5,6	5,3	5,3
	AI	5,9	5,8	5,8
Class Students Avg Score		29,8	26	22,5
Class AI avg score		29,9	29,4	29

Concerning references, those generated by students were generally more appropriate and connected to the sentences they referred to. This aspect, as expected, mainly happened in the top-scored papers, where evidently, the quality of the student's research and writing and the consequent overall quality of the article were closely linked to the quality of the bibliographic references analyzed and cited. The citation style was without mistakes from AI, and the references generated by AI were appropriate. However, from a scientific point of view, students' generated references (in top papers) were more connected to the context and more interesting for the reader. Due to the sample size constraints, in Table 3, we present a few simple statistical elements, with a preference for a more detailed and qualitative interpretation. A more substantial statistical significance is deferred to subsequent, more in-depth analyses.

Table 3. Simple stats on final results

	AI Avg improvement	AI Avg % improvement	AI Min Improvement	AI Max Improvement
Class 3 - Very good	**0,1**	**0,65%**	**-6,45%**	**10,71%**
Class 2 – Good	**3,4**	**13,08%**	**7,41%**	**20,00%**
Class 1 – Fair	**6,5**	**29,18%**	**20,83%**	**42,86%**

Table 4 presents a detailed series of results for each student, highlighting the outcomes achieved with the evaluation of the original text and the outcomes achieved with the text generated by artificial intelligence, starting from the same title and abstract provided by the student. The table also shows the differential in points between the two evaluations, with students who performed better than the artificial intelligence highlighted in red, and the detailed points relative to the final score gained by the students through the AI-generated text highlighted

in green. In the second column, next to each student's name, is the time the authors used artificial intelligence to regenerate the text, along with the methods described for repeating prompts. Time for generation comes from the interaction and exchange of chunks of text from ChatGPT™ to Jenni™, and vice versa.

Table 4. Detailed results by student, with time used by authors to generate the text with AI

		Student's Final Score	Student's Final Score with AI	Difference in points	
Stud reg.nr.	**AI-TIME(min.)**	Stud. Score	AI score		
Stud1	70	31	29	-2	C3
Stud2	58	31	30	-1	C3
Stud3	93	31	29	-2	C3
Stud4	140	31	31	0	C3
Stud5	55	29	29	0	C3
Stud6	81	29	30	1	C3
Stud7	166	28	31	3	C3
Stud8	74	28	30	2	C3
Class 3 - Very good	Avg: 92,13	29,8	29,9	Avg: 0,13	1
Stud9	69	27	29	2	C2
Stud10	81	27	30	3	C2
Stud11	178	26	29	3	C2
Stud12	59	25	30	5	C2
Stud13	67	26	29	3	C2
Stud14	144	26	28	2	C2
Stud15	76	26	30	4	C2
Stud16	81	25	30	5	C2
Class 2 - Good	Avg: 94,38	26,0	29,4	Avg: 3,38	27
Stud17	70	24	29	5	C1
Stud18	82	24	29	5	C1
Stud19	95	23	30	7	C1
Stud20	172	23	29	6	C1
Stud21	57	22	29	7	C1
Stud22	74	22	28	6	C1
Stud23	66	21	30	9	C1
Stud24	81	21	28	7	C1
Class 1 - Fair	Avg: 87,13	22,5	29,0	Avg: 6,50	52

4. DISCUSSION

The research questions addressed in this study were predominantly qualitative, bordering on foundational inquiries about the essence of teaching specific disciplines and evaluating outcomes using traditional methodologies. A primary question centers not exclusively on the performance of AI engines in generating academic texts but on the level of text creation required for bachelor's level students unaided by additional tools. The quality of the outcomes produced by generative pre-trained transformers (GPTs) models, trained explicitly for academic writing, is becoming increasingly evident and impressive. The main goal, however, was to identify the areas where LLMs provide significant improvements and understand and enhance educational processes.

Here, we can readily corroborate the findings presented in Table 2, noting a significant increase in improvement as the quality of the student-generated papers decreases. As hypothesized, the students with the poorest performance would benefit tremendously from using AI-generated texts in terms of their final grade, increasing their scores by approximately 30%, with peaks nearing 43%. This essentially amounts to an advancement of two grade categories.

For this, it is primarily evident that students with lower performance scores benefit the most from using artificial intelligence tools. This is predominantly true for the lengthier sections of an academic paper, such as the introduction, abstract, and conclusions. The longer the text, the greater the advantage compared to what a student produces natively. As highlighted by previous studies, the linguistic aspect is a significant area of improvement. The term 'linguistic area' does not solely refer to the use of the English language but also encompasses vocabulary, grammatical structures, and coherence and consistency in academic language. Generative models, also trained using scientific articles, produce results in this particular language and mode of expressing scientific thought that is decidedly superior to what a student can learn in a traditional academic course.

Another intriguing finding from our sample pertains to the distinct net improvement across the three categories. The sole instances of performance deterioration, relative to the grades assigned by the instructor, involve the three top students. We have recorded only one case of identical grading (albeit with slightly varied weighting), whereas, in the remainder of the sample, the outcomes of the AI-generated texts are invariably superior, if not significantly so. The inference here is relatively straightforward: those who benefited most in terms of grading are the ones who initially received the lowest scores. However, examining the specifics of the areas where the AI underperformed is particularly noteworthy compared to the "human-generated" papers. This phenomenon was observed among the students in the C3 category (grades from 28 to 31), and what is intriguing is the detailed analysis of the aspects where AI fell short. Given the lengthy process of recreating a document of equivalent length for each selected paper, we know that quality issues may compromise these data. Nevertheless, the fact that the AI system has, on average, penalized top-performing students (<10%) and subsequently demonstrated equal or vastly superior performance compared to over 90% of the students corroborates our foundational hypothesis.

Table 4 provides a detailed overview of the time, in minutes, that the authors spent recreating a text that was then submitted for evaluation. As said, some standard prompts have been used to uniform the results and exchange data between Jenni (more on the position of the academic reference in the text) and ChatGPT (more on the generation of long texts). As noted in the table, and as was relatively easy to anticipate, the individual and average times for the various students and the three classes are substantially the same. The differences are mainly due to the need for reformulation of sentences, the verification of sources and the improvement of the sentence where the GPT started to become repetitive and trivial in the use of certain words. Here, the difference between a general-purpose AI engine trained on documents that encompass the entirety of human knowledge (ChatGPT) and a specialized AI engine trained on books, articles and, in general, more specific resources (Jenni.AI) has been used to merge the results provided by the two engines.

In all three classes, the time required to create the paper is 52 to 95 minutes if we exclude five outliers (in yellow), which we want to comment on later. The average of the classes is substantially the same, and it takes around one hour and a half to achieve an excellent final grade. This result indicates that artificial intelligence could generate a substantially excellent paper in a fraction of the allocated time, starting from both excellent indications (titles and

introductions provided by the class of the top students) and fair/mediocre ones (those provided by the C1 class). All generated texts were done within the three-hour maximum time limit, although only the texts for two students required the entire time.

Outliers represent an interesting aspect of the lengthy process described above: for students 4,7, 11, 14 and 20, we needed a much longer time, closer to the three-hour threshold, and their generation time has been considerably superior to all the other students. The reason for this extra time that we had to devote to their paper's generation derives from detecting the well-known problem of "hallucinations" in those texts, i.e., text generated by the LLM that appears coherent and plausible but is false or nonsensical. We had to repeatedly request improvements and rephrasing for those students because the AI engine generated some hallucinations, which implied more effort to produce a credible and evaluable final version. Hallucinations, a part of undermining people's trust in AI, constitute a serious issue, especially for those who do not want to check the results of the generated content. Once identified, the problem was resolved through a lengthy series of prompts with the system, which extended the time required and demanded more attention to ensure uniformity in the quality of the final result.

For all other students analyzed, there were no such cases of hallucinations or generation of incoherent texts. However, verifying these issues significantly added to the time required for all 24 texts analyzed. Without this lengthy and costly verification phase, the pure generation of text, even of good quality, would have taken between 25% and 40% of the time reported in Table 4. This highlights the impressive quality even generalist language models achieved in generating academic text.

With a large student cohort, the assessment of their prowess in academic writing is impracticable in an offline setting. This element does not diminish the validity of AI tools in supporting educational activities, nor does it significantly augment the considerable apprehensions surrounding this synergy. It can be asserted that, within this sample and context, all indications suggest a radical shift in the perception of learning assessment through producing written material non-face-to-face. The impact is particularly pronounced on traditional education, homework assignments, and educational approaches involving content creation processes. Our findings reveal that AI performs these tasks in a remarkably superior fashion compared to human efforts, too, in a fraction of the time. Notably, this enhancement is substantial for most participants, culminating in an improvement for students with lower scores ranging from 20% to over 40%. As a bitter observation, we can say that it serves as a significant incentive for ethically weak students to complete tasks of this nature in times that are incomparable to what is usually considered the respective educational workload.

Related to this last comment is another line of inquiry concerned with the validity of student evaluation methods used at various educational levels, from elementary school to bachelor's degree programs. An additional expected contribution was to contemplate whether it still makes sense to allow students to work unsupervised on creating an essay, which could be more effectively accomplished by artificial intelligence. The results of our limited research are self-evident. Suppose we aim to ensure a fair evaluation that certifies a student's knowledge and skills in a specific subject. In that case, the methods of content production not supervised by an instructor do not guarantee this outcome. It has been known for some time that even LLM developers, despite being pressured to provide methods of differentiation and distinction between human-generated and AI-generated content, have not been able to guarantee such results. Even with some cases of "hallucinations" by AI, it is impossible to distinguish a text generated by a human from one generated by artificial intelligence with certainty in areas such as scientific production linked to university courses, leading one to wonder if it will ever be

feasible. Although linguistic aberrations and errors produced by LLM, known as hallucinations, are well-documented, it is statistically impossible at present to accurately recognize content generated by a student and to verify its originality. The outcome of this experiment confirms that medium to low-quality scientific texts generated by artificial intelligence result in significant improvements, especially in students with medium to low final performance. Various educational institutions at different levels due to the continuous progress of these AI tools.

Various educational institutions at different levels have underestimated this phenomenon, primarily due to some educators' unawareness of the potential use of these tools for generating content. It remains common to find course syllabi where the final grade is derived from projects, papers, or texts that students are expected to create independently, individually or in groups. In some contexts, this task becomes essentially replicable by an AI engine, producing inconceivable results within the time and effort constraints set by the credit system. This leads to invalidated outcomes and can discourage students who opt not to use such tools, thereby choosing to gauge their capabilities, resulting in widespread demotivation. The established "hourglass" structure for academic essays includes several sections where AI can substantially contribute, making even lengthy academic essays (such as theses) highly susceptible to the technical and ethical dilemmas highlighted in this brief study. Given the number of ECTS credits allocated to theses and considering the inevitable prolonged timeframe typically involved in thesis production, it is evident that there are many beneficial ways to use AI tools. This raises questions about whether the traditional form of a thesis, as it has been known for decades, still holds relevance.

The technical response is somewhat predictable: it is nonsensical to envision a thesis where 40 to 80% of the content can be AI-generated with minimal student involvement. This level of involvement is certainly not commensurate with the expectations set by current evaluation mechanisms, particularly regarding the "originality of contribution" expected by supervisors and thesis committees. This issue is currently insufficiently acknowledged but leads to another straightforward consideration. If students can use these tools to generate results, relying on their ethical principles becomes difficult. It is a painful discussion, but beyond cynicism, it is a matter of extreme efficiency. Therefore, the preliminary findings of this research illuminate our perspective and research questions. We were able, quite simply, to produce scientific papers of a lower-medium level in one-twentieth of the theoretical time, which consistently outperformed those produced by humans. Moreover, the less effective and efficient students showed the most significant improvement. Concerns were raised about demotivation and, conversely, about students' continued growth if an AI tool can achieve such results with practically no effort.

Therefore, this research's preliminary findings illuminate our perspective and research questions. They suggest that LLMs match and sometimes surpass human efforts in generating high-quality academic texts, especially on more complex topics. Furthermore, the study reveals that in the blind review process, papers crafted by LLMs were mainly superior in terms of academic quality to their human-written counterparts. This revelation opens up a discourse on the implications of such technologies in academic settings, particularly in maintaining academic integrity and fostering critical thinking skills. Moreover, we must delve into educators' evolving role in an AI-assisted educational landscape. Generative AI technologies are altering educational methodologies, content creation, and evaluation processes, especially in disciplines profoundly affected by these advancements. We need to contribute more to understanding AI's impact on education and profoundly re-evaluate pedagogical strategies in the face of these disruptive technologies.

In summary, this study contributes to the global warning that arises from various voices beyond the inevitable enthusiasm related to using AI tools in education. Integrating large language models (LLMs) into academic writing education can offer valuable support for teaching and learning. However, evaluation criteria and methods must be revised to ensure fair assessment and to encourage students to enhance their skills and knowledge in scientific production. Despite various attempts in recent months, it has become evident that a text of medium complexity, such as that required for the academic writing course, when generated by an AI system, cannot be distinguished with absolute certainty from a human-generated text. This raises several issues, particularly concerning the evaluation of written work.

Numerous considerations can be drawn from these results, and many scholars from various perspectives are attempting to verify the superiority of AI in replacing humans in specific tasks. The outcome of our research is limited but very clear: in academic writing, up to a certain level of complexity in the composition of specific components of scientific papers—whether linguistic, conceptual, or about sources—AI demonstrates markedly superior performance and is indistinguishable from human work. The implications for the field of education are vast and extend beyond the scope of this limited research. However, the conclusion is quite evident in a university-level bachelor's course where students must produce content independently. There are no conditions under which the work produced by the student can be considered original and personal.

The domain most significantly impacted by this AI-driven "superiority" in academic writing is undoubtedly that of thesis composition at all academic levels. Currently, the requirements set by various universities for thesis projects vary widely. The established "hourglass" structure for academic essays includes several sections where AI can substantially contribute, rendering a thesis highly susceptible to the technical and ethical dilemmas highlighted in this study. Given the number of ECTS credits allocated to theses and considering the inevitably prolonged timeframe typically involved in thesis production, we question whether the traditional form of a thesis, as it has been known for decades, still holds relevance. The technical answer is somewhat predictable: it is nonsensical to envision a thesis where 40 to 80% of the content can be AI-generated with minimal student involvement. Such involvement is certainly not commensurate with the expectations set by evaluation mechanisms, particularly concerning the "originality of contribution" expected by supervisors and thesis committees.

Furthermore, there is an urgent need for ongoing professional development and training for educators to understand and fully utilize AI writing tools in their teaching. This is crucial to ensure that these tools are used effectively and ethically and to adapt evaluation criteria to maintain academic integrity and the meaningful assessment of students' work. Integrating AI-powered writing tools in academic writing instruction can improve efficiency and enhance writing performance. However, balancing an approach that incorporates traditional teaching practices is also essential. Most sources indicate that integrating AI tools in academic writing instruction requires careful consideration and a gradual approach. Reforming writing in educational institutions at any level requires careful consideration because it will involve changes in curriculum standards, teaching practices, student assessments, teacher preparation, and education policies. These modifications necessitate appropriate reflection and experimentation, and educators need to weigh the impressive capabilities of AI writing tools against their important limitations. Case studies discuss the reactions and reflections of learners on the use of AI-assisted writing tools in academic settings, highlighting the positive impact of AI tools on learners’ writing performance and efficiency but also raising concerns about

potential negative effects on their writing skills over time and emphasizes the need for proper instruction on the appropriate application of AI tools in academic writing tasks.

In the more specific and speculative sections, AI's role is limited; however, it can serve as a valuable aid for rephrasing, paraphrasing, and summarizing texts in English. This is immensely beneficial for non-native English speakers and can lead to an undeniable enhancement in the overall quality of the writing and the understanding of academic English, as already supported by numerous bibliographic references. In other parts of the thesis, however, there is no denying that attributes such as "originality," "independent contribution," and "personal interpretation" are, regrettably, left to the ethical principles of the student.

5. CONCLUSIONS

This paper examined the effectiveness of integrating AI-assisted tools in improving learners' writing performance and efficiency in academic contexts. The study also addressed teachers' concerns about potential adverse effects on academic writing skills, emphasizing the importance of providing instruction on the appropriate application of AI tools for academic writing tasks. The findings indicate that using AI-assisted tools can significantly enhance learners' writing performance and efficiency, as AI-generated texts offer numerous suggestions for improvement. However, the superior quality of AI-generated texts compared to students' texts challenges the validity of traditional models for evaluating autonomously produced student papers. The study underscores the importance of incorporating AI tools within a structured learning environment that includes opportunities for discussion, drafting, and revision to support students' development of critical thinking and writing skills. Furthermore, it suggests that teaching practices should include formative feedback from instructors and peers, encouraging students to engage more deeply in drafting and revising to enhance their writing skills. Integrating AI tools in teaching academic writing can improve students' writing abilities while increasing their interest in the subject.

The study also acknowledges concer"s re'arding the potential detrimental effects on students' academic writing skills over time. To address these concerns, explicit instruction on the appropriate use of AI tools should be provided, and continuous monitoring and assessment of their impact on learners' writing skills should be conducted over an extended period. The case study demonstrated the potential benefits of using large language models and AI-assisted tools in teaching academic writing. Future research will aim to involve students more extensively in producing, comparing, and commenting on AI-generated texts and to expand the sample size to statistically strengthen the results presented in this article. Regrettably, the creation of academic texts related to literature reviews appears to be diminishing as a teaching possibility for evaluation purposes. Large language models, particularly for non-native English students, generate good to excellent linguistic texts, rich in appropriate references suitable for a bachelor's level. Future efforts should focus on elevating the level of analysis for master's and PhD students, concentrating on more specific topics and sections of scientific articles beyond mere literature reviews.

REFERENCES

Chong, D., 2021. Research on Artificial Intelligence-based English Writing Blended Teaching Mode. *Journal of Physics: Conference Series*, Vol. 1852. https://doi.org/10.1088/1742-6596/1852/3/032018.

Englert, C. S., Zhao, Y., Dunsmore, K., Collings, N. Y. and Wolbers, K., 2007. Scaffolding the writing of students with disabilities through procedural facilitation: Using an Internet-based technology to improve performance. *Learning Disability Quarterly*, Vol. 30, No. 1, pp. 9-29.

Fitria, T. N., 2021. Grammarly as AI-powered English writing assistant: students' alternative for writing English. *Metathesis: Journal of English Language Literature and Teaching*, Vol. 5, No. 1, pp. 65-78. doi: 10.31002/metathesis.v5i1.3519

Goldenberg, L., Meade, T., Cooperman, N. and Midouhas, E., 2011. Impact of a technology-infused middle school writing program on sixth-grade students' writing ability and engagement. *Middle Grades Research Journal*, Vol. 6, No. 2, pp. 75-97.

Hsiao, J. C. and Chang, J. S., 2023. Enhancing EFL reading and writing through AI-powered tools: design, implementation, and evaluation of an online course. *Interactive Learning Environments*, pp. 1-16. doi: 10.1080/10494820.2023.2212716

Huang, A. Y., Lu, O. H. and Yang, S. J., 2023. Effects of artificial Intelligence–Enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom. *Computers & Education*, Vol. 194, 104684.

Jiang, R. (2022). How does artificial intelligence empower EFL teaching and learning nowadays? A review on artificial intelligence in the EFL context. *Frontiers in Psychology*, Vol. 13, 1049401. doi: 10.3389/fpsyg.2022.1049401

Little, C. W., Clark, J. C., Tani, N. E. and Connor, C. M., 2018. Context and Implications Document for: Improving writing skills through technology-based instruction: a meta-analysis. Review of Education, Vol. 6, No. 2, pp. 202-204. https://doi.org/10.1002/rev3.3115

Liu, C., Hou, J., Tu, Y. F., Wang, Y. and Hwang, G. J., 2023. Incorporating a reflective thinking promoting mechanism into artificial intelligence-supported English writing environments. *Interactive Learning Environments*, Vol. 31, No. 9, pp. 5614-5632.

Meunier, F., Pikhart, M. and Klimova, B., 2022. New perspectives of L2 acquisition related to human-computer interaction (HCI). *Frontiers in Psychology*, Vol. 13, 1098208. doi: 10.3389/fpsyg.2022.1098208

Rowley, K. and Meyer, N., 2003. The effect of a computer tutor for writers on student writing achievement, *Journal of Educational Computing Research*, Vol. 29, No. 2, pp. 169-187.

Sánta, B., 2015. Introducing the Literary Critic: The CARS Model in the Introductions of Academic Papers in Literary Criticism. *International Journal of English Language & Translation Studies*, Vol. 3, No. 2, pp. 84-93. Retrieved from http://www.eltsjournal.org

Seufert, S., Burkhard, M. and Handschuh, S., 2021. Fostering Students' Academic Writing Skills: Feedback Model for an AI-enabled Support Environment. *Innovate Learning Summit Proceedings. Online, US,* pp. 49-58.

Song, C. and Song, Y., 2023. Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. Frontiers in Psychology, Vol. 14. https://doi.org/10.3389/fpsyg.2023.1260843

Stanford University, 2023. *AI Tools in Teaching and Learning*. Available at: https://teachingcommons.stanford.edu/news/ai-tools-teaching-and-learning

Sun, Z., Anbarasan, M. and Praveen Kumar, D. J., 2021. Design of online intelligent English teaching platform based on artificial intelligence techniques. *Computational Intelligence*, Vol. 37, No. 3, pp. 1166-1180.

Swales, J., 1990. *Genre Analysis*. Cambridge University Press.

Warren, S. J., Dondlinger, M. J. and Barab, S. A., 2008. A MUVE towards PBL writing: Effects of a digital learning environment designed to improve elementary student writing. *Journal of Research on Technology in Education*, Vol. 41, No. 1, pp. 113-140.

Wu, L., Wu, Y. and Zhang, X., 2021. L2 learner cognitive psychological factors about artificial intelligence writing corrective feedback. *English Language Teaching*, Vol. 14, No. 10, pp. 70-83. doi: 10.5539/elt.v14n10p70

Yan, D., 2023. Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, Vol. 28, pp. 13943–13967. doi: 10.1007/s10639-023-11742-4