Education and Information Technologies
https://doi.org/10.1007/s10639-024-12912-8

Comparing ChatGPT's correction and feedback comments with that of educators in the context of primary students' short essays written in English and Greek

Emmanuel Fokides[1] · **Eirini Peristeraki**[1]

Received: 9 April 2024 / Accepted: 12 July 2024

Abstract
This research analyzed the efficacy of ChatGPT as a tool for the correction and provision of feedback on primary school students' short essays written in both the English and Greek languages. The accuracy and qualitative aspects of ChatGPT-generated corrections and feedback were compared to that of educators. For the essays written in English, it was found that ChatGPT outperformed the educators both in terms of quantity and quality. It detected more mistakes, provided more detailed feedback, its focus was similar to that of educators, its orientation was more balanced, and it was more positive although more academic/formal in terms of style/tone. For the essays written in Greek, ChatGPT did not perform as well as educators did. Although it provided more detailed feedback and detected roughly the same number of mistakes, it incorrectly flagged as mistakes correctly written words and/or phrases. Moreover, compared to educators, it focused less on language mechanics and delivered less balanced feedback in terms of orientation. In terms of style/tone, there were no significant differences. When comparing ChatGPT's performance in English and Greek short essays, it was found that it performed better in the former language in both the quantitative and qualitative parameters that were examined. The implications of the above findings are also discussed.

Keywords ChatGPT-generated feedback · Instructor-provided feedback · Primary school · Short essays

✉ Emmanuel Fokides
fokides@aegean.gr

Eirini Peristeraki
eiriniperister@gmail.com

[1] University of the Aegean, Department of Primary Education, 1 Dimokratias Str., 85132 Rhodes, Greece

Published online: 27 July 2024

1 Introduction

Feedback (i.e., the provision of information with the aim of enhancing student performance), has been a focal point of educational research for numerous years. Studies have consistently underscored its significance in pedagogical contexts (e.g., Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006; Ryan et al., 2023). In addition, there is a considerable body of research examining its effectiveness in enhancing student performance (e.g., Henderson et al., 2019; Stern & Solomon, 2006; Weaver, 2006). Various factors have to be considered, including the timing, scope, and specificity of feedback relative to students' errors. Educators have to navigate these intricate decision-making processes, a task that has several challenges and does not guarantee successful outcomes (Crosthwaite et al., 2020).

However, the emergence of technological innovations has prompted a paradigm shift in education. Among these innovations, artificial intelligence (AI) and large language models (LLMs) afford educators the means to streamline tasks, such as text correction and grading. In this respect, AI can become instrumental in streamlining the feedback process and significantly alleviate the pedagogical burden on educators (Jia et al., 2022).

One of the most popular AIs is ChatGPT. Developed by OpenAI, this state-of-the-art natural language processing (NLP) system is based on the learning model known as the Generative Pre-trained Transformer (GPT). Drawing from an extensive collection of texts, it has been subjected to comprehensive training, allowing it not only to understand the context but also to produce responses with a high degree of coherence and a touch of creativity. Its versatility and adaptability render it a useful tool across a multitude of applications, ranging from the automation of customer service to the delivery of educational tutoring frameworks. It exhibited proficiency in detecting linguistic errors (Abdullayeva & Musayeva, 2023), but also in providing suggestions for improvements to text style and structure.

The discourse surrounding the quality of AI-generated feedback quality, particularly compared to instructor-generated one, remains active. For example, there is research suggesting that the effects of AI-generated feedback on students differ from those of educator-generated one (Han & Sari, 2022; Wang & Han, 2022). Conversely, other studies have indicated no significant differences (Escalante et al., 2023). The quality of AI-generated feedback has yielded contradictory findings. Some researchers asserted that educators offer superior feedback compared to AI systems like ChatGPT (e.g., Steiss et al., 2024), highlighting that without specific training in feedback generation, AI feedback is of limited value (e.g., Yoon et al., 2023). However, others have concluded that AI feedback is not only less time-consuming but also superior in quality in certain categories compared to expert feedback (Jacobsen & Weber, 2023). This has led to recommendations for a blended approach that leverages the strengths of both feedback forms (Escalante et al., 2023).

To contribute to the ongoing discourse, it was considered interesting to conduct a case study comparing the corrections and feedback provided by an AI, specifically ChatGPT, with those offered by educators on student assignments. For reasons that will be further elaborated in the coming sections, it was decided to focus on short

essays written by primary school students in both the English and Greek languages. The structure of this paper is organized as follows. Initially, a review of the literature explores the function of feedback, followed by the literature that discusses the application of AI in the realm of correction and feedback. Then, the study's research methodology is presented followed by the results. The subsequent discussion of the results concludes the work.

2 The role of feedback in the educational process

Hattie and Timperley (2007), defined feedback as information given by an agent (e.g. teachers, peers, and parents) concerning the performance or comprehension of a given subject matter. Its significance within the educational paradigm has been the subject of extensive scholarly discourse, with numerous studies examining its efficacy and the characteristics that encapsulate constructive and effective feedback modalities. Most academics support that feedback serves a pivotal role in enhancing the clarity with which students perceive their errors, thus, mitigating the risk of error perpetuation (Ai, 2017). There are several feedback types (e.g., formal, informal, summative, peer feedback, self-feedback, corrective, performance, explicit, and implicit) and several feedback classifications. For example, Hattie and Timperley (2007) in their seminal work, argued that there are four key feedback categories:

- Outcome-oriented feedback. This category refers to feedback that assesses the accuracy of an assignment. It informs learners about the correctness of their work, but also provides directions for improvement (e.g. "You must include X in your work").
- Process-focused feedback. This type focuses on the process of completing a task (e.g. "You need to process the task based on the theory we examined to make it more understandable").
- Self-regulation feedback. This category encompasses feedback that influences a student's ability to evaluate their performance and adjust their learning strategies accordingly. It also impacts their self-efficacy and, subsequently, their perception of achievement (e.g., "Your grasp of the theory is evident. Now, focus on its practical application within your work.").
- Personalized feedback. Finally, this type of feedback addresses the individual characteristics of the learner, potentially extending beyond the scope of the task itself. It affirms the student's efforts and can contribute positively to their motivation.

Other scholars have delineated feedback into three distinct types: positive, negative, and non-corrective (Lin et al., 2022). Positive feedback commends student achievements and motivates them to build upon their success. Conversely, negative feedback offers a critique of student performance, intending to rectify ineffective strategies (Stern & Solomon, 2006; Weaver, 2006). Effective feedback has to encompass both commendation and critique, fostering a balanced and constructive learning environment. Non-corrective feedback is designed to bolster student proficiency in subsequent, similar tasks (Ryan et al., 2021). Moreover, feedback may

entail comments that either verify responses or stimulate further processing (Shute, 2008). Verification feedback ascertains the correctness of responses, while processing feedback delves into the rationale behind those responses, with the latter being deemed more effective in promoting deeper learning (Hattie & Timperley, 2007; Shute, 2008). It is also worth noting that feedback addressing incorrect responses was considered more impactful in enhancing learning than feedback affirming correct responses (Clariana et al., 2000).

The style in which feedback is presented significantly influences student outcomes. Demonstrating kindness and exercising discretion are pivotal in bolstering the performance of students, particularly those facing learning challenges (McLaren et al., 2011). While feedback must incorporate constructive criticism, it must also be dispensed in a fashion that preserves the equilibrium of the student–teacher relationship (Wang et al., 2008). Empirical studies have shown that the delivery of praise requires careful consideration, as it may diminish students' willingness to accept their errors (Maclellan, 2005). General praise, such as "Good effort," fails to convey substantive feedback and may be counterproductive. Hattie and Timperley (2007) argued that when students receive nondescript praise following a successful task or neutral comments after an unsuccessful attempt, they might perceive this as an indication of the teacher's disinterest. To avoid this, it is recommended that feedback should be specific and articulate the aspects of the work a student has executed proficiently (e.g., "The structure of your work is commendable") (Lin et al., 2022). However feedback must be customized to align with each student's unique requirements, a sentiment echoed by scholars who advocated that one should be mindful of the diverse interpretations that different students may have in response to the same feedback (Henderson et al., 2019). Furthermore, it is crucial to take into consideration the distinct linguistic and cultural backgrounds of students, as this awareness is integral to furnishing quality feedback (Osakwe et al., 2022).

Research indicated that feedback provided in the form of questions is more effective in enhancing students' self-regulatory abilities as compared to statements (Hattie, 2012). Despite this evidence, it appears that this strategy is not widely implemented by educators. Moreover, the timeliness of feedback is critical; prompt feedback is more beneficial than a delayed one. Carless et al. (2011) found that students are prone to disregard feedback given after the completion of an assignment, due to a disconnection with the relevant content. Consequently, delayed feedback can act as a deterrent to student engagement (Poulos & Mahony, 2008).

The complexity of crafting effective feedback is amplified by the diverse and extensive range of possible responses. Due to this complexity, researchers suggested the provision of personalizing feedback, that acknowledges individual learning styles and can positively influence students' perceptions, the acquisition of knowledge, and their interactions with educators (Vasilyeva et al., 2007).

All in all, feedback is an intricate construct, designed to enlighten students regarding their perceptions, study habits, and dispositions, and how these elements coalesce to impact their academic outcomes (Boud & Falchikov, 2007; Henderson et al., 2019). In addition, feedback illuminates potential areas for enhancement (Parikh et al., 2001). Yet, providing effective feedback seems to be a challenging task for educators (Matcha et al., 2019). They frequently default to offering superficial

remarks that provide inadequate inspiration or direction for students (Weaver, 2006). Nevertheless, it is crucial to recognize that the impact of feedback is not uniform; rather, it is modulated by a multitude of individual and situational variables (Narciss et al., 2014).

As far as feedback related to young students' writing skills is concerned, there are some additional issues that need to be considered, given that, despite its significance, writing is an inherently complex task. As a result, many primary school students face difficulties (Graham, 2018). Moreover, younger students often spend minimal time on planning and revising their texts, resulting in shorter writings with numerous errors (Troia, 2006). Given these challenges, merely assigning grades to students' writings is insufficient to enhance the quality of their compositions (Parr & Timperley, 2010). On the other hand, feedback on writing has proven to be a powerful tool for increasing both the quantity and quality of writing among elementary and middle school students (Barrett et al., 2020). Besides, feedback that empowers young writers and enhances their understanding is essential for preventing literacy problems, which are best addressed in the early years of schooling (Graham et al., 2015a, b). Effective feedback on writing can focus on various aspects such as word choice, ideas, and organization (Schirmer & Bailey, 2000). These, alongside the provision of clear goals and steps for improvement can significantly foster students' writing skills (Brookhart, 2008). Research indicated that adults, such as teachers, are the most effective providers of feedback regarding students' text quality compared to peers and self-assessment (Graham et al., 2015a, b). However, young and struggling students may not receive adequate feedback by their teachers (Han & Xu, 2020) or not fully comprehend it or even perceive it negatively (Marrs et al., 2016). Additionally, students who find writing challenging often do not make good use of the feedback they were provided (Altstaedter & Doolittle, 2014).

3 The utilization of automated writing evaluation systems and artificial intelligence in task correction and feedback provision for students' writing

The scholarly work surrounding the implementation of AI in educational contexts is rapidly expanding. A degree of skepticism exists, as some argue that current evidence supporting its utility and alignment with established pedagogical principles remains inadequate (Zawacki-Richter et al., 2019). Others call for a more systematic application of AI in educational practices to ensure its efficacy (Gong et al., 2020). In contrast to the skeptics, empirical studies indicated that AI fostered positive attitudes among students regarding its usage (Chiu et al., 2022), while educators have experienced enhanced autonomy and an augmentation of their role (Chatterjee & Bhattacharjee, 2020).

Concentrating on the domain of automated task correction and grading of written tasks, the role of AI has been scrutinized through various research endeavors. For example, Grammarly, an AI-powered English language writing aid, was found to significantly bolster student academic outcomes by reducing the frequency of spelling and grammatical errors (Sanosi, 2022). Grammarly's effects on various

dimensions of writing skills, such as task achievement, coherence and cohesion, lexicon, and grammatical accuracy were also the focus of another study (Wei et al., 2023). It was revealed that the group of students who utilized Grammarly demonstrated superior performance across all aspects of writing skills. Yet, it was also evident that the learners' initial proficiency levels significantly influenced their subsequent writing outcomes. The e-rater engine, an integral component of the Criterion writing service, has demonstrated a 90% correction accuracy rate, coupled with a 0.76 correlation with human grading outcomes (Azmi et al., 2019). The efficacy of Juku, an automated writing evaluation system employed in China for English language instruction, has received positive feedback from both students and teachers, although there were instances of inadequacy in properly evaluating written tasks in terms of structure, coherence, and content (Lu, 2019). Chang et al., 2024) reviewed the evidence currently available on the performance of AIs, across a variety of tasks. They reported that, for writing tasks, AIs performed consistently across different genres, including argumentative and creative writing. There was also evidence that AIs can successfully evaluate text quality without the use of reference texts and that they outperform most existing automated algorithms (Chen et al., 2023). Yet, others have proposed that AI's proficiency in error correction is limited to specific categories (Fitria, 2021).

Overall, it seems that automated text scoring systems promote consistency and objectivity in assessment practices (Hussein et al., 2019). Meanwhile, they can address grammatical and syntactic errors, thereby streamlining the overall pedagogical process (Link et al., 2022). Then again, text evaluation requires taking into account many parameters such as relevance to the question, content, and coherence, areas in which AIs need to be improved (Ramesh & Sanampudi, 2022). Moreover, there is a challenge associated with the evaluation of creative writing and the expression of original ideas, a realm where LLMs have not yet matched human evaluators. This limitation is further pronounced due to the prevailing deficit in their linguistic diversity, a critical issue considering the variable nature of text structures across different languages. Consequently, LLMs may not consistently deliver a level of correction on par with a meticulous and diligent human rater (Murphy, 2019; Wang et al., 2022).

Besides correction, the provision of automatically generated feedback has been at the heart of several studies. For example, Jia et al. (2022) utilized the Insta-Reviewer platform and concluded that it can indeed generate feedback comparable to that of educators. Another study, focusing on the evolution of written expression, has established that such feedback can significantly bolster self-regulatory practices and contribute to the refinement of writing skills (Osawa, 2023). Furthermore, Cavalcanti et al. (2021) literature review has revealed that while a majority of the reviewed studies (65.07%) concluded that automated feedback catalyzed enhancements in student performance, a substantial proportion (46.03%) failed to affirm its efficacy in diminishing the workload of educators. In addition, in an overwhelming 82.53% of the studies, there was no evidence of the superiority of teacher-generated feedback over its AI-generated counterpart.

Then again, there are drawbacks; text overcorrection, the potential for inducing cognitive overload in students, and the delivery of inadequate explanations are

among the documented concerns (Barrot, 2023). Others postulated that automated feedback falls short when compared to that provided by humans, attributed to a lack of tailored recommendations and, on occasion, the provision of inaccurate information. In addition, repeated sentences were observed in scenarios where there was a need for differentiated feedback (Jia et al., 2022). There were also cases in which although students expressed favorable perceptions of such feedback, this was not translated into academic progress (Huang & Renandya, 2020). Investigations have illuminated a concerning trend where students either neglected to verify the validity of the feedback they received or became excessively reliant upon it, leading to suboptimal academic outcomes (Koltovskaia, 2020). Others suggested that AI systems must evolve to dispense personalized feedback, attuned to the individual's personality traits and linguistic competencies (Conati et al., 2021). Moreover, the suggestion has been made that AI-generated feedback should adopt a less directive approach to foster self-motivation and self-correction, as these are beneficial particularly for students who have strong learning motivation or limited language abilities (Liang et al., 2023).

It is important to note that studies examining the use of AI in the context of written tasks of young learners, specifically those in primary education, are limited and do not directly address the quality of corrections or feedback provided by these systems. For instance, Jang et al. (2023) investigated whether an AI-infused feedback mechanism, which provided diagnostic scaffolding to young writers, could support their metacognitive control. Additionally, the work of Ali et al. (2023) demonstrated that the use of AI, specifically ChatGPT, motivated primary school students to develop their reading and writing skills, although its impact on listening and speaking skills was found to be neutral.

Finally, there are studies comparatively examining automated and teacher feedback, reaching interesting conclusions. Tian and Zhou (2020) supported that despite automated feedback being usually more extensive than teacher feedback, students tend to ignore it. In addition, it seems teacher feedback has a positive effect on students' psychological state, while automated feedback seems to lead to higher language proficiency performances, as it focuses more on grammar and syntax (Han & Sari, 2022; Wang & Han, 2022). Other researchers suggested reconciling the benefits of teachers' feedback with that of AI; teachers could strategically leverage automated feedback to better assess learning needs (Di Placito & Mortensen, 2023).

4 ChatGPT

ChatGPT is one of the most popular AIs. The literature related to its applications is growing rapidly, while the views are often conflicting indicative of the intense debate caused by the advent of LLMs. For example, Aydın and Karaarslan (2022) suggested that the process of writing an academic paper will require less human effort, allowing scientists to focus on their research more efficiently, while Floridi and Chiriatti (2020) characterized ChatGPT as uninformed science fiction. Others pointed out the weaknesses, risks, limitations, and social implications of using ChatGPT (e.g., Borji, 2023).

ChatGPT can be used in many educational settings, such as language learning (Athanassopoulos et al., 2023), feedback on student assignments (Fuchs, 2023), but also the consolidation of a clearer, fairer grading system (Altamimi, 2023; Glaser, 2023). Its utilization within the realm of tertiary education has garnered considerable scholarly attention. On one hand, there are concerns regarding its potential to facilitate plagiarism, erode the credibility of written assessments, compromise contextual understanding, undermine academic integrity, and diminish the development of cognitive skills (e.g., Farrokhnia et al., 2023). On the other hand, scholars recognized its contributions such as the provision of automated correction and feedback for student assignments (e.g., Mizumoto & Eguchi, 2023). Research has also been conducted to explore the pedagogical implications of ChatGPT at the school level, for example, for enhancing specific skills (Woo et al., 2023) and for augmenting the overall learning experience (Zhang & Tur, 2023).

The spectrum of errors detected by ChatGPT extended across numerous linguistic dimensions, encompassing grammatical, lexical, spelling, and punctuation-related inaccuracies (Abdullayeva & Musayeva, 2023). In an interesting study, the authors used ChatGPT-3 to automatically score 12,100 essays written in English from individuals speaking 11 different languages (Mizumoto & Eguchi, 2023). They found that it has a certain level of accuracy and reliability, providing valuable support for human evaluations. Furthermore, they argued that by utilizing linguistic features the accuracy of the scoring could be enhanced. In the study conducted by Fang et al. (2023), ChatGPT's performance in correcting grammatical errors across three different languages was assessed. The results and subsequent human evaluations demonstrated that ChatGPT possesses exceptional error detection capabilities. However, it was also observed that ChatGPT tends to over-correct and does not consistently adhere to the principle of minimal edits.

In the realm of feedback provision, there is literature suggesting that ChatGPT has the potential to facilitate feedback practices (e.g., Katz et al., 2023). However, this body of literature is currently limited. For example, it exhibited the capacity to offer recommendations aimed at the enhancement of stylistic and structural elements within a text, potentially elevating its readability. A noteworthy finding was that feedback from ChatGPT typically includes a task summary and its rationale, in contrast to the predominance of brief evaluative remarks found in teacher feedback (Dai et al., 2023). This characteristic positions ChatGPT as a tool with significant potential to advance the feedback process across a variety of educational assessments, including but not limited to multiple-choice, essay, fill-in-the-blanks, and short-answer formats (González-Calatayud et al., 2021; Sein, 2022).

There are also studies comparatively examining the differences in feedback generated by ChatGPT and humans. For instance, Banihashem et al. (2024) found that ChatGPT provided more descriptive feedback, offering detailed information on the essay's writing style. In contrast, peer feedback focused on identifying problems within the essay. The findings suggested a complementary role for ChatGPT and students in the feedback process. Moreover, there was no significant relationship between the quality of the essays and the quality of the feedback provided by either ChatGPT or peers, implying that the quality of the essays did not influence the feedback's quality. The findings in the study by Steiss et al. (2024) were markedly

different; human raters consistently excelled in delivering high-quality feedback across all categories except for criteria-based feedback. Furthermore, the research highlighted distinctive variations in feedback quality between AI and human raters, depending on the quality of the essays. Jansen et al. (2024) conducted an evaluation on the effectiveness of feedback for students' argumentative writing, comparing the performance of ChatGPT-3.5 with that of expert-generated feedback. Their findings revealed that AI-generated feedback was deemed useful in 59% of cases, in contrast to the 88% usefulness rating for expert feedback. Notably, 23% of participants expressed a preference for providing AI-generated feedback to students. Moreover, others noted that, without training, ChatGPT is of limited value, especially in the context of providing feedback related to the coherence and cohesion of essays (Yoon et al., 2023).

It should also be acknowledged that the ultimate endorsement of the feedback generated by ChatGPT remains within the purview of educators (O'Cain et al., 2023). Furthermore, the reliability of the outcomes heavily relies on both the prompt (i.e., a set of directives for ChatGPT elucidating the task at hand and the manner of its execution) and the input text (De Winter et al., 2023). Finally, it is worth mentioning that ChatGPT's processing of the linguistic diversity and subtle nuances inherent in human language is not uniformly accurate, a limitation that could affect feedback quality (Fuchs, 2023).

5 Method

What can be concluded from the above presentation of the literature, is that the debate surrounding the capabilities and limitations of AI-generated corrections and feedback is still unresolved. This ongoing debate stems from the multifaceted nature of feedback, which, to be deemed effective and comprehensive, demands consideration of several parameters. In addition, while AI-generated corrections, grading, and feedback have demonstrated proficiency in particular domains, they exhibited deficiencies, notably in the realm of linguistic diversity. Moreover, the literature comparatively examining AI and human-generated corrections and feedback is rather limited. In light of these considerations, a study was designed and implemented to provide answers to the following research questions:

- RQ1a-c. How do the corrections made by ChatGPT on primary school students' short essays written in English and Greek compare to those made by educators in terms of (a) grammatical, syntactical, and spelling mistakes found, (b) the feedback comments (henceforth referred as feedback) provided to them in terms of focus, orientation, and style/tone, and (c) grading?
- RQ2. Does ChatGPT exhibit any variance in its approach to providing corrections and feedback for texts in English as compared to those in Greek?

Concerning the above RQs, some points need clarification. Firstly, ChatGPT was chosen because of its popularity and because there is some evidence suggesting that it can perform better than other AIs in certain tasks (e.g., Borji & Mohammadian, 2023;

Lossio-Ventura et al., 2024; Wu et al., 2023; Zhong et al., 2023). Secondly, it was considered important to focus on texts sourced from primary school learners. This decision stemmed from the recognition that primary education fundamentally differs from other educational levels. Primarily, the curriculum at the primary level is focused on the development of basic literacy and numeracy skills, centering on core disciplines such as reading, writing, and maths, that are fundamental for students' success in all subsequent academic pursuits. Therefore, the quality of feedback and the accuracy of corrections that students receive during this period are of paramount importance.

Thirdly, in the context of primary education, educators typically furnish comprehensive feedback comments on assignments related to language skills, which often take the form of written essays. These essays play a pivotal role in fostering the development of students' linguistic capabilities. Indeed, in Greece, students are tasked with the production of essays as an integral component of their Greek language or EFL courses. This pedagogical exercise may take the form of homework or an in-class activity. When assigned as homework, the essays demonstrate a coherent structure and contain minimal errors, attributable to the ample time provided for their composition. Furthermore, this setting allows for parental guidance or the intervention of private tutors to facilitate revisions. In contrast, when essay writing occurs within the classroom, students face temporal constraints, as they typically have 20 to 30 min to complete the task. As a result, their essays are short (just a couple of paragraphs) and not so well-rehearsed. In this respect, it can be argued that such unassisted texts offer a more authentic reflection of students' language proficiency, allowing educators to provide targeted corrections and substantive feedback. For that matter, the study elected to focus on short essays derived from the latter scenario.

In addition, it was decided to compare the proficiency of ChatGPT's corrective and feedback mechanisms in English (a language in which one can argue that it was thoroughly trained) with those in Greek, where its training was presumed to be less comprehensive. This investigative necessity was not only grounded on the literature presented in the preceding sections but also surfaced during the preliminary phase of the study. Upon experimenting with various prompts targeting text correction and feedback, a differentiation in performance was noted with ChatGPT v3.5 (the freely accessible version of the tool). In the English texts, the AI displayed a heightened accuracy in error detection and delivered detailed feedback. In contrast, with Greek inputs, ChatGPT v3.5 failed to identify errors, erroneously flagged correct words and passages as incorrect, and the overall quality of its responses bore childish grammatical and syntactical mistakes.

Finally, feedback, as established in the previous sections, is inherently a multifaceted concept. In the pursuit of a more comprehensive understanding, this study investigated three principal feedback components: (i) focus, which pertains to the particular facets of the students' writings that the feedback addressed, (ii) orientation, for which the classification of Hattie and Timperley (2007) served as a guiding framework, and (iii) style/tone, ascertained through the linguistic choices employed within the feedback.

To examine the RQs, the study employed a mixed-methods research approach, as will be presented in the section "Procedures and data preparation."

5.1 Participants

An invitation to participate was issued to primary school and EFL teachers via social media platforms, including details for the objectives and methods of the study. Consequently, 20 individuals expressed their willingness to participate, allocated into two groups of equal size (ten primary school teachers and ten EFL instructors who also fulfilled their teaching roles in primary schools). All participants had more than ten years of teaching experience ($M=14.75$, $SD=3.80$), their age range was between 37 and 51 ($M=43.15$, $SD=4.39$), and most were females ($n=13$). All participants were native Greeks. Ethical clearance for the project was granted by the university's ethical committee, and each participant provided informed consent prior to their engagement in the study.

5.2 Materials

Several teachers, not participating in the study, were asked to supply genuine student short essays. Out of these, 20 Greek and 20 English texts were randomly selected, belonging to students eight to eleven years old, with each text averaging approximately 200 words in length. As far as the English texts were concerned, the language level according to the Common European Framework of Reference was A2 to B1. All personal data (e.g., students' names, dates, and schools) were removed. Because the texts were handwritten and the subsequent requirement for analysis by ChatGPT, the texts were transcribed verbatim into a digital document.

5.3 Instruments

As mentioned at the begging of the "Method" section, ChatGPT v.3.5 was deficient in its ability to accurately correct texts written in Greek. In response to this limitation, ChatGPT v.4 turbo (preview 1106) was employed. As v.4 represented the latest version available during the time the study was conducted, it was assumed that its performance would surpass that of its predecessor. Given that access to this version necessitated a subscription, it was decided to use GhostWriter (https://www.ghostwriter-ai.com/), a paid add-on for Microsoft Word, which allows access to this version.

To gather data from ChatGPT related to the correction of students' texts (RQ1a-c), it was imperative to formulate a thorough prompt, as the efficacy of ChatGPT in accomplishing any task relies on the clarity of the task's outline and the specificity of the instructions provided. To this end, several prompts underwent rigorous testing using a small set of students' essays. The outcomes were subjected to validation, leading to subsequent refinements in the prompt. The prompt that was finally used for correcting students' English and Greek texts was the following:

> "I want you to act as an EFL teacher [Greek primary school teacher]. Below there is a text written by one of your students. Since the text is in English [Greek], use all the knowledge you have about the English [Greek] language. Make a list of the grammar mistakes, syntax mistakes, and spelling mistakes you detected. [Keep in mind that, in the Greek language, the incorrect placement of a stress mark or lack of a stress mark in a word that should have one, is considered a spelling mis-

take]. Do not list the same mistake twice. Do not group similar mistakes, list them separately. Also, in a second list, I want you to record comments regarding the strengths of the text, that is the positive elements you found in its grammar, spelling, syntax, expression, structure, and content. In a third list, I want you to record your comments on the text regarding the weaknesses you found in its grammar, spelling, syntax, expression, structure, and content. Then, write summative comments as feedback to the student, using as a basis the mistakes you found, as well as the comments you made about the strengths and weaknesses of the text. On the basis of the severity of the mistakes you found, as well as on the basis of the comments you made and the feedback you provided, assess the text on a 1-10 grading scale. Do not justify your grade. Do not re-write the text with corrections. The text I want you to check is the following: '…'"

Comments on the prompt.

- The initial directive, "I want you to act as an EFL teacher [Greek primary school teacher]," mandated that ChatGPT adopt the persona of an EFL educator or a primary school teacher, necessitating behavior befitting said roles. This type of directive is prevalent in contexts where ChatGPT is expected to demonstrate expertise within a specific domain.
- Subsequently, the instruction, "Since the text is in English [Greek], use all the knowledge you have about the English [Greek] language," served as a mandate for ChatGPT to engage its full linguistic capabilities in English or Greek.
- To avoid redundancy, the clause, "Do not list the same mistake twice," was integrated. This was particularly pertinent as students often repeat the same mistake within their writings.
- Moreover, the condition, "Do not group similar mistakes, list them separately," addressed an observed tendency of ChatGPT to present errors as collective grammatical or syntactical issues without specification.
- The guideline, "Then, write summative comments as feedback to the student, using as a basis…," provided ChatGPT with a framework for constructing comprehensive feedback. Please note that it was purposively not instructed to provide a specific type/category of feedback to avoid narrowing the scope of its response.
- In the realm of assessment, the directive, "On the basis of the severity of the mistakes you found, …assess the text on a 1–10 grading scale," delineated the criteria ChatGPT should employ in evaluating student texts. More detailed instructions could have been given (e.g., a specific number of points could have been allocated for each type of mistake or the positive/negative aspects of the text). Yet, no rule of thumb could have been followed. Moreover, a text might have many mistakes but the structure and content might be good or vice versa. As teachers deal with each case on an individualized basis, it was decided not to be very specific on how ChatGPT graded the texts.
- The mandate, "Do not justify your grade…," was based on the assumption that the feedback provided sufficient explanation for the assigned grade.

- Lastly, the instruction, "Do not re-write the text with corrections…," was included to prevent ChatGPT from generating revised versions of student texts, a practice that had been noted in initial prompt formulations.

In addition to basic functionalities, GhostWriter provides a suite of customization options for modulating various dimensions of ChatGPT's response, without the need to include them in the prompt. To more closely mirror the response of an educator, the decision was made to adjust the response style to "instructive." Moreover, the creativity parameter was increased to 0.8 from the default 0.7 to infuse the responses with a heightened level of inventiveness.

For collecting data from the educators, a total of 40 Google Forms were used, equal to the number of texts. Within each form, participants were presented with the text in need of correction, instructions on how to correct it (identical to the instructions given to ChatGPT), followed by a sequentially organized array of fields. Namely, there was a field for recording the mistakes they found, one for recording their positive comments, another one for recording their negative ones, followed by a field for providing their overall feedback. In the final field, the respondents could provide their grade.

5.4 Procedures and data preparation

Educators were granted access to Google Forms and were tasked with the correction of the short essays (20 in Greek or 20 in English, depending on their expertise) within 24 h. Although, in reality, educators need less time to assess and correct students' assignments, the necessity of performing these corrections in a digital format was identified as a contributing factor to procedural delays; therefore, it was decided to provide them a day to complete this task. The compiled results were exported to a spreadsheet. The authors corrected students' essays using ChatGPT, with the outcomes being integrated into the previously utilized spreadsheet. During this phase, ChatGPT's outputs were reviewed to verify the adherence of its responses to the specific guidelines it was given and to prevent instances where ChatGPT might furnish off-topic or wholly inaccurate responses.

In the quantitative component of the data analyses, the mistakes identified by both the educators and ChatGPT were enumerated. Additionally, the word count of the feedback provided on each text was calculated. As noted in a previous section, ChatGPT had, on occasion, mistakenly identified accurate words and textual fragments as erroneous. Instances of such inaccuracies were also enumerated (for both the English and Greek texts, and both the educators and ChatGPT). The dataset which included 40 responses generated by ChatGPT (20 texts per language X 2 languages) and 400 responses provided by participating educators (10 educators per language X 2 languages X 20 texts per language) was exported to SPSS 29 for statistical analysis.

The qualitative part involved three stages. In the first, NVivo v.1.7 was employed as a tool to perform the thematic analyses of the feedback of both the educators and ChatGPT, in order to determine their focus. In addition, a

content analysis was undertaken, again using NVivo, to ascertain the orientation characterizing the feedback provided by educators and ChatGPT. As stated earlier, this analysis was anchored in the framework of Hattie and Timperley (2007), which delineates feedback into four distinct types: outcome-oriented, process-focused, self-regulation, and personalized feedback. Both stages were conducted by a pair of skilled coders to mitigate the influence of subjectivity and to bolster the overall reliability and credibility of the data interpretation process. These individuals underwent extensive training across multiple sessions utilizing a representative subset of the dataset. This training process continued until they reached a high degree of intercoder reliability, as reflected by Cohen's kappa coefficient of 0.81. The results of these two stages of the qualitative analyses presented in the coming section, were derived from averaging the findings of both coders.

In the third stage, LIWC-22 (https://www.liwc.app/) was employed, to perform lexical analysis of the educators' and ChatGPT's feedback, in order to determine its style/tone. The Linguistic Inquiry and Word Count (LIWC) tool analyzes texts by evaluating each constituent word against a dictionary that includes categories of words that possess psychological significance, including but not limited to, emotions, cognitive processes, and social language (Pennebaker et al., 2022). The occurrence of these lexical categories offers insights into the author's psychological state or condition. As effective feedback is also related to linguistic style, features extracted by LIWC were of relevance to the study, as they elucidated the structural and emotional attributes of the feedback. Although LIWC-22 can analyze over 100 dimensions, four summary measures were selected:

- Analytical thinking quantifies the extent to which an individual's language reflects structured, logical, and hierarchical thought processes. High scores are indicative of language that aligns with academic performance and reasoning capabilities. Conversely, a score falling below 50 suggests a warmer, more approachable, and friendly linguistic style.
- Clout assesses the demonstration of social dominance, leadership, and confidence. This measure provides insights into the writer's social influence and perceived authority.
- Authenticity measures the degree of genuineness. Low Authenticity scores (<50) typically indicate prepared texts (e.g., speeches) or texts in which individuals exercise caution and socially reserved behavior. In contrast, high Authenticity scores are often found in unguarded discourse, such as casual dialogues among friends.
- Emotional tone consolidates the spectrum of affective expression into a singular metric. Higher values correspond to a more positive sentiment, whereas values below the midpoint are suggestive of a predominately negative emotional tone.

Please note that the comments related to the strengths and weaknesses of students' texts were not analyzed. Their purpose was to provide a basis for both the educators and ChatGPT to formulate their feedback.

6 Results

6.1 Quantitative analyses

Table 1 presents descriptive statistics for the study's variables. Prior to conducting One-way ANOVA tests, which aimed to discern disparities in the corrections and feedback provided by educators and ChatGPT, as well as between the two languages within the educators' and ChatGPT's corrections, an assessment of the data's suitability for this statistical method was undertaken. As a non-normal distribution of the dataset and a breach of the homogeneity of variance were observed, the Mann–Whitney U test was used, a robust non-parametric alternative. The results are presented in Tables 2 and 3.

Table 1 Descriptive statistics for the study's variables

Language		Educators				ChatGPT			
		min	*max*	*M*	*SD*	*min*	*max*	*M*	*SD*
English	Mistakes correct	4	38	14.79	5.94	7	30	17.90	5.76
	Mistakes wrong	0	0	0.00	0.00	0	0	0.00	0.00
	Feedback	2	141	38.69	18.69	121	268	189.25	37.79
	Grade	1	10	5.89	1.89	3	6	4.40	0.75
Greek	Mistakes correct	2	61	23.01	13.67	6	35	17.95	7.92
	Mistakes wrong	0	0	0.00	0.00	0	5	0.90	1.17
	Feedback	7	330	68.38	45.71	85	200	127.15	30.35
	Grade	2	10	6.01	1.75	4	6	5.55	0.69

Table 2 Comparison between the educators and ChatGPT

Language		Mistakes correct	Mistakes wrong	Feedback	Grade
English	Mean rank educators	107.42	110.50	100.51	115.81
	Mean rank ChatGPT	141.28	110.50	210.45	57.45
	Mann–Whitney U	1384.50	2000.00	1.00	939.00
	z	-2.271	0.000	-7.367	-3.959
	p	0.023	1.000	<0.001	<0.001
	η^2	0.023	0.000	0.247	0.071
	Effect size interpretation	small-medium	-	very large	medium-large
Greek	Mean rank educators	112.20	104.50	102.86	112.27
	Mean rank ChatGPT	93.53	170.50	186.95	92.78
	Mann–Whitney U	1660.50	800.00	471.00	1645.50
	z	-1.252	-11.237	-5.634	-1.326
	p	0.211	<0.001	<0.001	0.185
	η^2	0.007	0.574	0.144	0.008
	Effect size interpretation	-	very large	large	-

For the interpretation of the effect sizes, the following cutoff values were used: 0.010-small, 0.059-medium, 0.138 or higher-large (Cohen, 2013)

Table 3 Comparison between the English and Greek texts

Grader		Mistakes correct	Mistakes wrong	Feedback	Grade
Educators	Mean rank English	163.60	200.50	153.98	197.32
	Mean rank Greek	237.41	200.50	247.02	203.68
	Mann–Whitney U	12,619.000	20,000.000	10,696.00	19,364.00
	z	-6.390	0.000	-8.049	-0.557
	p	<0.001	1.000	<0.001	0.577
	η^2	0.102	0.000	0.162	0.001
	Effect size interpretation	medium-large	-	large	-
ChatGPT	Mean rank English	20.95	14.50	28.85	13.58
	Mean rank Greek	20.05	26.50	12.15	27.43
	Mann–Whitney U	191.00	80.00	33.00	61.500
	z	-0.244	-4.039	-4.519	-3.973
	p	0.820	<0.001	<0.001	<0.001
	η^2	0.001	0.408	0.511	0.395
	Effect size interpretation	-	very large	very large	very large

6.2 Qualitative analyses

6.2.1 Feedback's focus

During the thematic analyses of the data regarding the focus of the feedback, the following themes emerged for both the English and Greek texts and for both educators and ChatGPT:

- Language mechanics. This theme encompassed the technical aspects of writing that contribute to its overall quality, including codes such as comments on grammar, spelling, and punctuation. It reflected the students' grasp of language rules and their ability to apply them correctly.
- Effectiveness of expression. This theme dealt with the precision and clarity of expression in the students' writing. It involved codes such as accurate word choice, specificity in detail, expression of students' emotions and experiences, and maintaining a clear message throughout the text.
- Content structure. This theme highlighted the importance of organizing thoughts in a logical and reader-friendly manner. It included codes such as text coherence and paragraph and text structure.

Table 4 Comparison between educators' and ChatGPT's feedback focus in both languages

Theme	English (%)		Greek (%)	
	Educators	ChatGPT	Educators	ChatGPT
Language mechanics	49.9	43.6	31.2	23.3
Expression	18.4	20.6	23.3	17.8
Content structure	8.5	4.7	7.3	10
Support and guidance	23.1	31.2	38.2	42.5

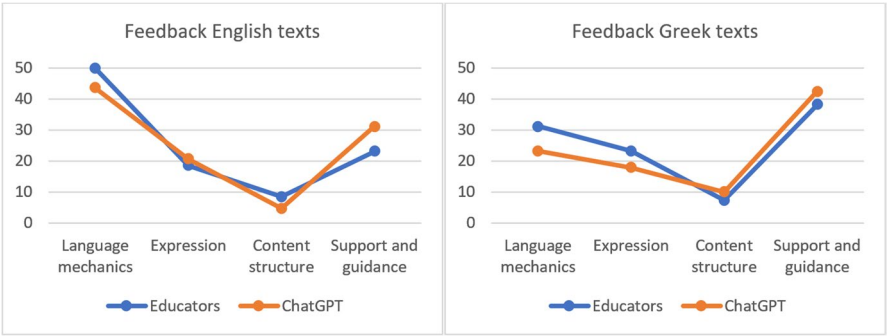

Fig. 1 Comparison between educators' and ChatGPT's feedback focus

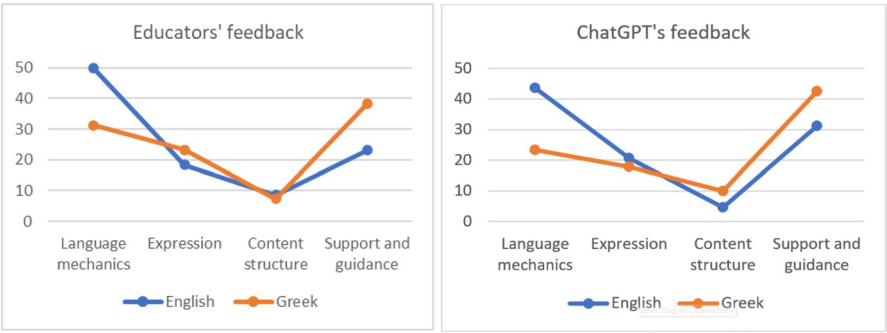

Fig. 2 The educators' and ChatGPT's feedback focus on the English and Greek texts

- Support and guidance. This theme centered on the teacher's role in providing support, constructive feedback, and encouragement for personal growth.

Please note that for the clarity of the results' presentation, the full set of themes and codes that have been developed can be found in the Appendix (Tables 8, 9, 10 and 11). Table 4 presents the percentages of the themes found in the educators' and ChatGPT's feedback, while Figs. 1 and 2 offer a comparative presentation of the percentages of the themes about the corrections made by educators and ChatGPT across the English and Greek texts.

6.2.2 Feedback's orientation

Tables 5 and 6, along with Figs. 3 and 4, present the findings of the feedback's content analysis for determining its orientation using the classification suggested by Hattie and Timperley (2007). The calculation of percentages adhered to the formula: instances where evidence of a category's presence was identified divided by the total number of texts wherein feedback was rendered. Please note that certain feedback entries touched upon more than one feedback category; thus, the total count of categorized instances surpassed the actual number of texts for which feedback was provided.

Table 5 Orientation of the educators' feedback in the English and Greek texts

Language	Category	*n*	%	Example quotes
English	outcome-oriented	182	91.0	"Study more the prepositions in grammar!"
	process-focused	56	28.0	"Be careful with subject personal pronouns in English we must use them almost always before our verbs!"
	self-regulation	58	29.0	"Overall a great essay with a beginning, main part, and a conclusion."
	personalized	68	34.0	"Good job [student's name]!
Greek	outcome-oriented	180	90.0	"Watch your spelling!"
	process-focused	91	45.5	"When you finish writing, re-read…"
	self-regulation	88	44.0	"Apply what we learned when writing your essay…"
	personalized	155	75.5	"You had some very nice ideas…"

Table 6 Orientation of ChatGPT's feedback in the English and Greek texts

Language	Category	*n*	%	Example quotes
English	outcome-oriented	20	100.0	"Pay attention to verb tenses and ensure that the subject and verb agree in your sentences."
	process-focused	19	95.0	"When you're writing about past events, make sure to use the past simple tense."
	self-regulation	16	80.0	"Keep practicing your English, and don't hesitate to look up words or rules if you're unsure."
	personalized	20	100.0	"Your positive energy is definitely a strong point in your writing."
Greek	outcome-oriented	20	100.0	"Watch out for spelling mistakes, you had quite a lot."
	process-focused	0	0.0	-
	self-regulation	0	0.0	-
	personalized	20	100.0	"Congratulations on your effort to write this text."

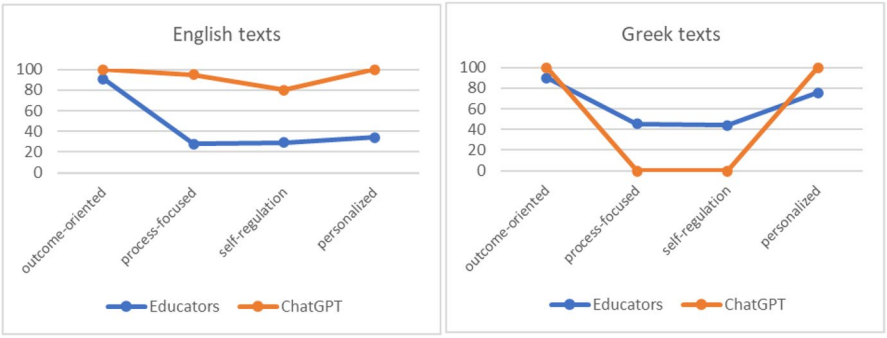

Fig. 3 Comparison between the educators' and ChatGPT's feedback orientation

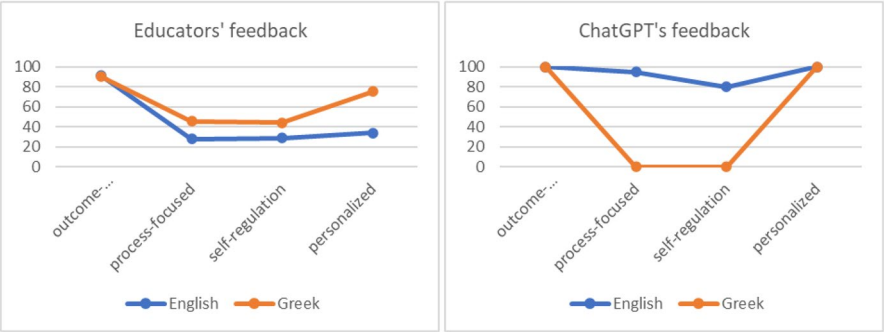

Fig. 4 The educators' and ChatGPT's feedback orientation in the English and Greek texts

6.2.3 Feedback's tone/style

Table 7, as well as Figs. 5 and 6 present the results of the lexical analyses using LIWC-22. For the interpretation of the results, the following criteria apply:

- Analytical thinking. Values above 50 indicate an academic/formal style, while values below 50 indicate a warm and friendly tone.
- Clout. Values above 50 reflect authority and confidence.

Table 7 The results of the lexical analyses

Language	Feedback provider	Analytical thinking	Clout	Authenticity	Emotional tone
English	Educators	49.70	91.77	4.32	78.15
	ChatGPT	66.14	96.74	14.46	97.51
Greek	Educators	54.51	94.62	13.68	89.55
	ChatGPT	57.35	98.66	4.94	99.00

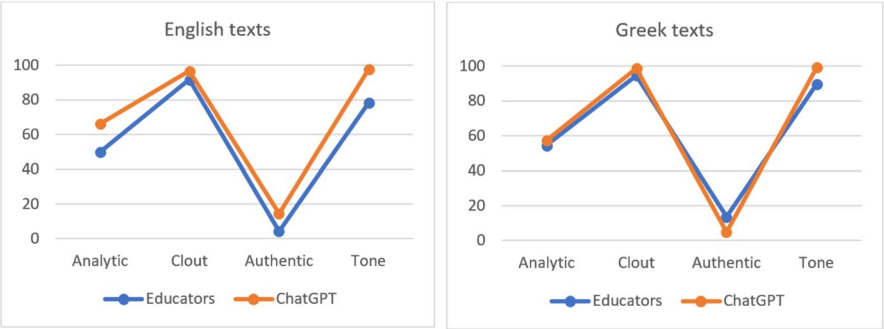

Fig. 5 Comparison between the educators' and ChatGPT's lexical analysis of their feedback

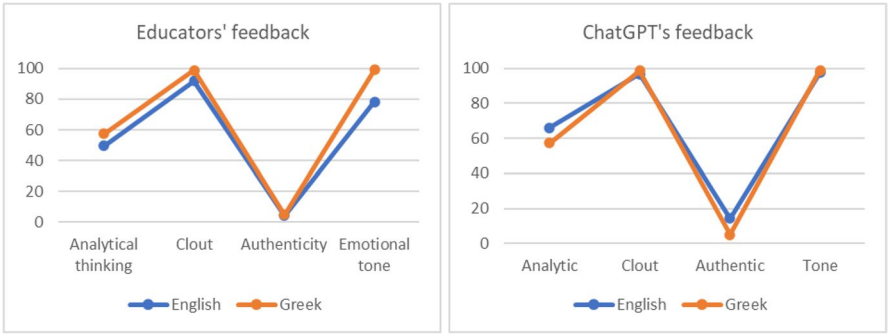

Fig. 6 The educators' and ChatGPT's lexical analysis of their feedback in the English and Greek texts

- Authenticity. Values below 50 indicate caution and social reservation, while values above 50 indicate a casual and unprepared style.
- Emotional tone. Values above 50 reflect a positive tone.

6.3 Answers to the research questions

Taking together the above results and to answer the RQs, the following can be noted:

- RQ1a: The quantitative analysis demonstrated that ChatGPT outperformed educators in identifying errors within English texts ($Mean\ rank_{Educators} = 141.28$, $Mean\ rank_{ChatGPT} = 107.42$, $p = 0.023$, $\eta^2 = 0.023$-small-medium). Conversely, the analysis of Greek texts indicated no statistically significant disparities, as both educators and ChatGPT detected an equivalent number of mistakes ($Mean\ rank_{Educators} = 112.20$, $Mean\ rank_{ChatGPT} = 93.53$, $p = 0.211$). However, it is noteworthy that, in the Greek texts, ChatGPT erroneously flagged certain words or phrases as incorrect, an error not observed in the judgments of educators ($Mean\ rank_{Educators} = 104.50$, $Mean\ rank_{ChatGPT} = 170.50$, $p < 0.001$, $\eta^2 = 0.574$-very large). This phenomenon was not replicated in the English texts, where neither educators nor ChatGPT falsely recognized errors ($Mean\ rank_{Educators} = 110.50$, $Mean\ rank_{ChatGPT} = 110.50$, $p = 0.999$).
- RQ1b. Concerning feedback's word count, ChatGPT was consistently more comprehensive than the educators for both the English ($Mean\ rank_{Educators} = 100.51$, $Mean\ rank_{ChatGPT} = 210.45$, $p < 0.001$, $\eta^2 = 0.247$-very large) and Greek texts ($Mean\ rank_{Educators} = 102.86$, $Mean\ rank_{ChatGPT} = 189.95$, $p < 0.001$, $\eta^2 = 0.144$-large). Despite this, the variance in feedback focus was marginal. ChatGPT demonstrated a slight preference for providing support and guidance, whereas educators exhibited a minor inclination toward language mechanics. This pattern was observed in the assessment of both English and Greek texts (see Table 4 and Fig. 1). In the English texts, ChatGPT delivered more comprehensive feedback in terms of orientation, as it maintained a balance between all four feedback categories (see Table 5

and Fig. 3). On the other hand, educators predominantly favored an outcome-oriented approach. However, a contrasting pattern emerged within the Greek texts. ChatGPT's feedback was exclusively outcome-oriented and personalized, while the feedback rendered by educators showed a more equitable distribution across different categories, although it was somehow unbalanced as there was a noticeable inclination towards outcome-oriented and personalized feedback. The lexical analyses revealed that, in the English texts, ChatGPT's style/tone while being more academic/formal (ChatGPT = 66.14, educators = 49.70), was more positive than that of educators (ChatGPT = 97.51, educators = 78.15) (see Table 7 and Fig. 5). Both the educators and ChatGPT adopted a rather authoritarian/confident (values > 90) and reserved/non-casual tone (values < 15). In the Greek texts, the only significant difference was that ChatGPT's tone was slightly more positive (ChatGPT = 99.00, educators = 89.55). As in the English texts, both the educators and ChatGPT adopted a rather authoritarian/confident (values > 90) and reserved/non-casual tone (values < 15), while in terms of analytical thinking the style of both was neither academic nor friendly (values ≈ 55).

- RQ1c. Concerning the assignment of grades, it was observed that educators assigned higher grades compared to ChatGPT in the English texts ($Mean\ rank_{Educators} = 115.81$, $Mean\ rank_{ChatGPT} = 57.45$, $p < 0.001$, $\eta^2 = 0.071$-medium-large) but there were no differences in the Greek ones ($Mean\ rank_{Educators} = 112.27$, $Mean\ rank_{ChatGPT} = 92.78$, $p = 0.185$). It should be noted that in all cases the grades were rather low.
- RQ2: The quantitative analysis substantiated that ChatGPT detected a comparable number of errors in texts of both languages ($Mean\ rank_{English} = 20.95$, $Mean\ rank_{CGreek} = 20.05$, $p = 0.820$). Nonetheless, it is noteworthy that ChatGPT erroneously identified correct words and phrases as errors in the Greek texts, an issue that was absent in the English ones ($Mean\ rank_{English} = 14.50$, $Mean\ rank_{Greek} = 26.50$, $p < 0.001$, $\eta^2 = 0.408$-very large). Moreover, the feedback ChatGPT provided in the English texts was notably more elaborate in terms of word count ($Mean\ rank_{English} = 28.85$, $Mean\ rank_{Greek} = 12.15$, $p < 0.001$, $\eta^2 = 0.511$-very large). On the other hand, ChatGPT assigned statistically higher grades to the Greek texts ($Mean\ rank_{English} = 13.58$, $Mean\ rank_{Greek} = 27.43$, $p < 0.001$, $\eta^2 = 0.395$-very large). As a sidenote, when comparing the educators' feedback, it was observed that they found more mistakes in the Greek texts, did not erroneously identify correct words and phrases as errors in both languages, provided more detailed feedback in the Greek texts, and there were no differences in their grading. Delving into the qualitative aspects, it was observed that ChatGPT's feedback focus diverged; in the English texts, ChatGPT concentrated on language mechanics, whereas in the Greek ones, the emphasis was on providing support and guidance. Then again, a similar pattern was observed in educators' feedback. Regarding the orientation of feedback, the analysis indicated that ChatGPT delivered a more equitable distribution of feedback across the four categories within the English texts, suggesting a comprehensive approach (see Table 6 and Fig. 4). In contrast, the examination of the Greek texts revealed a narrower orientation, with ChatGPT

dispensing feedback only in the form of outcome-oriented and personalized comments. The lexical analysis indicated that there were no significant differences in ChatGPT's feedback for both languages (see Table 7 and Fig. 6); its feedback's style/tone was slightly academic/formal, authoritarian/confident, reserved/non-casual, and rather positive, a trend almost identical to that of educators. It should be noted that ChatGPT's feedback for the essays written in Greek contained spelling, grammar, and syntax mistakes. Yet, these mistakes were considerably fewer than those made when ChatGPT v3.5 was used in the study's preliminary stage.

7 Discussion

This study explored two RQs. The first aimed to elucidate the variances between educators and ChatGPT in providing corrections and feedback on short essays written by students in both English and Greek. The objective of the second was to delve into the differences in how ChatGPT approaches the task of correcting texts across these two languages. The empirical evidence has yielded an array of findings that merit further discussion.

7.1 General comments

As discussed in a previous section, the role of the prompt is of paramount importance, given that the outcomes rely on its precision and clarity (De Winter et al., 2023). The prompt crafted for this study was designed to strike a balance between precision and utility, deliberately avoiding the over-specification of the feedback focus, orientation, and style expected from ChatGPT. In other words, the prompt was engineered to align with the standard perception of an educator's feedback mechanism. A more detailed prompt might probably have produced feedback far superior to that of educators. However, such specificity would have not allowed for a fair comparison with educator-provided feedback, except under the condition that they were similarly directed and it was ascertained that such directives were adhered to. Even so, it must be noted that the feedback from educators, under these provisions, would likely not mirror their authentic professional practices, as they tend to offer limited and superficial comments (Weaver, 2006) or brief evaluative statements (Dai et al., 2023). This tendency was substantiated within the scope of the present study, with instances recorded where participant feedback was composed of a mere seven or even two words (see Table 1). Nonetheless, it was anticipated that ChatGPT would deliver extensive feedback in terms of word count. It is well-documented that AI-generated feedback typically exhibits greater length (e.g., Tian & Zhou, 2020), but crucially, AIs are not subject to the temporal and physical constraints inherent to humans. This quantitative finding underscores the aforementioned advantage but does not provide enough evidence for its quality. This issue will be discussed in a coming section.

Another result that warrants attention is the grading of the short essays by educators and ChatGPT. It was observed that ChatGPT exhibited a more strict approach than educators when evaluating the English texts, though low scores were awarded by both. In the assessment of the Greek texts, ChatGPT's grading closely mirrored that of the educators, albeit both, once again, tended to lower scores. These findings appear to partially resonate with the research conducted by Azmi et al. (2019), which reported a 0.76 correlation between the grading outcomes of an AI system and human evaluators. What also needs to be discussed is the issue of objectivity in grading. There is evidence suggesting that automated text scoring systems may enhance consistency and promote objectivity in evaluation practices, contributing to a more transparent and fairer grading framework (e.g., Altamimi, 2023; Glaser, 2023; Hussein et al., 2019). The lower grades dispensed by ChatGPT (in both languages) accurately reflect the many mistakes, as well as the problems in expression and structure the texts had. However, the application of such strict standards in primary education raises pedagogical concerns, while it contrasts with the typically lenient grading approach adopted by teachers at this level of education. In fact, in Greece, during the early stages of primary education, the teachers are instructed not to assign grades on students' work, to prevent diminishing their motivation. While ChatGPT's grading could be deemed more objective in comparison to that of the educators involved in the study, an argument could be made advocating for the integration of subjective considerations to align more closely with the pedagogical goals of primary education. Equally intriguing is the deviation of the participating educators from expected grading norms; they too assigned low grades, a phenomenon potentially attributable to their awareness of participating in a research study, which might have "forced" them to adopt a stricter grading stance.

7.2 Comments regarding RQ1

The primary objective of the study, as suggested in RQ1, was to comparative analyze the corrections and feedback rendered by ChatGPT and educators. To this end, a suite of metrics was evaluated to determine the outcomes of this comparison.

Error detection proficiency Central to the task of correcting student texts is the ability to accurately identify errors related to spelling, grammar, and syntax, while abstaining from erroneously flagging any correct word or phrase as mistakes. There is literature suggesting that AIs demonstrated high correction accuracy rates (Azmi et al., 2019), addressing grammatical, syntactic, lexical, spelling, and punctuation-related mistakes (Abdullayeva & Musayeva, 2023; Link et al., 2022). Furthermore, it has been argued that such precision could substantially enhance academic performance by mitigating the recurrence of student errors (Sanosi, 2022). However, a strand of the literature suggested that AIs' proficiency may be limited to specific categories of mistakes (Fitria, 2021). The findings of the study reveal a divided landscape. In the English texts, ChatGPT's proficiency was commendable, exceeding that of human educators, albeit the margin of statistical significance was moderate ($p = 0.023$). In contrast, in the Greek

texts, the variance between the error detection of educators and ChatGPT was not statistically significant, though educators detected more errors. Given that, the findings, although in line with past research (at least for the English texts), do not indicate a significant advantage of AIs over humans, other than speed. A point of particular concern is the observation that, in the Greek texts, ChatGPT erroneously identified as mistakes correctly written words or phrases. This phenomenon echoes concerns previously expressed regarding the capacity of AIs to match humans in correction tasks (Murphy, 2019; Wang et al., 2022), raising concerns about their language proficiency and the reliability of the feedback they provide, issues that will be discussed in a coming section.

Feedback length The metric of feedback length, while previously addressed, warrants further emphasis due to the striking disparity observed in word count. In an examination of the Greek texts, the feedback provided by ChatGPT was nearly double than that of the educators. This discrepancy was even more pronounced in the English texts, where ChatGPT's feedback exceeded that of educators almost by a factor of five. These observations are consistent with literature which suggested that automated feedback systems tend to generate more comprehensive responses (Dai et al., 2023; Tian & Zhou, 2020).

Feedback focus Feedback constitutes a pivotal component in the pedagogical process, particularly within the domain of essay writing. Common-sense dictates that it necessitates a comprehensive approach that addresses the multi-faceted nature of student writing, which poses a formidable challenge to educators (Matcha et al., 2019). On one hand, there is literature arguing that AIs furnish feedback comparable to that of educators (Jia et al., 2022). On the other hand, the intrinsic complexity of evaluating creative writing has led scholars to advocate for the refinement of AI capabilities (Ramesh & Sanampudi, 2022). Various studies have documented instances where AI fell short in adeptly assessing written tasks with respect to their structure, coherence, and content (Lu, 2019), and in some cases, offered explanations that lacked adequacy (Barrot, 2023). The analysis revealed that both the educators' and ChatGPT's feedback for both languages focused on four themes, namely language mechanics (meaning grammar, syntax, punctuation, and spelling), content structure, expression, and the provision of support and guidance. Given the brevity of the texts, commentary on structure and expression was anticipated to be minimal. Consequently, it was expected that the majority of comments would involve language mechanics and the provision of support and guidance. The analysis unveiled that in the context of the English texts, educators and ChatGPT exhibited a matching pattern of focus, prioritizing language mechanics, followed by support and guidance. The comments related to expression and content structure were notably fewer, aligning with expectations. The scenario diverged when assessing the Greek texts. Despite a similarity in functionality between educators and ChatGPT, there was a noticeable pivot towards comments on support and guidance, with observations on language mechanics receding considerably. This

difference in the findings both corroborates and conflicts with research supporting that AIs focus more on grammar and syntax (Han & Sari, 2022; Wang & Han, 2022). Furthermore, the outcomes related to the Greek texts were unanticipated; because of the presence of numerous issues in language mechanics, it was presumed that both educators and ChatGPT would generate a substantial volume of comments addressing these concerns. Nonetheless, it is noteworthy that ChatGPT paralleled the approach of educators concerning feedback focus.

Orientation Hattie and Timperley (2007) suggested that feedback can be categorized as outcome-oriented (that assesses the accuracy of the assignment), process-focused (that focuses on the process of completing a task), self-regulation (that provides students with the means to self-evaluate their performance), and personalized (that addresses the individual characteristics). It can be supported that a balanced distribution of these feedback types is highly desirable as it indicates a holistic and comprehensive approach. Indeed, while feedback illuminates potential areas for enhancement (Parikh et al., 2001), it must be customized to align with each student's needs (Henderson et al., 2019) as this can positively influence their knowledge acquisition (Vasilyeva et al., 2007). Studies have shed light on the potential of AI-generated feedback to strengthen self-regulatory practices and refine writing skills (Osawa, 2023). Yet, AIs' feedback is quite directive and, thus, does not adequately foster self-motivation and self-correction (Liang et al., 2023). Others suggested further advancements are needed for AIs to be able to dispense personalized feedback (Conati et al., 2021). Upon analyzing feedback methodologies, a clear dichotomy emerged. In the English texts, ChatGPT demonstrated a balanced feedback approach, whereas educators tended to employ an unbalanced, outcome-focused strategy, often neglecting the other three orientations. Conversely, in Greek texts, educators provided a more balanced feedback orientation, while ChatGPT's was limited to outcome-oriented and personalized feedback. In this respect, while it was supported that ChatGPT's feedback is superior to that of educators (Jacobsen & Weber, 2023), the findings of this study do not fully support this notion; in terms of orientation ChatGPT was more comprehensive than the educators in the English texts, while in the Greek texts, educators outperformed ChatGPT in the diversity of feedback provided.

Style/tone The domain of student feedback extends beyond mere content critique; it encompasses how educators address students when providing feedback, as this significantly influences their performance. Though there is no rule of thumb regarding the style/tone of feedback, kindness and discretion are desirable as they can enhance the performance of students, especially those facing learning problems (McLaren et al., 2011). Yet, general praise or indiscriminate approval does not constitute "good" feedback, as it potentially undermines students' willingness to acknowledge and learn from their mistakes (Maclellan, 2005). Moreover, the delivery of feedback should maintain the delicate balance in the student–teacher dynamics (Wang et al., 2008). Considering these factors and taking into account the young age of students, an optimal approach to feedback is one that carefully

navigates between formality and approachability. Such feedback is neither overly casual nor rigorous; it exudes a sense of educator confidence without crossing into authoritarianism. Upon analyzing the results, it became apparent that in the English texts, the feedback generated by ChatGPT assumed a more academic and formal tone in comparison to that provided by the educators. Despite this, both demonstrated a tendency towards authoritative confidence and a reserved/non-casual demeanor. Nevertheless, it is noteworthy that ChatGPT's feedback was characterized by a relatively more positive tone than that of the educators. Consequently, there is no clear conclusion regarding who's style was better. The same applies in the case of the Greek texts, as the findings indicated an absence of significant differences. Both the educators and ChatGPT were neither too academic/formal nor friendly, maintained a confident, reserved/non-casual posture, and their feedback was imbued with a positive sentiment.

In sum, in the English texts, ChatGPT identified a greater number of errors and dispensed more extensive feedback in terms of word count. Furthermore, ChatGPT's emphasis within its feedback was markedly similar to that of educators, focusing mostly on language mechanics. Additionally, ChatGPT's feedback was characterized by a more holistic and balanced approach in terms of orientation, encapsulating comments that spanned all four feedback categories uniformly. In terms of style, there is no clear conclusion regarding who's style was better. In the Greek texts, ChatGPT's performance in error detection was comparable to that of educators. Despite this, it incorrectly identified certain correct words and phrases as errors, which can be considered a significant weakness. ChatGPT's focal points in feedback paralleled those of educators, although it focused less on language mechanics. Nonetheless, its feedback dispensation was characterized by a narrower orientation compared to educators, concentrating on merely two of the four feedback categories, This, denotes a diminished level of comprehensiveness and balance. In terms of style/tone, there were no significant differences.

Therefore, it can be supported that ChatGPT demonstrated proficiency in evaluating English texts, surpassing the performance of educators in some aspects. In contrast, in the Greek texts, one can support that ChatGPT did not perform as well as educators did.

7.3 Comments regarding RQ2

Research has highlighted the need to consider the unique linguistic and cultural backgrounds of students to deliver quality feedback (Osakwe et al., 2022). Then again, studies found limitations in AIs' capabilities to process linguistic structures in various languages, raising concerns about their efficacy to deliver a level of correction similar to that of humans (Murphy, 2019; Wang et al., 2022). ChatGPT was also found to have such limitations, that may affect its feedback quality (Fuchs, 2023). Consequently, it was justified to comparatively examine ChatGPT's performance on texts written in English versus those in Greek, with the latter being a less commonly spoken language and, hence, presumably less

represented in its training datasets. This examination was anticipated to provide insights into the proficiency of AI in handling languages with varied degrees of complexity and prevalence.

Although in terms of accurate error detection ChatGPT performed equally well in both languages, it evidenced shortcomings in the Greek texts as it inaccurately flagged correct words as mistakes. This is a first and rather strong indication that it did not perform well in the Greek language. Moreover, compared to Greek texts, ChatGPT offered more extensive feedback in the English ones. Given that the difference was significant, one cannot ascribe this variance to differences in the texts alone; one can support that this difference serves as yet another indication of ChatGPT's linguistic limitations.

ChatGPT's feedback on the English texts was focused mainly on language mechanics, while on the Greek texts, it was skewed towards providing support and guidance, neglecting language mechanics. In fact, its language mechanics-related comments in the Greek texts amounted to merely half of the ones made in the English texts and it was one of the few cases in which the educators surpassed ChatGPT. This can be viewed as another linguistic limitation of ChatGPT. The orientation of feedback provided for the Greek texts exhibited a lack of balance, as it was confined to addressing merely two out of the four types of feedback (outcome-oriented and personalized). In contrast, in the feedback distribution in the English texts, a more equitable dispersion was observed across all four categories, suggesting a comprehensive strategy. This discrepancy highlights yet another significant variation in the feedback mechanism of ChatGPT across different languages and underscores the need for further development in AI-generated feedback systems to achieve truly equitable and holistic educational support.

However, the lexical analysis revealed no significant disparities; in both languages, the feedback style/tone was slightly academic, authoritarian/confident, reserved/non-casual, and positive. Finally, though it is not part of the analysis, it was observed that ChatGPT made several spelling and grammar mistakes in the feedback it provided on Greek texts.

As a result of the above, it can be concluded that ChatGPT appears to possess a heightened aptitude for the English language compared to Greek; thus, its language processing capabilities evidently vary across different linguistic contexts.

7.4 Implications for research and practice

The results of the study hold significant relevance for researchers, AI specialists, and educators. The methodology necessitated the transcription of students' handwritten essays into a digital format prior to assessment by ChatGPT. In real educational settings, expecting young learners to compose essays using a word processor is impractical to the point of being unfeasible. Furthermore, it is not viable for educators to assume the task of transcription as they can easily provide feedback directly on physical documents. Consequently, there is a demand for tools that can

seamlessly convert handwritten materials into digital texts. This necessity likely demands enhancements in existing Optical Character Recognition (OCR) technology to achieve heightened accuracy and processing speed.

ChatGPT made mistakes in its Greek responses, both in terms of error identification and grammar/syntax mistakes in its feedback. This observation warrants attention as it may bear implications for the reliability and efficacy of ChatGPT in languages that are less prevalently spoken, and suggests the necessity for further refinement of its language processing capabilities.

Although certain features characterize effective feedback, it is intrinsically subjective; some educators prioritize language mechanics over comments on expression, and others may favor outcome-oriented feedback over process-focused one. To address this variance, it is essential to equip educators with tools that allow the extensive customization of AI feedback parameters. Such customization would enable AI-generated feedback to align with individual teaching styles, pedagogical objectives, and the need to provide personalized feedback.

Educators frequently face the challenge of balancing heavy workloads and the provision of timely feedback. Such a balancing act is critical, as delayed feedback has a negative impact on students' engagement (Poulos & Mahony, 2008). Given that ChatGPT was able not only to provide quite comprehensive feedback for the texts written in English and adequate feedback for the texts written in Greek but also rather fast, it can be supported that AIs can alleviate the educators' workload as others suggested (Jia et al., 2022). However, as ChatGPT made errors when processing Greek texts, educators are forced to allocate additional time to reviewing and correcting its output. Moreover, they have to critically assess and refine the output before endorsing it, as they are the ones to decide on the appropriateness of AI-generated feedback (O'Cain et al., 2023). These necessities cast doubt on the alleged efficiency gains from employing AIs for feedback provision. The findings of Cavalcanti et al. (2021) reinforce this skepticism, as in 46.03% of the studies they reviewed there was no evidence that automatic feedback eased instructors' workload.

The need for critical assessment is also important due to the profound ethical implications involved. There is a vigorous debate centered on the potential of AI to facilitate cheating and plagiarism among students (e.g., Dehouche, 2021; Farrokhnia et al., 2023), necessitating a reconsideration of the educational paradigm. This raises the question of whether educators' unconditional reliance on AI for student feedback could also be viewed as a form of academic dishonesty. On a broader scale, the abilities of AI to design lesson plans, advise on pedagogical approaches, propose subject-specific exercises, evaluate assignments, and deliver detailed feedback (as demonstrated in this study), call for reflection on the effects on teachers' role and the need to redefine their position in an AI-integrated educational landscape. In fact, it is rather possible that, in the near future, teachers will have to harness the technological capabilities of AI not merely as supplementary tool but as a transformative agent that transcends traditional pedagogical boundaries. To fully leverage AI, educators must delve beyond its surface functionalities and effectively engage with its core mechanisms. For example, AI's ability to personalize learning experiences stands as

a cornerstone of its educational utility. By analyzing datasets encompassing student performance, learning styles, and behavioral patterns, it can deliver tailored instructional materials. In addition, teachers can utilize AI to identify knowledge gaps, predict future learning trajectories, and provide targeted interventions, thus fostering a more effective educational environment. Furthermore, AI's capacity to automate administrative tasks significantly liberates educators from the time-consuming activities that often detract from instructional time. By doing so, a more human-centric approach can be promoted, that allows teachers to focus on developing critical thinking, creativity, and socio-emotional skills among students.

Yet, to see beyond the surface of AI, teachers should cultivate a mindset of adaptive expertise, where they are not only adept at using AI tools but also proficient in modifying their instructional approaches based on AI-driven insights. To achieve this, continuous professional development aimed at understanding the underlying principles of AI technologies is strongly advised. This, will empower teachers to critically assess AI tools and integrate them effectively into their pedagogical strategies. Collaborative endeavors between educators and AI developers are also needed, to design AI systems that are pedagogically sound.

7.5 Limitations and future work

This study encompasses limitations that warrant recognition. The selection of ChatGPT introduces a degree of uncertainty regarding the potential performance of alternative AIs. Some may argue that the prompt could have been more detailed and specific. On the other hand, one can counterargue that, in reality, it is very hard for educators, regardless of their experience, to consider several parameters and incorporate them into their feedback. In addition, as mentioned in the previous section, the feedback depends on one's preferences. Acknowledging the varied approaches to feedback, the study opted for a less detailed prompt and minor style adjustments with the intent to simulate the feedback of an educator. Furthermore, an intricate prompt would have skewed the data, potentially creating an undue advantage for ChatGPT. Although several feedback dimensions were analyzed, others might have opted for the examination of parameters not included in this study. Moreover, the corpus of student essays, although reflective of their linguistic proficiency, was quantitatively constrained. A larger set of essays would have undoubtedly afforded a more robust comparative analysis of the feedback provided by educators and ChatGPT. This limitation extends to the number of educators involved in the study, with the small cohort potentially affecting the diversity of feedback styles captured.

The aforementioned limitations may serve as guidelines for future research endeavors. A larger sample of participating educators can facilitate the documentation of a broader spectrum of feedback modalities. Moreover, the incorporation of more essays, encompassing a variety of topics, would substantially enrich the heterogeneity of the texts subject to feedback provision. Furthermore, the investigation of more intricate prompts is necessary, contingent upon the provision

of equally detailed guidelines to educators. In light of the multifaceted nature of feedback, researchers can explore and evaluate multiple analytic methodologies. The efficacy of ChatGPT in dispensing feedback calls for a comparative analysis with other AIs. Finally, it is crucial to assess the proficiency of AI systems in offering feedback across a multitude of languages. Through such testing, researchers can gain a deeper insight into the constraints of current AI capabilities and, consequently, propose enhancements to the systems' performance in feedback provision.

8 Conclusion

The present study sought to evaluate the efficacy of ChatGPT in its role as a corrective and feedback tool for short essays written by primary school students in both the English and Greek languages. A comparative analysis was conducted between the precision and effectiveness of corrections and feedback generated by ChatGPT and those of educators. The findings revealed that, in the case of the English-language essays, ChatGPT demonstrated superior performance over educators, surpassing them in both the volume and quality of its interventions. ChatGPT exhibited a higher error detection rate and furnished more comprehensive feedback. Furthermore, its focus mirrored that of the educators, but it maintained a more balanced orientation and adopted a more constructive yet academically inclined style and tone. Conversely, for essays composed in Greek, ChatGPT's proficiency did not match that of the educators. Despite identifying a comparable number of errors, it mistakenly identified correct words and phrases as errors. Additionally, ChatGPT's emphasis on language mechanics was less pronounced than that of the educators, and it provided feedback that was less equitably oriented. No notable discrepancies were observed regarding the style and tone. An assessment of ChatGPT's performance across essays in both languages indicated a marked superiority in English, with ChatGPT excelling in all measured quantitative and qualitative aspects. In conclusion, the study underscores ChatGPT's potential as an educational aid, particularly for English language learners, while also highlighting areas for improvement in its application to Greek language essays. In this respect, the study is anticipated to contribute to the pedagogical discourse on the integration of AI-driven solutions in educational practices and establish a foundation for future technological implementations in the realm of academic assessment.

9 Appendix

The themes and codes of the qualitative analyses. Please note that *n* represents the number of comments in which a code appeared, although it was possible for a code to appear more than once in a feedback.

Table 8 Themes and codes found in the educators' feedback for the English texts

Theme	Code	*n*	%	Example quotes
Language mechanics (49.9%)	grammar concerns	132	13.9	"There were few grammar and spelling mistakes"
	spelling accuracy	73	7.7	"…pay close attention to your spelling"
	vocabulary usage	57	6.0	"…there are some vocabulary mistakes."
	syntax precision	37	3.9	"There were syntax problems."
	error awareness	30	3.2	"Be careful with the mistakes you made due to haste"
	pronoun use	25	2.6	"Be careful with the use of pronouns"
	punctuation accuracy	23	2.4	"Some commas were missing."
	word choice and accuracy	22	2.3	"Check the wrong use of words."
	prepositional use	21	2.2	"Study more the use of prepositions."
	tense usage	20	2.1	"…confused with past tense."
	use of Greek expressions	20	2.1	"Be careful with the Greek expressions you are using."
	linking word use	11	1.2	"…you could use more linking words for better structure."
	singular-plural confusion	3	0.3	"'Actors' is plural, so use 'are' not 'is'…"

Table 8 (continued)

Theme	Code	*n*	%	Example quotes
Expression (18.4%)	engagement and clarity	32	3.4	"Very engaging content, with proper expression of ideas."
	idea expression	31	3.3	"You had great ideas…"
	positive expression	21	2.2	"…overall, in your essay, you used nice expressions."
	avoidance of repetition	19	2.0	"Try to moderate the use of the same words many times."
	content relevance and interest	14	1.5	"…interesting content."
	narrative clarity	12	1.3	"…your letter had a good storyline."
	cohesion and coherence	11	1.2	"…but the errors in grammar and the coherence of the text, make it hard to understand."
	tone and style	7	0.7	"You have a nice friendly tone"
	expressive weaknesses	5	0.5	"…many mistakes regarding expression that made the text difficult to understand…"
	contextual clarity	4	0.4	"Mixed context, needs more clarity…"
	topic understanding	4	0.4	"You did a great job analyzing the advantages and the disadvantages…"
	content compliments	3	0.3	"…overall great text with a beginning, main part, and a conclusion"
	contextual flow	3	0.3	"Very good contextual flow…"
	semantic accuracy	3	0.3	"…be more specific to avoid semantic errors."
	appropriate greetings and closures	2	0.2	"You should have written a greeting at the beginning."
	text expansion	2	0.2	"You could have described a few more jobs."
	introduction appreciation	2	0.2	"Great introduction about the film and how it impacted the industry."

Table 8 (continued)

Theme	Code	*n*	%	Example quotes
Content structure (8.5%)	structure and organization	58	6.1	"Your essay had a suitable structure."
	text consistency	20	2.1	"…you have a nice consistency in your writing."
	paragraph organization	3	0.3	"You should also organize your ideas in paragraphs…"
Support and guidance (23.1%)	encouragement for further study/reading	71	7.5	"Try to read more letter examples to give you inspiration."
	encouragement and positive reinforcement	60	6.3	"Can't wait for your next letter."
	attention to detail	25	2.6	"Pay attention to your mistakes and try not to repeat them."
	engagement with English media/speakers	13	1.4	"I'd highly encourage you to watch some English movies, and videos, or read some English books…"
	adherence to instructions	13	1.4	"Following the instructions I gave you regarding grammar"
	encouragement for revision	11	1.2	"Try to rewrite your essay."
	text review before submission	10	1.1	"I think you should try to check your text before handling it for evaluation…"
	thinking in English	5	0.5	"You need to cultivate the skill of thinking in English."
	encouragement for progress	5	0.5	"…looking forward to your improvement."
	practice recommendations	3	0.3	"Keep practicing and you will become better…"
	avoid hasty writing	3	0.3	"…try not to write hastily."

Table 9 Themes and codes found in the educators' feedback for the Greek texts

Theme	Code	*n*	%	Example quotes
Language mechanics (31.2%)	attention to spelling errors	85	12.5	"Pay attention to spelling."
	punctuation use	38	5.6	"Try using punctuation marks."
	need for grammatical improvement	35	5.1	"Beware of grammatical errors."
	stress marks	20	2.9	"Do not forget the stress marks!"
	grammar and spelling issues	12	1.8	"There are quite a lot of grammar and spelling mistakes."
	capitalization of initial letters	6	0.9	"…always, the first word in a sentence is capitalized."
	connectives utilization	5	0.7	"Use different connecting words"
	syntax issues	4	0.6	"…but also, the syntax of your sentences."
	use conjunctions	3	0.4	"Try using conjunctions to link sentences."
	lexical diversity	2	0.3	"You need to have a richer vocabulary"
	use of synonyms and pronouns	2	0.3	"…try to use pronouns or synonyms for certain words."

Table 9 (continued)

Theme	Code	*n*	%	Example quotes
Expression (23.3%)	expression and clarity	35	5.1	"In some places, your expression was not good and the text was not easy to understand."
	emotional expression	15	2.2	"You presented your feelings very nicely."
	presentation of ideas	14	2.1	"Try presenting your ideas more thoroughly."
	word repetition avoidance	13	1.9	"Watch out for word repetition."
	proper salutation and closure	12	1.8	"You addressed the mayor properly…"
	event description	10	1.5	"You described the celebration nicely…"
	warning against informality	9	1.3	"try to avoid informality."
	recognition of good ideas	8	1.2	"While your ideas are generally good"
	commentary on ideas and organization	7	1.0	"While I see that you have nice ideas, you don't express them in an organized and structured way."
	knowledge of historical events	4	0.6	"Read what we celebrate on October 28!"
	personal opinion and experience	4	0.6	"Well done for presenting… your opinion!"
	organizing ideas effectively	4	0.6	"Organize your ideas to make the text more appealing."
	use of short sentences for clarity	4	0.6	"Try using shorter sentences."
	thematic relevance and adherence	3	0.4	"Well done! You presented the topic nicely."
	understanding of the subject	3	0.4	"…have a good understanding of the topic you are analyzing."
	politeness in expression	2	0.3	"The most important thing: we express ourselves politely."
	effective argumentation	2	0.3	"In your text, there is persuasive argumentation."
	proper letter formatting	2	0.3	"When writing a letter, we follow certain rules."
	logical flow	2	0.3	"Sentences should be linked to each other conceptually."
	style appropriateness	1	0.1	"Use the right style for the occasion."
	critique on readability	1	0.1	"In general, your text is not so easy to read."
	recognition of detailed description	1	0.1	"Your text was detailed."
	thematic relevance	1	0.1	"Focus on the subject…"
	chronological narration	1	0.1	"Narration of events in chronological order."

Table 9 (continued)

Theme	Code	*n*	%	Example quotes
Content structure (7.3%)	cohesion and coherence	32	4.7	"Your text was correct in terms of content and coherence."
	commentary on structure	18	2.6	"Your text is very well structured."
Support and guidance (38.2%)	positive reinforcement	86	12.6	"Congratulations, you did great!"
	suggestions for improvement	37	5.4	"I want you to try to keep grammatical and syntactic rules in mind when writing."
	acknowledgment of effort	28	4.1	"Bravo for the effort!"
	constructive critisism	25	3.7	"…try not to repeat the same mistakes…"
	attention to detail	20	2.9	"Don't rush and pay attention to details."
	encouragement of practice	20	2.9	"Keep practicing and your writing will become better…"
	proofreading post-completion	15	2.2	"Re-read your text to make corrections."
	encouragement to continue the effort	10	1.5	"Keep trying."
	encouragement for improvement	9	1.3	"Your text raises awareness and your ideas are not bad, they just need further development."
	grammar rules and revision	5	0.7	"Try re-reading the grammar rules."
	vocabulary review	2	0.3	"It is necessary to review your vocabulary."
	comments on reading practices	1	0.1	"You can improve in many ways by reading books."
	use of a dictionary or asking for help	1	0.1	"Use a dictionary if you do not know the spelling of a word."
	focus on the task	1	0.1	"Try to be more focused."

Table 10 Themes and codes found in ChatGPT's feedback for the English texts

Theme	Code	*n*	%	Example quotes
Language mechanics (43.6%)	spelling accuracy	20	5.5	"Also, try to clarify your ideas by using the correct words and spellings."
	verb tense usage	20	5.5	"Pay attention to verb tenses…"
	preposition correctness	19	5.2	"A little tip for prepositions: they can be tricky…"
	grammar improvement	18	5.0	"However, to make your advice more effective and understandable, focusing on improving your grammar…"
	subject-verb agreement	18	5.0	"Be mindful of subject-verb agreement…"
	punctuation importance	18	5.0	"Punctuation to help your reader follow your advice with ease."
	article usage	16	4.4	"Use articles where necessary, as this helps to clarify your meaning."
	pronoun usage	12	3.3	"Watch out for your pronoun usage."
	singular/plural awareness	8	2.2	"Watch out for the plural forms of nouns; 'people' is already plural…"
	vocabulary compliment	6	1.7	"It's clear that you have a good grasp of the vocabulary…"
	capitalization	3	0.8	"Capitalization is important, so always use 'I' instead of 'i'"
Expression (20.6%)	direct address technique	11	3.0	"Your recent email was received with great warmth…"
	writing clarity	14	3.9	"Consider revising sentence construction for better clarity…"
	expressive enthusiasm	10	2.8	"Your enthusiasm for history and your teacher, Mrs. Kanela, shines through…"
	specific detailing	10	2.8	"You've done a good job of recounting the events of your birthday party to Tom…"
	consistency in writing	7	1.9	"…strive for clear and concise sentence structures to ensure that your reader can easily follow your thoughts."
	Idea organization	6	1.7	You've done a good job of organizing your ideas into two clear categories…"
	topic relevance	6	1.7	"You've done a good job of organizing your ideas into two clear categories and staying on topic throughout the text."
	clear summarization	5	1.4	"You have managed to convey the basic timeline of exams, which is the core of the message…"
	personal experience inclusion	5	1.4	You have a good start with the structure of your text, keeping it simple and focused on your personal experiences."
Content structure (4.7%)	sentence structure	17	4.7	"Consider revising sentence construction for better clarity…"

Table 10 (continued)

Theme	Code	*n*	%	Example quotes
Support and guidance (31.2%)	encouragement of practice	20	5.5	"Remember, practice is key to improvement."
	error identification	20	5.5	"However, to make your advice more effective and understandable, focusing on improving your grammar…"
	constructive criticism	20	5.5	"However, there are opportunities for improvement in spelling and grammar."
	positive reinforcement	18	5.0	"You have done a wonderful job…"
	improvement opportunities	18	5.0	"However, to make your advice more effective and understandable, focusing on improving your grammar…"
	proofreading suggestion	10	2.8	"Watch out for spelling errors—double-check words you're unsure about to improve your writing."
	reading as a learning tool	7	1.9	"…try to read as much as you can in English—it's one of the best ways to improve."

Table 11 Themes and codes found in ChatGPT's feedback for the Greek texts

Theme	Code	*n*	%	Example quotes
Language mechanics (23.3%)	attention to spelling	20	9.1	"I would suggest that you practice a little more in spelling…"
	grammatical accuracy	15	6.8	"Revise your text for grammatical and spelling mistakes…"
	proper punctuation	14	6.4	"Also, remember to use punctuation marks correctly because they help the reader follow your thoughts."
	active voice	1	0.5	"I recognize the effort to use the active voice…"
	use of capital letters	1	0.5	"Also, remember to use capital letters when necessary."
Expression (17.8%)	enthusiasm	14	6.4	"You have managed to convey your enthusiasm, and this is very important in expression."
	encouragement for sharing	14	6.4	"First of all, I would like to congratulate you on your willingness to share your experience…"
	expression of thoughts and emotions	10	4.6	"It's wonderful that you can express your feelings and describe how you feel…"
	formality in writing	1	0.5	"Additionally, always ensure to write clearly and avoid the use of informal expressions in formal letters."
Content structure (10%)	text organization	11	5.0	"…organize your text into paragraphs…"
	coherence and cohesion	8	3.7	"Try to organize your thoughts in a clearer way…"
	appreciation of the content's value	3	1.4	"It is very significant that you express your opinions and propose solutions."

Table 11 (continued)

Theme	Code	*n*	%	Example quotes
Support and guidance (42.5%)	praise for effort	20	9.1	"Congratulations on your effort to write this text!"
	suggestions for improvement	20	9.1	"I would like to give you some advice to improve your writing next time."
	continuous improvement and persistence	18	8.2	"With perseverance and patience, you will see great improvement"
	constructive criticism	7	3.2	"I noticed that there are some mistakes in spelling and syntax that need correction."
	positive reinforcement at the end of comments	7	3.2	"I eagerly await reading your next text!"
	asking for help	5	2.3	"Pay close attention to spelling—tools like spellcheck can be very helpful here."
	improvement over time	5	2.3	"You will get better every time."
	oral rehearsal	4	1.8	"Try to read your text out loud after writing it…"
	personal growth through writing	3	1.4	"Continue the good work and remember that with each text you write and correct, you become better."
	creative thinking	2	0.9	"It is very important that you express your opinions and suggest solutions."
	missteps as learning opportunities	1	0.5	"Every mistake is an opportunity to learn."
	improvement through reading	1	0.5	"Also, the use of a dictionary and reading books can help you with spelling…"
Community and environmental engagement (6.4%)	environmental concern	7	3.2	"It's wonderful that you care about the environment and are trying to find ways to protect it."
	community involvement	5	2.3	"Your willingness to share the spirit of the holidays and to want to help those in need is important."
	ideas for improvement	2	0.9	"Your suggestions for the community improvement…"

Author contributions All authors contributed equally to the study.

Funding The study received no funding.

Data availability Data will be made available on reasonable request.

Declarations

Ethical statement The research was conducted in accordance with all pertinent legislation and institutional protocols. The Research Ethics Committee of the Department of Primary Education, University of the Aegean reviewed and approved the methodologies and practices of this study. All participants were briefed and their informed consent was obtained; they retained the right to withdraw at any point. The privacy and rights of the individuals involved were protected; no personal data were collected and/or processed.

Competing interests The authors declare that have no conflict of interest.

References

Abdullayeva, M., & Musayeva, Z. M. (2023). The impact of Chat GPT on student's writing skills: An exploration of Ai-assisted writing tools. *Proceedings of the International Conference of Education, Research and Innovation* (Vol. 1, No. 4), 61-66. ICERI. https://doi.org/10.5281/ZENODO.7876800

Ai, H. (2017). Providing graduated corrective feedback in an intelligent computer-assisted language learning environment. *ReCALL, 29*(3), 313–334. https://doi.org/10.1017/S095834401700012X

Ali, J. K. M., Shamsan, M. A. A., Hezam, T. A., & Mohammed, A. A. (2023). Impact of ChatGPT on learning motivation: teachers and students' voices. *Journal of English Studies in Arabia Felix, 2*(1), 41–49. https://doi.org/10.56540/jesaf.v2i1.51

Altamimi, A. B. (2023). Effectiveness of ChatGPT in essay autograding. *Proceedings of the 2023 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, 102–106. IEEE. https://doi.org/10.1109/iCCECE59400.2023.10238541

Altstaedter, L. L., & Doolittle, P. (2014). Students' perceptions of peer feedback. *Argentinian Journal of Applied Linguistics, 2*(2), 60–76.

Athanassopoulos, S., Manoli, P., Gouvi, M., Lavidas, K., & Komis, V. (2023). The use of ChatGPT as a learning tool to improve foreign language writing in a multilingual and multicultural classroom. *Advances in Mobile Learning Educational Research, 3*(2), 818–824. https://doi.org/10.25082/AMLER.2023.02.009

Aydın, Ö., Karaarslan, E. (2022). OpenAI ChatGPT generated literature review: Digital twin in healthcare. In Ö. Aydın (Ed.), *Emerging Computer Technologies 2* (pp. 22–31). İzmir Akademi Dernegi. https://doi.org/10.2139/ssrn.4308687

Azmi, A. M., Al-Jouie, M. F., & Hussain, M. (2019). AAEE-Automated evaluation of students' essays in Arabic language. *Information Processing & Management, 56*(5), 1736–1752. https://doi.org/10.1016/j.ipm.2019.05.008

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *INternational Journal of Educational Technology in Higher Education, 21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4

Barrett, C. A., Truckenmiller, A. J., & Eckert, T. L. (2020). Performance feedback during writing instruction: A cost-effectiveness analysis. *School Psychology, 35*(3), 193–200. https://doi.org/10.1037/spq0000356

Barrot, J. S. (2023). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning, 36*(4), 584–607. https://doi.org/10.1080/09588221.2021.1936071

Borji, A., & Mohammadian, M. (2023). Battle of the Wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard. GPT-4, Claude, and Bard. *SSRN, 2023*. https://doi.org/10.2139/ssrn.4476855

Borji, A. (2023). A categorical archive of ChatGPT failures. *arXiv preprint* arXiv:2302.03494. *https://doi.org/10.48550/ARXIV.2302.03494*

Boud, D., & Falchikov, N. (Eds.). (2007). Rethinking assessment in higher education: Learning for the longer term. Routledge.

Brookhart, S. M. (2008). How to give effective feedback to your students. ASCD.

Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education, 36*(4), 395–407. https://doi.org/10.1080/03075071003642449

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y. S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2*, 100027. https://doi.org/10.1016/j.caeai.2021.100027

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology, 15*(3), 1–45. https://doi.org/10.1145/3641289

Chatterjee, S., & Bhattacharjee, K. K. (2020). Adoption of artificial intelligence in higher education: A quantitative analysis using structural equation modelling. *Education and Information Technologies, 25*(5), 3443–3463. https://doi.org/10.1007/s10639-020-10159-7

Chen, Y., Wang, R., Jiang, H., Shi, S., & Xu, R. (2023). Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint* arXiv:2304.00723. *https://doi.org/10.18653/v1/2023.findings-ijcnlp.32*

Chiu, T. K. F., Meng, H., Chai, C.-S., King, I., Wong, S., & Yam, Y. (2022). Creation and evaluation of a pretertiary artificial intelligence (AI) curriculum. *IEEE Transactions on Education, 65*(1), 30–39. https://doi.org/10.1109/TE.2021.3085878

Clariana, R., Wagner, D., & Murphy, L. (2000). Applying a connectionist description of feedback timing. *Educational Technology Research and Development, 48*, 5–22. https://doi.org/10.1007/BF02319855

Cohen, J. (2013). Statistical power analysis for the behavioral sciences. *Academic Press*. https://doi.org/10.4324/9780203771587

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence, 298*, 103503. https://doi.org/10.1016/j.artint.2021.103503

Crosthwaite, P., Storch, N., & Schweinberger, M. (2020). Less is more? The impact of written corrective feedback on corpus-assisted L2 error resolution. *Journal of Second Language Writing, 49*, 100729. https://doi.org/10.1016/j.jslw.2020.100729

Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y.-S., Gasevic, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. *EdArXiv.* https://doi.org/10.35542/osf.io/hcgzj

De Winter, J. C., Dodou, D., & Stienen, A. H. (2023). ChatGPT in education: Empowering educators through methods for recognition and assessment. *Informatics, 10*(4), 87. https://doi.org/10.3390/informatics10040087

Dehouche, N. (2021). Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics, 21*, 17–23. https://doi.org/10.3354/esep00195

Di Placito, M. L., & Mortensen, E. (2023). Applying AI efforts to student assessments: That is, alternative innovations! *The Interdisciplinary Journal of Student Success, 2*, 93–108.

Escalante, J., Pack, A., & Barrett, A. (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

Fang, T., Yang, S., Lan, K., Wong, D. F., Hu, J., Chao, L. S., & Zhang, Y. (2023). Is ChatGPT a highly fluent grammatical error correction system? A comprehensive evaluation. *arXiv preprint* arXiv:2304.01746.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, 1–15. https://doi.org/10.1080/14703297.2023.2195846

Fitria, T. N. (2021). Grammarly as AI-powered English writing assistant: Students' alternative for writing English. *Metathesis: Journal of English Language, Literature, and Teaching, 5*(1), 65. https://doi.org/10.31002/metathesis.v5i1.3519

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Fuchs, K. (2023). Exploring the opportunities and challenges of NLP models in higher education: Is Chat GPT a blessing or a curse? *Frontiers in Education, 8*, 1166682. https://doi.org/10.3389/feduc.2023.1166682

Glaser, N. (2023). Exploring the potential of ChatGPT as an educational technology: An emerging technology report. *Technology, Knowledge and Learning, 28*(4), 1945–1952. https://doi.org/10.1007/s10758-023-09684-4

Gong, X., Tang, Y., Liu, X., Jing, S., Cui, W., Liang, J., & Wang, F.-Y. (2020). K-9 artificial intelligence education in Qingdao: Issues, challenges and suggestions. *Proceedings of the 2020 IEEE International Conference on Networking, Sensing and Control (ICNSC)*, 1–6. IEEE. https://doi.org/10.1109/ICNSC48988.2020.9238087

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences, 11*(12), 5467. https://doi.org/10.3390/app11125467

Graham, S. (2018). Instructional Feedback in Writing. In *The Cambridge handbook of instructional feedback,* 145–168. Cambridge University Press. https://doi.org/10.1017/9781316832134.009

Graham, S., Harris, K. R., & Santangelo, T. (2015a). Research-based writing practices and the common core: Meta-analysis and meta-synthesis. *Elementary School Journal, 115*, 498–522. https://doi.org/10.1086/681964

Graham, S., Hebert, M., & Harris, K. R. (2015b). Formative assessment and writing: A meta-analysis. *Elementary School Journal, 115*, 523–547. https://doi.org/10.1086/681947

Han, T., & Sari, E. (2022). An investigation on the use of automated feedback in Turkish EFL students' writing classes. *Computer Assisted Language Learning*, 1–24. https://doi.org/10.1080/09588221.2022.2067179

Han, Y., & Xu, Y. (2020). The development of student feedback literacy: The influences of teacher feedback on peer feedback. *Assessment & Evaluation in Higher Education, 45*(5), 680–696. https://doi.org/10.1080/02602938.2019.1689545

Hattie, J. (2012). Visible learning for teachers: Maximizing impact on learning. Routledge, Taylor & Francis Group.

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

Henderson, M., Phillips, M., Ryan, T., Boud, D., Dawson, P., Molloy, E., & Mahoney, P. (2019). Conditions that enable effective feedback. *Higher Education Research & Development, 38*(7), 1401–1416. https://doi.org/10.1080/07294360.2019.1657807

Huang, S., & Renandya, W. A. (2020). Exploring the integration of automated feedback among lower-proficiency EFL learners. *Innovation in Language Learning and Teaching, 14*(1), 15–26. https://doi.org/10.1080/17501229.2018.1471083

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, e208. https://doi.org/10.7717/peerj-cs.208

Jacobsen, L. J., & Weber, K. E. (2023). The promises and pitfalls of ChatGPT as a feedback provider in higher education: An exploratory study of prompt engineering and the quality of AI-driven feedback. OSF preprints, 2023. https://doi.org/10.31219/osf.io/cr257

Jang, E. E., Hunte, M., Barron, C., & Hannah, L. (2023). Exploring the role of self-regulation in young learners' writing assessment and intervention using BalanceAI automated diagnostic feedback. In *Fundamental considerations in technology mediated language assessment*, 31–48. Routledge. https://doi.org/10.4324/9781003292395-4

Jansen, T., Höft, L., Bahr, L., Fleckenstein, J., Möller, J., Köller, O., & Meyer, J. (2024). Empirische Arbeit: Comparing Generative AI and Expert Feedback to Students' Writing: Insights from Student Teachers. *Psychologie in Erziehung Und Unterricht, 71*(2), 80–92. https://doi.org/10.2378/peu2024.art08d

Jia, Q., Young, M., Yunkai, X., Jialin, C., Chengyuan L., Rashid, P., & Gehringer, E. (2022). InstaReviewer: A data-driven approach for generating instant feedback on students' project reports. *Proceedings of the 15th International Conference on Educational Data Mining*, 1-12. International Educational Data Mining Society. 10.5281/ZENODO.6853099

Katz, A., Wei, S., Nanda, G., Brinton, C., & Ohland, M. (2023). Exploring the efficacy of ChatGPT in analyzing student teamwork feedback with an existing taxonomy. *arXiv preprint* arXiv:2305.11882.

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing, 44*, 100450. https://doi.org/10.1016/j.asw.2020.100450

Liang, K.-H., Davidson, S., Yuan, X., Panditharatne, S., Chen, C.-Y., Shea, R., Pham, D., Tan, Y., Voss, E., & Fryer, L. (2023). ChatBack: Investigating methods of providing grammatical error feedback in a GUI-based language learning chatbot. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 83–99. https://doi.org/10.18653/v1/2023.bea-1.7

Lin, J., Dai, W., Lim, L.-A., Tsai, Y.-S., Mello, R., Khosravi, H., Gasevic, D., & Chen, G. (2022). Learner-centred analytics of feedback content in higher education. *EdArXiv*. https://doi.org/10.35542/osf.io/ub5dy

Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning, 35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323

Lossio-Ventura, J. A., Weger, R., Lee, A. Y., Guinee, E. P., Chung, J., Atlas, L., Linos, E., & Pereira, F. (2024). A comparison of chatgpt and fine-tuned open pre-trained transformers (opt) against widely used sentiment analysis tools: Sentiment analysis of COVID-19 survey data. *JMIR Mental Health, 11*, e50150. https://doi.org/10.2196/50150

Lu, X. (2019). An empirical study on the artificial intelligence writing evaluation system in China CET. *Big Data, 7*(2), 121–129. https://doi.org/10.1089/big.2018.0151

Maclellan, E. (2005). Academic achievement: The role of praise in motivating students. *Active Learning in Higher Education, 6*(3), 194–206. https://doi.org/10.1177/1469787405057750

Marrs, S., Zumbrunn, S., McBride, C., & Stringer, J. K. (2016). Exploring elementary student perceptions of writing feedback. *Journal on Educational Psychology, 10*(1), 16–28.

Matcha, W., Gašević, D., Uzir, N. A., Jovanović, J., & Pardo, A. (2019). Analytics of learning strategies: Associations with academic performance and feedback. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 461–470. https://doi.org/10.1145/3303772.3303787

McLaren, B. M., DeLeeuw, K. E., & Mayer, R. E. (2011). A politeness effect in learning with web-based intelligent tutors. *International Journal of Human-Computer Studies, 69*(1–2), 70–79. https://doi.org/10.1016/j.ijhcs.2010.09.001

Mizumoto, A., & Eguchi, M. (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

Murphy, R. (2019). Artificial intelligence applications to support K-12 teachers and teaching: A review of promising applications, challenges, and risks. *RAND Corporation*. https://doi.org/10.7249/PE315

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education, 71*, 56–76. https://doi.org/10.1016/j.compedu.2013.09.011

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

O'Cain, A., Fedoruk, B. D., Masri, Z., Frost, R., & Alahmar, A. (2023). A system for the improvement of educational assessment using intelligent conversational agents. *SSRN Electronic Journal*, 1–9. https://doi.org/10.2139/ssrn.4393234

Osakwe, I., Chen, G., Whitelock-Wainwright, A., Gašević, D., Pinheiro Cavalcanti, A., & Ferreira Mello, R. (2022). Towards automated content analysis of educational feedback: A multi-language study. *Computers & Education: Artificial Intelligence, 3*, 100059. https://doi.org/10.1016/j.caeai.2022.100059

Osawa, K. (2023). Integrating automated written corrective feedback into e-portfolios for second language writing: Notion and notion AI. *RELC Journal,* 00336882231198913. https://doi.org/10.1177/00336882231198913

Parikh, A., McReelis, K., & Hodges, B. (2001). Student feedback in problem based learning: A survey of 103 final year students across five Ontario medical schools. *Medical Education, 35*(7), 632–636. https://doi.org/10.1046/j.1365-2923.2001.00994.x

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15*(2), 68–85. https://doi.org/10.1016/j.asw.2010.05.004

Pennebaker, J. W., Boyd, R. L., Booth, R. J., Ashokkumar, A., & Francis, M. E. (2022). *Linguistic Inquiry and Word Count: LIWC-22*. Pennebaker Conglomerates. https://www.liwc.app

Poulos, A., & Mahony, M. J. (2008). Effectiveness of feedback: The students' perspective. *Assessment & Evaluation in Higher Education, 33*(2), 143–154. https://doi.org/10.1080/02602930601127869

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Ryan, T., Henderson, M., Ryan, K., & Kennedy, G. (2021). Designing learner-centred text-based feedback: A rapid review and qualitative synthesis. *Assessment & Evaluation in Higher Education, 46*(6), 894–912. https://doi.org/10.1080/02602938.2020.1828819

Ryan, T., Henderson, M., Ryan, K., & Kennedy, G. (2023). Identifying the components of effective learner-centred feedback information. *Teaching in Higher Education, 28*(7), 1565–1582. https://doi.org/10.1080/13562517.2021.1913723

Sanosi, A. B. (2022). The impact of automated written corrective feedback on EFL learners' academic writing accuracy. *Journal of Teaching English for Specific and Academic Purposes, 301*. https://doi.org/10.22190/JTESAP2202301S

Schirmer, B. R., & Bailey, J. (2000). Writing assessment rubric: An instructional approach with struggling writers. *Teaching Exceptional Children, 33*(1), 52–58. https://doi.org/10.1177/004005990003300110

Sein, M. (2022). AI-assisted knowledge assessment techniques for adaptive learning environments. *Computers and Education: Artificial Intelligence, 3*, 100050. https://doi.org/10.1016/j.caeai.2022.100050

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Stern, L. A., & Solomon, A. (2006). Effective faculty feedback: The road less traveled. *Assessing Writing, 11*(1), 22–41. https://doi.org/10.1016/j.asw.2005.12.001

Tian, L., & Zhou, Y. (2020). Learner engagement with automated feedback, peer feedback and teacher feedback in an online EFL writing context. *System, 91*, 102247. https://doi.org/10.1016/j.system.2020.102247

Troia, G. A. (2006). Writing instruction for students with learning disabilities. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 324–336). Guilford.

Vasilyeva, E., Puuronen, S., Pechenizkiy, M., & Rasanen, P. (2007). Feedback adaptation in web-based learning systems. *International Journal of Continuing Engineering Education and Life-Long Learning, 17*(4/5), 337. https://doi.org/10.1504/IJCEELL.2007.015046

Wang, X., Lee, Y., & Park, J. (2022). Automated evaluation for student argumentative writing: A survey. *arXiv preprint* arXiv:2205.04083. *https://doi.org/10.48550/ARXIV.2205.04083*

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies, 66*(2), 98–112. https://doi.org/10.1016/j.ijhcs.2007.09.003

Wang, Z., & Han, F. (2022). The effects of teacher feedback and automated feedback on cognitive and psychological aspects of foreign language writing: A mixed-methods research. *Frontiers in Psychology, 13*, 909802. https://doi.org/10.3389/fpsyg.2022.909802

Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education, 31*(3), 379–394. https://doi.org/10.1080/02602930500353061

Wei, P., Wang, X., & Dong, H. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. *Frontiers in Psychology, 14*, 1249991. https://doi.org/10.3389/fpsyg.2023.1249991

Woo, D. J., Susanto, H., Yeung, C. H., Guo, K., & Fung, A. K. Y. (2023). Exploring AI-generated text in student writing: How does AI help? *arXiv preprint* arXiv:2304.02478.

Wu, H., Wang, W., Wan, Y., Jiao, W., & Lyu, M. (2023). Chatgpt or Grammarly? Evaluating ChatGPT on grammatical error correction benchmark. *arXiv preprint* arXiv:2303.13648

Yoon, S. Y., Miszoglad, E., & Pierce, L. R. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. *arXiv preprint* arXiv:2310.06505.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0

Zhang, P., & Tur, G. (2023). A systematic review of ChatGPT use in K-12 education. *European Journal of Education*, 1–22. https://doi.org/10.1111/ejed.12599

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). Can ChatGPT understand too? a comparative study on ChatGPT and fine-tuned Bert. *arXiv preprint* arXiv:2302.10198.