ExELL Explorations in English Language and Linguistics

ISSN 2303-4858

13.2 (2025): 159–180
https://doi.org/10.51558/2303-4858.2025.13.2.159
Original scientific article

Iterative refinement: Using the generative AI chatbot Copilot as a feedback tool – Evidence from the ELT classroom

Victoria Eibinger
Hannes Fromm
Margit Reitbauer
University of Graz, Austria

ABSTRACT

Recent advances in artificial intelligence have opened the door to new applications in the teaching of writing. This article addresses a current research gap by combining automated AI-based feedback, provided by Microsoft Copilot, with teacher feedback and student self-correction. The present study investigates the effectiveness of AI-assisted feedback on EFL students' writing skills. Data were collected from 41 university students. The feedback from Copilot and the subsequent revisions made by students were analysed using MAXQDA according to categories such as genre conventions, accuracy, lexical scaffolding, and content. Results suggest that considerable improvements in writing skills, especially in the areas of lexical scaffolding and line of argumentation, can be achieved through AI-assisted feedback.

ARTICLE HISTORY

Received
9 November 2025
Accepted
28 November 2025

KEYWORDS

AI chatbot;
EFL writing instruction;
AI-assisted feedback;
AI-literacy.

1. Introduction

Academic writing skills play a crucial role in the development of language proficiency. Enhancing these skills requires learners to develop proficiency in writing organization, coherence, grammar, and vocabulary (Campbell, 2019). Learners who can effectively communicate their ideas gain an advantage across professional domains (Yoon, 2011). The complex and demanding cognitive process of writing necessitates multiple cycles

of revision for students to produce cohesive and coherent texts that meet the standards of a given genre (MacArthur & Graham, 2016). We assume that this revision process can be most effectively supported by providing learners with real-time feedback on their drafts.

However, providing students with feedback is often time-consuming, particularly in large classes or when dealing with multiple drafts (John & Woll, 2020). Consequently, teachers' workload increases (Kerman et al., 2022), though the incorporation of AI feedback tools such as Copilot could help mitigate this effect. In line with Bruning and Horn (2000), we argue that despite the support offered by AI tools, teachers must remain in the feedback loop and provide additional input to sustain student motivation. By combining teacher feedback, AI-generated feedback, and self-feedback, we aim to enhance students' motivation during the revision process.

The use of artificial intelligence for writing instruction holds many possibilities and is revolutionizing the way in which feedback can be provided (Harunasari, 2023; Imran & Almusharraf, 2023; Shen & Chen, 2025). Since AI chatbots like Copilot can provide real-time feedback and guidance on vocabulary, grammar, and syntax, they can support students' writing development in a way that is tailored to their individual needs and abilities. We chose to test Copilot (a GPT 4.0-based AI chatbot) as a feedback tool because it is very efficient in terms of its well-structured presentation of feedback and is superior to AI tools such as Grammarly and ProWritingAid in terms of checking grammar mistakes (Schmidt-Fajlik, 2023). In addition, AI chatbots can provide feedback not only on grammar and vocabulary, but also on coherence and students' overall language level (Escalante et al., 2023). The feedback from AI chatbots can also highlight strong and weak points in students' performance, which allows them to better focus on areas that still need improvement (Beccaluva et al., 2023). Furthermore, the nature of AI-generated feedback makes it possible for students to ask for clarification and explanations (Alsaweed & Aljebreen, 2024).

We opted for the use of corrective iterative feedback to provide continuous support for learners. Our aim was to close the gap between their current performance and the target performance as discussed by Hattie and Timperley (2007). However, it is questionable whether Hattie and Timperley's (2007: 88–90) important feedback questions, "Where am I going?" (Feed Up), "How am I going?" (Feed Back), and "Where to next?" (Feed Forward) can be answered by using AI. This provides another valid argument for teachers having to stay involved in the feedback process to guarantee the long-term development of writing skills. Although the chatbot at times includes evaluative and tutorial components, suggests alternative options, and poses engaging questions, the empathetic component that teachers who know about their learners' individual needs and personal learning trajectories can add is not comparable to AI-generated feedback and thus remains indispensable.

We hypothesize that the guidance on vocabulary, grammar, and line of argumentation provided by AI helps improve EFL students' writing skills over time. However, as with all didactic approaches, careful design and implementation are prerequisites for ensuring effectiveness. When used as complementary feedback tools, chatbots based on

LLMs such as ChatGPT and Copilot are valuable, as they have the potential to deliver feedback derived from (ideally frequently updated) big data into the classroom (Beck & Levine, 2023). By providing feedback in the form of human-like texts, chatbots can create a highly interactive writing experience for EFL learners. Combining teacher feedback, self-feedback, and AI-generated feedback enables the development of more engaging exercises, such as prompt writing, as well as creative and collaborative writing activities. Regular, individualized feedback can be provided in real time, thereby effectively supporting students' writing development.

AI chatbots also have the potential to create learning environments that mirror collaborative interaction and provide feedback and scaffolding. This important interplay between feedback, learner engagement, and collaborative learning contributes to the effectiveness of AI-supported writing instruction (Barrot, 2023; Huang & Tan, 2023). This form of instruction also facilitates assessment for teachers and, as a result, enhances teachers' self-efficacy and ultimately learners' writing skills (Mizumoto & Eguchi, 2023; Mizumoto et al., 2024; Shen & Teng, 2024). It can be assumed that teachers employing innovative teaching strategies create positive and supportive learning environments for their students and thus ultimately foster motivation and engagement with the writing process (Ryan & Deci, 2000; Wei, 2023). Again, it must be stressed that teachers need to stay in the loop since AI-assisted feedback has limitations in terms of commenting on certain aspects of the writing process, such as creativity, grammar, and genre conventions.

As far as the uptake of feedback provided by a chatbot is concerned, teachers need to make sure that students develop appropriate AI literacy skills to effectively navigate and integrate AI feedback into their writing tasks (Zhao et al., 2024). This requires them to be aware of the strengths, weaknesses, and potential biases of AI chatbots to corroborate the accuracy of AI responses throughout the writing and revision process. In addition, EFL learners need to be able to incorporate AI-generated text in their own writing in an ethically justifiable way (Warschauer et al., 2023).

Since AI chatbots sometimes produce so-called hallucinations, we argue that we should also endeavour to raise students' awareness of the shortcomings of AI-generated feedback to arrive at an integrated understanding of its potential as a language learning tool. We hypothesize that the learning process might be more effective if students became aware that the acquisition of writing skills can only happen when they critically engage with their drafts and the feedback provided by AI assistants, and engage in follow-up activities in cooperation with their teachers.

Virtually all stages of the writing process (i.e., pre-writing, during-writing, and post-writing) can be supported by AI chatbots (Su et al., 2023). In line with Steinhoff (2023), we argue that it is essential in this process that AI takes on the role of a writing tutor or a peer in collaborative writing rather than that of a ghostwriter. Keeping this in mind, tasks such as outline preparation, content revision, proofreading, and post-writing reflection can be facilitated in AI-assisted writing classes, the text type used in the present study. All these skills are particularly relevant in drafting opinion essays. So far, many studies have investigated the efficiency of feedback by comparing different feedback

types or feedback conditions to no-feedback groups (e.g., Aslam et al., 2025). However, research into traditional feedback forms combined with AI-assisted feedback remains scarce. It is this gap in research that we attempt to address in this article.

2. Methodology

The objective of this research is to assess the effectiveness of chatbots such as Copilot (based on GPT 4.0) as a technology that provides iterative feedback for students working on a writing task in an undergraduate EFL classroom. We collected data in the form of first drafts and revised versions of students' essays, as well as transcripts of students' chats with Copilot. A comparison of the first and the revised version allowed for the tracking and assessment of changes made based on AI feedback. We assessed the effectiveness of AI-generated feedback by investigating the uptake of AI suggestions and how this uptake affected the quality of student texts.

This study was conducted at the Department of English Studies at the University of Graz in the language course "Language Productive and Receptive Skills" designed for second-year undergraduate students of English at C1 level (according to the CEFR). 41 students of three parallel course groups (taught by two of the authors) consented to the processing of their assignments for this study.

The data analysed in this study stem from opinion essays, which is the genre students are taught in this class. This text type is chosen to foster critical thinking and argumentation skills using formal English. The specific type of essay taught in this course comprises five paragraphs: firstly, an introductory paragraph which functions as a general preamble, then zooms into the topic and culminates in a clear and concise thesis statement. This, in turn, serves as both the semantic anchor of the text as well as the point of departure for the subsequent three body paragraphs. These follow a deductive structure, which begins with a topic sentence containing one of three carefully chosen arguments in support of the thesis statement based on one of two propositions that students could choose from. The argument stated in each topic sentence is then supported, explained, and exemplified in primary and secondary support sentences in the body paragraphs. The essay is rounded off by a concluding paragraph, which succinctly reiterates the thesis statements and provides a brief recapitulation of the three supporting arguments.

At this point, we deem it important to point out that genre and text type conventions likely pose a challenge for the use of AI in the role of a ghostwriter or an officially allowed feedback assistant. The reason for this is that definitions and conceptions of genres such as essays, or even specific formats like five-paragraph essays, vary significantly across world regions, educational systems, and Anglophone cultures. For instance, certain definitions may value a line of argumentation exclusively in favour of the thesis statement. By contrast, others might demand the inclusion of counterarguments since counterevidence and granting a stage to opposing views is considered to bolster one's own credibility in academia.

However, it is unclear which definition a chatbot bases its feedback on. If, for example, the AI were to suggest the incorporation of counterarguments, it would thereby satisfy the definition it learned on the basis of large data samples – but it would infringe upon the criteria outlined in our course. This goes to show that definitions, standards, and conventions that AI chatbots rely on as points of reference are usually unknown to the user and might not align with the ones underlying a specific pedagogic or academic task – and may thus lead to less than satisfactory feedback.

Tailoring AI-generated feedback specifically to a given task would therefore require minute and comprehensive prompting. Otherwise, students using an AI to ghostwrite might risk exposure, while students using an AI as an academically sanctioned tutor or peer might receive unhelpful or misleading feedback. Prompting would therefore have to go well beyond merely copying the task description into the chat. It would likely have to include a clear definition of the text type required, which, as mentioned above, might diverge from definitions the AI operates with. When considering this, we suspected that the AI could potentially fall short of giving meaningful feedback without these specific requirements, so to speak, "in mind." Therefore, we took care to repeatedly communicate the rather specific criteria for five-paragraph essays to students prior to the assignment. This allowed us to assess both the suitability of AI-generated feedback for the task specified in the assignment, as well as students' awareness of text-type requirements.

2.1. *Procedure*

To obtain data on feedback provided by Copilot and subsequent changes to the essays, students were asked to submit four different types of language data on two separate occasions. As can be seen in Figure 1, students were first instructed to submit an outline of their essay to their teachers, which consisted of an introductory paragraph, topic sentences for each of the three supporting paragraphs, as well as a few keywords for their arguments, and a concluding paragraph. This AI-independent phase was included to ensure that students relied on their own critical thinking skills to create their lines of argumentation.

In a second step, students were then asked to draw on feedback from their course teachers to write the first full drafts of their essays. As a third step, students were instructed to upload their first drafts to the chat with Copilot and request feedback from the chatbot. Upon receiving AI-generated feedback, students were encouraged to revise their essays in the fourth and final step. They were then asked to submit their first drafts as well as their chat transcripts in addition to the final versions of their essays. These final versions were then corrected by their course teachers. After the elimination of five assignments in which students used ChatGPT 3.5 instead of Copilot as specified in the task description, 41 datasets were included in the corpus of this study.

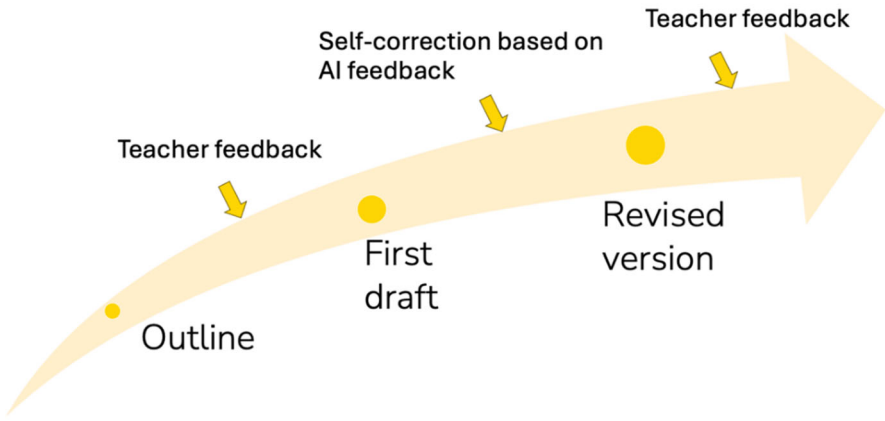

Figure 1: Task structure and stages of iterative feedback

2.2. *Methods*

Datasets obtained from students were coded using MaxQDA and analysed using Braun and Clarke's (2006) model for thematic analysis. All three authors were involved in the definition of categories for coding the language data in student texts and chat transcripts. Individual datasets were subsequently manually coded by one of the authors and, in a second coding wave, reviewed by another. In a third and final coding wave, discrepancies in coding were discussed among all three authors of this study and resolved by further specification of existing codes and subcodes or, if needed, the addition of new codes.

We organized codes into three general categories: student prompts, AI-generated feedback, and codes that function as descriptors for both AI-generated feedback, as well as changes in student texts. The first category contains all student contributions to chats, which were not further divided into subcategories. The second main category, AI-generated feedback, contains the codes "praise & motivation," "feedback on the process of writing," "summary & meta-structure," "phatic & organization of chat," and "AI-generated text," further divided into the subcodes of "sentences and phrases" and "full text or paragraph" (see Appendix).

The third and most comprehensive general category encompasses descriptors of feedback and changes. These descriptors were used to code both AI-generated feedback geared towards specific aspects of the essays or types of mistakes, as well as changes that students made to their texts in the revision process. Descriptor codes were further divided into four overarching categories: firstly, "lexical scaffolding" contains the subcodes of "register," "word choice," and "cohesion." Secondly, "accuracy" is comprised of "grammar," "syntax," "spelling," and "punctuation." Thirdly, "content" encompasses the subcodes of "coherence," "topicality," and "line of argumentation." The fourth and final descriptor category was labelled "genre conventions." Occasionally, one

stretch of text was assigned two codes if applicable, for example, when the AI assistant suggested the provision of counterarguments. While this clearly falls into the area of line of argumentation, it also infringes upon the text requirements specified in the task description and was therefore additionally coded as genre conventions.

During the process of coding, some of the codes and subcodes defined above emerged as particularly prevalent in both AI-generated feedback and changes in student revisions. Based on these patterns, we decided to add a second, separate type of coding in order to gather evaluative data in a structured way: to evaluate students' changes to their first drafts based on AI-generated feedback, the first full drafts and final versions of each essay were compared using the Track Changes function in MS Word. Adjustments were assessed based on the grading scale used in the course and outlined in the course handout (which was introduced and explained to students before the assignment of the essay writing task). Adjustments to texts were categorized as (a) "line of argumentation," (b) "word choice," (c) "cohesion and coherence," (d) "accuracy," and (e) "register" (see Figure 4 below). These categories were chosen because they represent the most frequent and significant types of changes made by students based on AI-generated feedback. The respective changes were assessed as either "neutral," "positive," "negative," or "no changes." This additional type of evaluative coding allowed us to assess the effectiveness and potential of AI as a feedback assistant for individual aspects of writing in a more differentiated way.

The main objective of this study is to assess the effectiveness of feedback provided by Copilot for improving students' writing skills. In order to do this, our data analysis is mainly guided by the following three research questions:

RQ1. Which aspects of students' texts does Copilot focus on in its feedback?

RQ2. Which aspects of the AI feedback were taken up by students?

RQ3. Which aspects of their writing did students improve based on AI feedback?

3. Results

To answer the three research questions, this section presents the results of the three analyses we conducted: first, the interactions between students and Copilot; next, the revisions students made after receiving feedback from the chatbot; and last, an evaluation of students' revisions according to the categories described above.

3.1. *Student-AI conversations: Which aspects of students' texts does Copilot focus on in its feedback?*

An analysis of the conversations between students and Copilot revealed that the code summary and meta-structure was the most prevalent one with 38.6% (432 instances out of 1119). This can be seen in Figure 2. The reason why this code occurred with such high frequency is that Copilot tends to structure its feedback into categories and subcategories such as "Introduction" and "Conclusion," "Counterarguments," "Supporting

Evidence," "Language and Style," "Transitions and Coherence," as well as "Text Type and Conventions." The prevalence of this code is therefore a result of Copilot's way of structuring its feedback.

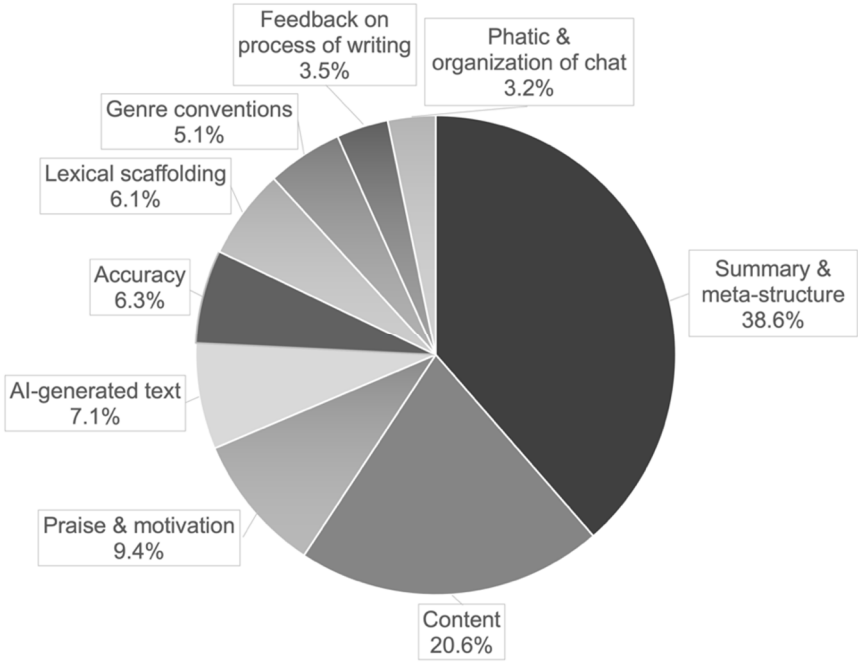

Figure 2: Codes used for conversations between students and Copilot

Interestingly, while showing some similarities, the headings used by Copilot to structure its feedback into sections were different in the individual conversations between students and the chatbot. For example, the category might be called "Evidence and Examples" in the interaction with one student but "Reasons and Evidence" in a chat with another. It is difficult to judge whether these variations are random or whether they might be due to differences in the way students reasoned in their texts and how they built their arguments. Similarly, a category referring to vocabulary choice and grammar mistakes was called "Grammar and Style" in one interaction and "Accuracy" in another.

Some students' feedback was organized into entirely different categories that were based on the main arguments that students put forward in their essays, such as "Engagement and Active Learning," "Professor-Student Interaction," and "Balancing Autonomy and Necessity." This mainly happened when students uploaded their essays in two parts rather than one, and might showcase Copilot's adaptability in structuring its feedback and tailoring it to the data it is provided with. This arguably represents an added value of AI as a feedback assistant in that it provides a perspective that is uncompromised by pre-defined marking criteria and feedback categories defined by teachers. This is in line with Baker (2016) as well as Maity and Deroy (2024), who argued that AI

feedback can adapt to individual learners' needs and learning styles, thereby adding value to the learning experience. In this way, it complements teacher feedback in an unprecedented way.

A clear focal point of the feedback and advice provided by Copilot is the category of content with 20.6% (231 instances), specifically the subcode line of argumentation with 210 occurrences. This suggests that Copilot primarily focuses on the logical structure and persuasive elements of students' essays. For example, it frequently draws students' attention to the fact that more precise examples, counterexamples, or evidence could strengthen their arguments. We found that Copilot is even capable of recognizing logical inconsistencies. For instance, it can point out that a student should write "homework at university should therefore not be mandatory" rather than "homework at university should therefore be mandatory."

As can be seen in Figure 2, another common code was praise & motivation with 9.4% (105 instances) since Copilot prefers to end its feedback on a positive note (e.g., "Overall, your essay is well thought out and presents a compelling case for voluntary homework. Keep up the good work! 👍"). This example illustrates that Copilot occasionally uses emojis, primarily thumbs-up and smiley-face emojis.

The next most frequent code was AI-generated text with 7.1%, subdivided into 65 instances where Copilot suggested individual sentences and phrases compared to only 15 instances where it provided longer stretches of text (full text or paragraph) to substitute passages in student essays. In other words, Copilot tended to give more general feedback in the form of bullet points; if it suggested specific ways in which students' texts could be improved, it primarily did so in the form of short phrases. This was also the case when students provided relatively vague prompts and simply asked for "feedback," "critical feedback," or "suggestions for improvement."

This represents a notable shift from GPT 3.5 to GPT 4.0. According to pilot studies we conducted, chatbots based on older LLMs, such as ChatGPT 3.5, seem to be more prone to rewriting students' texts for them when students' prompts are not specific about the form of feedback they prefer. This tendency was not observable in our current study with Copilot, which is based on GPT 4.0. As a result, there were fewer instances of the code full text or paragraph than we initially expected.

Other prominent categories included accuracy with 6.3% (71 instances), with a focus on grammar and syntax (29 and 25 instances, respectively), lexical scaffolding with 6.1% (68 instances), with a particular emphasis on the subcode word choice (40 instances), and genre conventions with 5.1% (57 instances).

While chatbots based on GPT 4.0 excel at finding more obvious, surface-level mistakes, they are not yet reliably able to detect and elaborate on more complex grammatical and syntactical mistakes, as this would require an in-depth understanding of a sentence's meaning and deep structure that AI tools are not yet capable of. We suspect that this is why, in our data, the accuracy code occurred with relatively low frequency. This was not necessarily the case because students made few grammatical (e.g., tense,

preposition) and syntactical (e.g., word order) mistakes; Copilot simply did not point them out consistently.

3.2. *Student revisions and uptake: Which aspects of the AI feedback were taken up by students?*

After receiving feedback from Copilot, the 41 students made 413 revisions in total. As can be seen in Figure 3, the majority of these revisions can be categorized as lexical scaffolding (209 instances; 50.6%). Of these, the subcode word choice was predominant (152 instances). This suggests that students respond strongly to feedback on vocabulary and correct word choice in their essays.

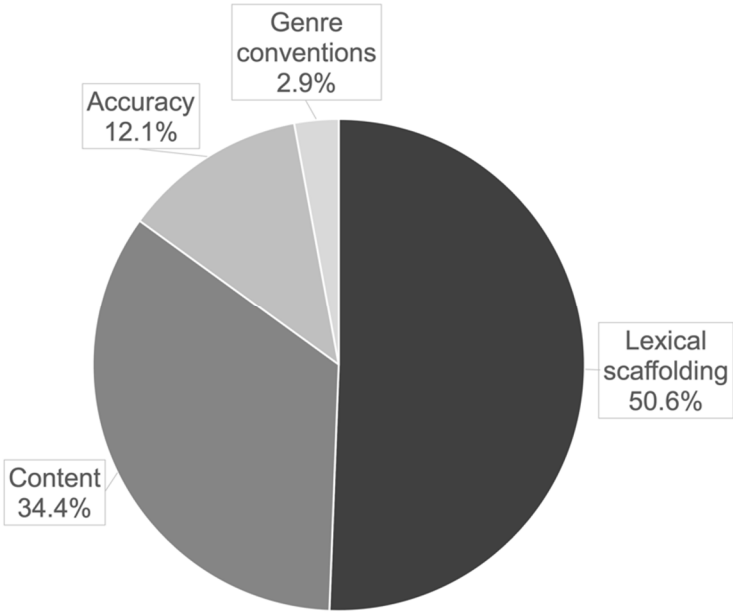

Figure 3: Students' revisions after receiving AI feedback

Considering that only 40 instances of word choice were coded in the student-AI conversations (see Table 1), it is noteworthy that students made 152 revisions to their choice of vocabulary, which is an uptake of 380%. The reason for this is that Copilot frequently provided three or more suggestions under a heading such as "Language and Style," which we did not code as separate instances. Each coded segment resulted in 3.8 revisions on average, which indicates that students adopted most of Copilot's suggestions.

Table 1. Revisions and uptake in relation to codes in student-AI conversations

Codes	Student-AI conversations	Revisions	Uptake
	n	*n*	%
Lexical scaffolding	**68**	**209**	**307**
Word choice	40	152	380
Cohesion	17	41	241
Register	11	16	145
Accuracy	**71**	**50**	**70**
Spelling	4	5	125
Punctuation	7	8	114
Grammar	29	22	76
Syntax	25	15	60
Unspecified	6	0	0
Content	**231**	**142**	**61**
Unspecified	3	6	200
Topicality	2	2	100
Coherence	16	11	69
Line of argumentation	210	123	59
Genre conventions	**57**	**12**	**21**
Total	**427**	**413**	

Moreover, the number of content revisions was notable, accounting for 34.4% of total revisions (142 instances), with the line of argumentation again being the focus of attention. Students made 123 revisions to their lines of argumentation altogether, which reflects students' high willingness to strengthen their arguments by including more specific examples and precise reasoning. A close examination of AI commentary on argumentation confirms its ability to critically assess students' arguments in most cases. The high number of revisions also highlights a certain lack of argumentative prowess on the part of the students.

Another interesting observation is that the uptake of accuracy-related feedback was very high, with 70%: the 71 accuracy instances in the AI feedback resulted in 50 revisions (12.1% of total revisions), which were comprised of 22 grammatical and 15 syntactical changes, with the rest divided between punctuation and spelling. Even though the number of revisions students made in this area is low compared to lexical scaffolding and

content, the high uptake suggests that students were quite receptive to making changes to the grammar, syntax, punctuation, and spelling in their essays.

In contrast, the uptake of feedback coded as relating to genre conventions was relatively low, with 21%. The 41 students only made 12 changes related to this code in total, despite 57 instances of genre conventions being coded in student-AI conversations. This is possibly because students recognized that many of Copilot's suggestions were not necessarily suitable for the type of opinion essay that students were instructed to write in this class. For example, the AI tool frequently urged students to include counterarguments in their body paragraphs and a call to action in their conclusions, which was not necessary in the case of this genre.

3.3. *Assessment of student revisions: Which aspects of their writing did students improve based on AI feedback?*

To answer the third research question, we evaluated the revisions that the 41 students made to their first drafts. As described in the "Methodology" section above, we focused on five categories and classified the collective revisions made by each student in each category as positive, negative, neutral, or no changes. Figure 4 shows that revisions across all five categories were predominantly positive, with negative changes being minimal. This indicates that AI feedback typically steers students in a constructive direction.

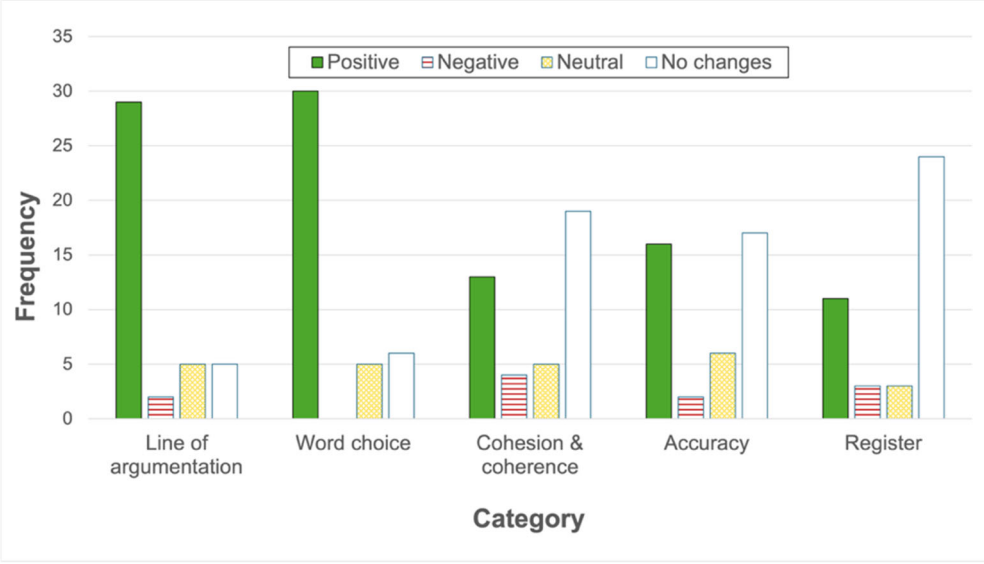

Figure 4: Overall assessment of revisions per category

Revisions were particularly successful in the categories of line of argumentation and word choice. There were 29 and 30 positive changes, respectively, and only 2 negative changes in total, both relating to the line of argumentation. This demonstrates that AI

feedback is especially effective in guiding students towards clearer argument construction and more appropriate vocabulary choices.

Of the remaining three categories, register was the one where the fewest changes were made. 24 of the 41 students opted not to make any changes at all. This is unsurprising as Copilot rarely provided feedback related to register (11 instances in total), perhaps because students were mostly competent in using formal language.

Similarly, in the cohesion & coherence category, there were 19 students who did not make any revisions. This suggests that students were either satisfied with the linking in their drafts, disagreed with Copilot's assessment, or found it challenging to implement AI suggestions related to the cohesion and coherence of their essays. The latter could be explained by the predominance of relatively vague instructions in this category, such as "Ensure smooth transitions between your paragraphs to improve the flow of your essay. This can be achieved by using transitional phrases or by ensuring the last sentence of one paragraph links to the first sentence of the next" (Microsoft, 2024).

In the accuracy category, the number of students who made positive revisions was similar to that of students who made no changes (16 positive, 17 no changes). While the uptake of accuracy feedback was high, as discussed above, it has to be noted again that Copilot does not always find and point out all grammar and syntax mistakes consistently. The number of negative changes was very low (2 students). This was also the case in the cohesion & coherence and register categories (4 and 3 negative changes, respectively). This underlines that AI feedback facilitated constructive revisions rather than leading to a decrease in the quality of students' essays.

The number of changes categorized as neutral was similar across all categories (between 3 and 6). This was because students occasionally improved certain areas of their writing only to include new mistakes and inaccuracies elsewhere in their essays. For example, they might include a specific example in one of their paragraphs to support an argument as suggested by Copilot, only to make a grammatical mistake when doing so. As a result, even if the student in question improved their line of argumentation and perhaps even corrected a grammar mistake pointed out by Copilot, the new grammar mistake they included would result in the "neutral" label being applied to their accuracy revisions.

4. Discussion of results

The results of the present study indicate that AI feedback can complement human feedback in several ways. It offers valuable insights into the way AI chatbots provide feedback on drafts of writing assignments and how students respond to this feedback. These insights, in turn, allow for conclusions regarding the targeted implementation of AI chatbots into the iterative feedback process. Specifically, they enable us to differentiate between features of writing that seem to be predestined for AI feedback on the one hand, such as certain aspects of argumentation, and on the other hand, areas which are better catered to by human instructors, such as more complex grammar and syntax and a deep

understanding of the language. Thus, the results of this study contribute to the current body of knowledge about the potential of AI in language teaching and help chart differences and overlaps between human teachers and AI chatbots in the role of feedback assistants.

4.1. *Student-AI chats and AI-generated feedback*

(RQ1) The findings from the first analytical stage, which focused on the student-AI conversations, point toward two types of codes which were particularly prevalent in AI feedback and together account for 59.2% of the total number of codes: the codes summary & meta-structure, as well as content. The fact that 38.6% stem from the former category suggests that AI chatbots provide their feedback in a structured way using framing devices, (sub-)headings, bullet points, introductory phrases, and connectors between items.

Notably, content, and within this theme, the subcategory line of argumentation in particular, stand out as the feedback categories that AI comments on most frequently. In many cases, feedback on content and argumentation is also more elaborate when compared to other categories, such as accuracy. This ability of AI chatbots to not only assess student arguments but also provide meaningful feedback for improvement has the potential to significantly reduce the workload of teachers.

To summarize, the capacity for generating long stretches of explanatory and organisational language is a key strength of AI feedback assistants that can productively complement human teachers. Firstly, this capacity helps structure feedback and meaningfully frame individual feedback items. Secondly, and arguably more importantly, it provides more elaborate explanations and clarification for reasoning as well as examples when critically assessing student arguments than human instructors typically do.

4.2. *Student revisions and uptake of feedback*

(RQ2 and RQ3) The results from the second and third stages of the analysis, in which we coded and assessed the revisions that students made to their opinion essays, suggest that students are quite open to making revisions in the areas of lexical scaffolding (50.6% of total revisions) and content (34.4%). In the former category, the changes were primarily made to stretches of text coded as word choice (36.8%). The majority of these revisions were positive: 30 of 41 students used word-choice-related AI feedback to improve their use of vocabulary, and not a single student made vocabulary revisions that were detrimental to the quality of their essay. This indicates that AI feedback on word choice has considerable potential for writing instruction, as the feedback seems to be appropriate, well accepted, and rarely appears to mislead students into making erroneous changes.

A second common subcode of the lexical scaffolding code was cohesion, which accounted for 9.9% of total student revisions. We coded revisions as cohesion if they were related to students' use of linking devices, such as transitional words and phrases,

pronouns, and other structures used to ensure the cohesiveness of a text. When we assessed whether students' revisions were positive or negative, we took both cohesion and coherence into account. We found that only 13 students made positive changes to their essays in this category, whereas 19 students did not make any changes at all. This means that there was a positive effect of cohesion and coherence-related AI feedback on student texts in only 34.1% of cases, which is in line with Alsaweed and Aljebreen's (2024) findings, who reported that incorrect or missing connectors were correctly coded by ChatGPT 3.5 in only 38% of cases. These results are in slight contrast to Allen and Mizumoto's (2024) findings, which suggested that using AI feedback for proofreading and editing considerably enhances the cohesion and clarity of students' texts. While we observed a similar positive effect on the clarity of students' essays (see the discussion of the content category below), our findings with respect to cohesion and coherence were less positive. This could be due to both the relatively vague feedback provided by Copilot, as well as a strong focus on cohesion and coherence in the ELT programme at the University of Graz.

In the content category, students mainly revised their lines of argumentation. The high number of revisions in this category, which were primarily positive, showcases the importance of iterative feedback cycles. First, students improved the clarity and strength of their arguments as a result of teacher feedback on their essay outlines. In addition to their uptake of teacher feedback, the majority of students were able to further improve their lines of argumentation after receiving more extensive feedback from Copilot on their first drafts, for example, by including specific examples as suggested by the AI. This is in line with Allen and Mizumoto's (2024) findings in that it underlines generative AI's considerable potential to improve the clarity of students' texts and arguments.

Our data from the accuracy category also raise some interesting points of discussion. While the feedback provided by Copilot focused more on the line of argumentation, for example, than on accuracy, students' uptake was higher in the latter category. The total number of 50 accuracy-related revisions that students made was considerably lower than the 123 revisions related to line of argumentation. Looking at students' uptake of AI-feedback presents a different picture, however, because the 50 revisions that students made to their grammar, syntax, spelling, and punctuation were a result of only 71 suggestions offered by Copilot regarding students' accuracy. This represents a higher uptake of accuracy-related feedback than that of line of argumentation-related feedback.

It therefore seems that students are even more receptive to making changes to their grammar, syntax, punctuation, and spelling. We can speculate that this is because they found it easier to implement these accuracy-related changes rather than make changes to the content of their essays. It is also possible that students regarded Copilot's feedback as more credible with regard to accuracy, which would make sense considering that some of the argument-related feedback was not necessarily relevant to the specific type of opinion essay students were supposed to write. Suggestions that students knew they could safely ignore, for example, were related to the inclusion of counterarguments in their body paragraphs or a call to action in their conclusion. Follow-up interviews would

likely have provided further insight into why students chose to take up certain suggestions over others, offering an intriguing avenue for further research.

The second noteworthy finding that emerges from our accuracy data is that, despite the high uptake, fewer than half of the students managed to improve the grammar, syntax, punctuation, and spelling of their texts. One reason for this is that generative AI sometimes misses more complex grammar and syntax mistakes in students' texts. In line with Al-Garaady and Mahyoob (2023), we found that AI chatbots are relatively successful at discovering surface-level mistakes such as prepositions, spelling, and punctuation mistakes. They can also typically detect problems with subject-verb agreement (e.g., a missing third-person *-s*) or an obvious word order problem. However, more complex mistakes whose detection would require an integrated understanding of semantics and grammar, such as tense mistakes, are often not pointed out.

Interestingly, this inability to detect more complex mistakes also appears to cause AI chatbots to misidentify grammar or syntax errors as punctuation errors. For instance, if a student's sentence contained a grammatical or syntactical mistake, Copilot sometimes suggested adding an additional comma or deleting one that was already there, even if this did not resolve the problem. We speculate that, in these cases, the AI was able to identify that something was wrong with the sentence in question, but that AI technology is not yet advanced enough to always allow it to pinpoint the actual mistakes and rectify them. Therefore, like Mizumoto et al. (2024) and Teng (2024), we argue that AI chatbots have the potential to be a useful first resource for obtaining feedback related to linguistic accuracy in L2 contexts. However, feedback from human instructors is still crucial for addressing more complex grammatical and syntactical mistakes, which, again, underlines the importance of iterative feedback cycles.

Lastly, AI feedback was coded as genre conventions if it was only relevant for a specific type of text or essay, for instance, discussing contrasting viewpoints, including a call to action, or ending the essay with a suggestion for further research. As discussed above, students knew to ignore many of these suggestions, which resulted in comparatively low uptake. While this might imply that generative AI is not currently capable of guiding students in making genre-specific adjustments, it is more likely that the problem lies with students' lack of ability to prompt AI chatbots successfully. They apparently need specific instruction on writing more appropriate and precise prompts in order to receive AI feedback that is useful to the specific text genre they are working on. Otherwise, the AI will simply provide feedback with a generic text type in mind.

5. Conclusion

The present study examined the role and effectiveness of AI feedback assistants in L2 instruction in tertiary education. In our analysis, we focused on aspects of writing that were addressed by the AI as well as the types of feedback that the chatbot provided. We subsequently examined students' uptake of this feedback and improvements they made to their essays.

It has to be noted here that we solely investigated one specific text type, namely opinion essays, which means the results might not necessarily be generalizable to other text types. A further limitation is that we did not focus on students' prompt engineering (cf. Ekin, 2023; Giray, 2023), which might have allowed us to discover what kind of prompts lead to more effective, genre-specific feedback.

The results of our study point to a variety of ways in which AI feedback can complement feedback by human instructors. What is more, we were able to identify aspects of language instruction that readily lend themselves to AI assistance, whereas in the context of others, human instructors appear to be more reliable and achieve better results.

We see great potential for a collaborative approach to feedback involving both human instructors and AI assistants in the area of planning and developing lines of argumentation. Our results show that AI feedback has proven rather beneficial in this context: AI feedback assistants excel in this area because they provide clear and elaborate explanations of complex aspects of writing such as faulty argumentation or ineffective examples. Furthermore, we found that AI chatbots provide their feedback in clear and coherent structures, often presented in lists of bullet points. Our results show that students frequently improved their lines of argumentation upon receiving this type of AI-generated feedback.

AI-generated feedback can also contribute to teacher well-being by helping instructors better manage their time and cognitive resources. While punctuation, for instance, is a type of quick and uncomplicated correction, meaningful feedback on line of argumentation and content requires extensive elaborations, which is time-consuming for instructors. This type of feedback can be outsourced to AI-feedback assistants and can thus positively contribute to what we term teacher economy. By teacher economy, we understand the instructors' dilemma of trying to navigate the line between managing their own time and mental resources on the one hand, by, for example, resorting to abbreviations in marking, and providing students with explanations and elaborate guidance on the other hand.

Teachers, either unconsciously or simply because of limited personal resources, might at times fall short of providing this important type of feedback on more complex semantic structures (cf. Guo & Wang, 2024). This is where we see potential for the incorporation of AI-generated advice into the iterative writing process: our data suggest that AI-assistants such as Copilot can effectively provide this very type of sorely needed feedback on inaccurate, faulty or otherwise flawed lines of argumentation while at the same time fostering teacher well-being. This is in line with Guo and Wang (2024), who found that ChatGPT generated larger amounts of feedback than human instructors.

While tutoring argumentation clearly represents a strength of AI assistants, we suggest a cautious approach to delegating feedback on lines of argumentation to AI exclusively. The reason for this is that an over-reliance on AI in education could have a negative impact on the development of critical thinking skills (Barrot, 2023). This is why our study design included an outline that students produced without the help of AI and that was corrected by human instructors.

To avoid the deskilling of students with respect to the development of critical thinking and argumentation skills, we contend that, in the age of AI, one of the main reasons why writing longer coherent texts is still taught to language students at different levels is that the process of writing in and of itself fosters the development of critical thinking skills. We might consequently pose the question of whether it is a sensible idea to outsource the critical skill of assessing arguments to AI. It could be argued that, in the future, one of the most important areas within which we will rely on our critical thinking skills is to evaluate and assess AI-generated content. Collectively, we might have to ask ourselves: can we look to the very technology whose creations will often be in the focus of our critical analysis to teach future generations critical thinking in the first place?

In order to foster accountability early in language students, it is imperative that instructors emphasize the development of metacognitive awareness, critical thinking skills and metalinguistic awareness when dealing with AI chatbots. For instance, students need to be able to identify feedback that is not suitable to the specifications provided in their assigned tasks or assess the appropriateness of changes suggested by AI assistants. This type of double-checking requires active and critical engagement with the chatbot, which makes a certain level of critical awareness an essential prerequisite on the part of students. Provided that students have acquired basic critical thinking skills, subsequent active engagement with AI assistants can facilitate the further development of higher order critical thinking skills.

The need for the integration of the scaffolding of critical thinking skills ties in with Zhao et al.'s (2024) call for the development of a novel form of AI literacy that encompasses several dimensions. This new type of AI literacy requires users to evaluate AI output in ways that go beyond the technology's current capacities, specifically in the areas of pragmatics, safety, reflective understanding, socio-ethics, and contextual understanding (Zhao et al., 2024).

It is also crucial that instructors develop this new form of AI literacy in order to oversee the development of both writing skills as well as critical thinking skills in language students. The requirements for teachers in this context extend beyond mere knowledge and understanding of AI literacy – they also need to stay in the feedback loop and be able to educate their students about the limitations of current AI assistants.

To prevent pedagogically ineffective use of AI, such as ghostwriting, the iterative design of our task led students to work with AI assistants in the roles of tutors and partners (cf. Steinhoff, 2023). Since students were required to submit outlines and drafts, they could not rely on generative AI to write the essays for them. Rather, they actively engaged with the AI assistants in an iterative process of revising and editing their texts (cf. Teng, 2024). We suggest that this way of integrating modern technologies could be a promising avenue for writing instruction.

Generally, to support students in their critical engagement with AI feedback, teachers need to provide scaffolding and guidance (Celik et al., 2022). To ensure that teachers are capable of integrating generative AI in their teaching practices, it is indispensable that existing teacher training programs be modernized and adapted accordingly to

familiarize teachers with the potential of AI in language education (Dimitriadou & Lanitis, 2023; Royce, 2025; Zhang et al., 2023).

To conclude, recent research in this field as well as the findings of this study indicate that the future of writing instruction in ELT will be characterized by close collaboration with AI assistants in different roles and in a variety of contexts. The challenges that teachers and lecturers have been facing since the advent of LLMs at the end of 2021 are merely harbingers of more fundamental issues in language and writing instruction as well as education in general. For example, in the context of writing instruction, educators will likely have to ask themselves which genres will still need teaching in the future. Similarly, they might have to revisit debates on authorship and plagiarism both in and outside of the classroom.

In order to prepare teachers for imminent changes in the educational landscape, future research will need to build on studies such as the present one and continue to assess the vast and ever-growing potential of AI feedback assistants. The targeted and effective allocation of specific tasks to AI feedback assistants is a first step. However, there is also a need to further develop AI literacy and thereby to safeguard the development of students' critical thinking skills. Promising focal points for further research in this context are effective and purposeful prompting in the writing classroom as well as the development of students' ability to critically assess AI chatbots' responses and feedback.

References

Al-Garaady, Jeehaan, Mohammad Mahyoob (2023). ChatGPT's capabilities in spotting and analyzing writing errors experienced by EFL learners. *Arab World English Journal 9:* 3–17. https://dx.doi.org/10.24093/awej/call9.1

Allen, Todd, Atsushi Mizumoto (2024). ChatGPT over my friends: Japanese English-as-a-foreign-language learners' preferences for editing and proofreading strategies. *RELC Journal*. https://doi.org/10.1177/00336882241262533

Alsaweed, Waad, Saad Aljebreen (2024). Investigating the accuracy of ChatGPT as a writing error correction tool. *International Journal of Computer-Assisted Language Learning and Teaching 14*(1): 1–18. https://doi.org/10.4018/IJCALLT.364847

Aslam, Uswa, Maryam Dilawar, Muhammad Nadeem Anwar, Muhammad Zahid (2025). Comparative analysis of English language teachers' and Chat GPT-generated assessment and feedback on argumentative essays by Pakistani undergraduates. *Journal of Applied Linguistics and TESOL 8*(2): 178–188. https://jalt.com.pk/index.php/jalt/article/view/595

Baker, Ryan S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education 26*(2): 600–614. https://doi.org/10.1007/s40593-016-0105-0

Barrot, Jessie S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing 57*: 100745. https://doi.org/10.1016/j.asw.2023.100745

Beccaluva, Eleonora Aida, Fabio Catania, Fabrizio Arosio, Franca Garzotto (2023). Predicting developmental language disorders using artificial intelligence and a speech data analysis tool. *Human–Computer Interaction 39*(1–2): 8–42. https://doi.org/10.1080/07370024.2023.2242837

Beck, Sarah W., Sarah R. Levine (2023). Backtalk: ChatGPT: A powerful technology tool for writing instruction. *Phi Delta Kappan 105*(1): 66–67. https://doi.org/10.1177/00317217231197487

Braun, Virginia, Victoria Clarke (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology 3*(2): 77–101. https://doi.org/10.1191/1478088706qp063oa

Bruning, Roger, Christy Horn (2000). Developing motivation to write. *Educational Psychologist 35*(1): 25–37. https://doi.org/10.1207/S15326985EP3501_4

Campbell, Madelaine (2019). Teaching academic writing in higher education. *Education Quarterly Reviews 2*: 608–614. https://doi.org/10.31014/aior.1993.02.03.92

Celik, Ismail, Muhterem Dindar, Hanni Muukkonen, Sanna Järvelä (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends 66*(4): 616–630. https://doi.org/10.1007/s11528-022-00715-y

Dimitriadou, Eleni, Andreas Lanitis (2023). A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. *Smart Learning Environments 10*(1): 12. https://doi.org/10.1186/s40561-023-00231-3

Ekin, Sabit (2023). Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices. *TechRxiv*. https://doi.org/10.36227/techrxiv.22683919.v2

Escalante, Juan, Austin Pack, Alex Barrett (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education 20*(1): 57. https://doi.org/10.1186/s41239-023-00425-2

Giray, Louie (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering 51*(12): 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Guo, Kai, Deliang Wang (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies 29*: 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Harunasari, Siti Youlidhar (2023). Examining the effectiveness of AI-integrated approach in EFL writing: A case of ChatGPT. *International Journal of Progressive Sciences and Technologies 39*(2): 357–368. http://dx.doi.org/10.52155/ijpsat.v39.2.5516

Hattie, John, Helen Timperley (2007). The power of feedback. *Review of Educational Research 77*(1): 81–112. https://doi.org/10.3102/003465430298487

Huang, Jingshan, Ming Tan (2023). The role of ChatGPT in scientific communication: Writing better scientific review articles. *American Journal of Cancer Research 13*(4): 1148–1154.

Imran, Muhammad, Norah Almusharraf (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology 15*(4): ep464. https://doi.org/10.30935/cedtech/13605

John, Paul, Nina Woll (2020). Using grammar checkers in an ESL context: An investigation of automatic corrective feedback. *CALICO Journal 37*(2): 193–196. https://doi.org/10.1558/cj.36523

Kerman, Nafiseh Taghizadeh, Omid Noroozi, Seyyed Kazem Banihashem, Morteza Karami, Harm Jochempje Albertus Biemans (2022). Online peer feedback patterns of success and failure in argumentative essay writing. *Interactive Learning Environments 32*(2): 1–13. https://doi.org/10.1080/10494820.2022.2093914

MacArthur, Charles A., Steve Graham (2016). Writing research from a cognitive perspective. MacArthur, Charles A., Steve Graham, Jill Fitzgerald, eds. *Handbook of Writing Research.* (2nd ed.). New York: The Guilford Press, 24–40.

Maity, Subhankar, Aniket Deroy (2024). Generative AI and its impact on personalized intelligent tutoring systems. *arXiv*: 2410.10650. https://doi.org/10.48550/arXiv.2410.10650

Microsoft. (2024). *Copilot* [Large language model]. https://copilot.microsoft.com/

Mizumoto, Atsushi, Masaki Eguchi (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics 2*(2): 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mizumoto, Atsushi, Natsuko Shintani, Miyuki Sasaki, Mark Feng Teng (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics 3*(2): 100116. https://doi.org/10.1016/j.rmal.2024.100116

Royce, Christine Anne (2025). Becoming AI literate: Preparing future educators for the use of AI. Keeley, Krista LaRue, ed. *AI Applications and Strategies in Teacher Education*. Hershey, PA: IGI Global Scientific Publishing, 1–20. https://doi.org/10.4018/979-8-3693-5443-8

Ryan, Richard M., Edward L. Deci (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist 55*(1): 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Schmidt-Fajlik, Ronald (2023). ChatGPT as a grammar checker for Japanese English language learners: A comparison with Grammarly and ProWritingAid. *AsiaCALL Online Journal 14*(1): 105–119. https://doi.org/10.54855/acoj.231417

Shen, Xiaolei, Mark Feng Teng (2024). Three-wave cross-lagged model on the correlations between critical thinking skills, self-directed learning competency and AI-assisted writing. *Thinking Skills and Creativity 52*: 101524. https://doi.org/10.1016/j.tsc.2024.101524

Shen, Yanan, Liu Chen (2025). 'Critical chatting' or 'casual cheating': How graduate EFL students utilize ChatGPT for academic writing. *Computer Assisted Language Learning*: 1–29. https://doi.org/10.1080/09588221.2025.2479141

Steinhoff, Torsten (2023). Künstliche Intelligenz als Ghostwriter, Writing Tutor und Writing Partner Zur Modellierung und Förderung von Schreibkompetenzen im Zeichen der Automatisierung und Hybridisierung der Kommunikation am Beispiel von ChatGPT. [Artificial intelligence as ghostwriter, writing tutor, and writing partner for modelling and fostering of writing skills in the context of automation and hybridization of communication using the example of ChatGPT]. Albrecht, Christian, Jörn Brüggemann, Tabea Kretschmann, Christel Meier, eds. *Personale und funktionale Bildung im Deutschunterricht: Theoretische, empirische und praxisbezogene Perspektiven* [Personal and Functional Education in German Classes: Theoretical, Empirical and Practical Perspectives]. Berlin: Metzler, 85–99. https://doi.org/10.1007/978-3-662-69640-8

Su, Yanfang, Yun Lin, Chun Lai (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing 57*: 100752. https://doi.org/10.1016/j.asw.2023.100752

Teng, Mark Feng (2024). A systematic review of ChatGPT for English as a foreign language writing: Opportunities, challenges, and recommendations. *International Journal of TESOL Studies 6*(3): 36–57. https://doi.org/10.58304/ijts.20240304

Warschauer, Mark, Waverly Tseng, Soobin Yim, Thomas Webster, Sharin Jacob, Qian Du, Tamara Tate (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *Journal of Second Language Writing 62*: 101071. https://doi.org/10.1016/j.jslw.2023.101071

Wei, Ling (2023). Artificial intelligence in language instruction: impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology 14*: 1261955. https://doi.org/10.3389/fpsyg.2023.1261955

Yoon, Choongil (2011). Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes 10*(3): 130–139. https://doi.org/10.1016/j.jeap.2011.03.003

Zhang, Chengming, Jessica Schießl, Lea Plößl, Florian Hofmann, Michaela Gläser-Zikuda (2023). Acceptance of artificial intelligence among pre-service teachers: A multigroup analysis. *International Journal of Educational Technology in Higher Education 20*(1): 49. https://doi.org/10.1186/s41239-023-00420-7

Zhao, Xin, Andrew Cox, Liang Cai (2024). ChatGPT and the digitisation of writing. *Humanities and Social Sciences Communications 11*: 482. https://doi.org/10.1057/s41599-024-02904-x

Authors' addresses:

Victoria Eibinger
University of Graz
Department of English Studies
Heinrichstraße 36/II, 8010 Graz
e-mail: victoria.eibinger@edu.uni-graz.at

Hannes Fromm
University of Graz
Department of English Studies
Heinrichstraße 36/II, 8010 Graz
e-mail: hannes.fromm@uni-graz.at

Margit Reitbauer
University of Graz
Department of English Studies
Heinrichstraße 36/II, 8010 Graz
e-mail: margit.reitbauer@uni-graz.at

Appendix: Codes used solely for student-AI conversations

Codes	Student-AI conversations
	n
Student prompt	81
Summary & meta-structure	432
Praise & motivation	105
AI-generated text	80
Sentences and phrases	65
Full text or paragraph	15
Feedback on the process of writing	39
Phatic & organization of chat	36
Total	**773**