Education and Information Technologies (2024) 29:8435–8463
https://doi.org/10.1007/s10639-023-12146-0

To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing

Kai Guo[1] · **Deliang Wang**[1]

Received: 28 April 2023 / Accepted: 14 August 2023 / Published online: 29 August 2023

Abstract

ChatGPT, the newest pre-trained large language model, has recently attracted unprecedented worldwide attention. Its exceptional performance in understanding human language and completing a variety of tasks in a conversational way has led to heated discussions about its implications for and use in education. This exploratory study represents one of the first attempts to examine the possible role of ChatGPT in facilitating the teaching and learning of writing English as a Foreign Language (EFL). We examined ChatGPT's potential to support EFL teachers' feedback on students' writing. To reach this goal, we first investigated ChatGPT's performance in generating feedback on EFL students' argumentative writing. Fifty English argumentative essays composed by Chinese undergraduate students were collected and used as feedback targets. ChatGPT and five Chinese EFL teachers offered feedback on the content, organisation, and language aspects of the essays. We compared ChatGPT- and teacher-generated feedback in terms of their amount and type. The results showed that ChatGPT produced a significantly larger amount of feedback than teachers and that compared with teacher feedback, which mainly focused on content-related and language-related issues, ChatGPT distributed its attention relatively equally among the three feedback foci (i.e., content, organisation, and language). Our results also indicated that ChatGPT and teachers displayed tendencies towards using different feedback types when evaluating different aspects of students' writing. Additionally, we examined EFL teachers' perceptions of using ChatGPT-generated feedback to support their own feedback. The five teachers reported both positive and negative perceptions of the features of ChatGPT feedback and the relation between ChatGPT and teacher feedback. To foster EFL students' writing skills, we suggest that teachers collaborate with ChatGPT in generating feedback on student writing.

Keywords EFL writing · Automated writing evaluation · Teacher feedback · ChatGPT · Human–machine collaboration

Extended author information available on the last page of the article

1 Introduction

Teacher feedback plays an important role in the teaching and learning of writing English as a Foreign Language (EFL). As a crucial source of students' perceived self-efficacy (Dujinhower et al., 2010), teacher feedback can bring many benefits to students, such as helping them identify areas in need of improvement and encouraging additional practice attempts (McMartin-Miller, 2014). However, providing feedback is time-consuming and can result in teacher burnout (Lee, 2014). Feedback given by teachers who teach large size classes can be limited (Yang et al., 2006), and such teachers often suffer from the tedium of correcting student essays (Hyland, 1990). Alternatives to teacher feedback have been used to help address these problems. Generally, two types of supplementary feedback are used in the writing classroom: *peer feedback* and *machine feedback*. The former involves classmates assessing each other's texts and providing comments and suggestions, and the latter involves using software to automatically generate feedback on students' writing. Compared with teacher feedback and peer feedback, the principal advantage of machine feedback is its efficiency: automated writing evaluation (AWE) systems can produce immediate feedback on students' writing (Shermis & Burstein, 2013).

In recent decades, a number of AWE systems, such as Grammarly (Koltovskaia, 2020), eRevise (Wang et al., 2020), and Pigai (Geng & Razali, 2020), have been developed, which have helped increase teachers' feedback, expedite the process of providing feedback (Wilson & Czik, 2016), accelerate the practice–feedback loop (Kellogg et al., 2010), and lessen teachers' feedback burden (Ranalli, 2018). Empirical evidence has shown that combining teacher feedback with AWE can enable teachers to be selective in the type of feedback they provide, thereby increasing students' writing motivation, writing persistence, and writing quality (Grimes & Warschauer, 2010; Link et al., 2022; Wilson & Czik, 2016). Notably, although studies have indicated the usefulness of AWE tools for correcting and improving surface-level features (e.g., grammatical and mechanical errors) of students' writing, existing AWE systems' capabilities of evaluating deep-level features (e.g., organisation and content) have been repeatedly questioned (Burstein et al., 2004; Fu et al., 2022; Hearst, 2000). As a result, the use of AWE feedback might lead to an excessive focus on surface-level features and inadequate attention to propositional content on the part of students (Li et al., 2015). In practice, many EFL teachers tend to use AWE as a complement to their feedback. That is, teachers use AWE systems to provide feedback on the formal features of student essays, thereby gaining more time to deal with higher-level issues (Stevenson, 2016). In other words, teachers collaborate with AWE systems in generating feedback on students' writing.

Successful collaboration between EFL teachers and AWE systems depends on two factors: the teachers' and systems' respective capabilities and the manner in which they collaborate. With regard to the first factor, as mentioned above, current AWE systems have been criticised for their inability to evaluate the deep-level features of students' writing. This deficiency has led to most teachers adopting a

collaboration mode in which they attend to content-related issues while machines deal with surface-level problems. Thanks to the advancement of artificial intelligence (AI) and natural language processing (NLP), machines' capabilities of understanding, analysing, and evaluating human written texts have greatly improved (Wambsganss et al., 2022; Zhu et al., 2020). When such advanced AI-enabled technologies are integrated into writing pedagogy and used to assess students' writing and generate automated feedback, the collaboration mode between teachers and machines may change. Before teachers can decide how to team up with AI-enabled machines, a necessary first step will be to examine AI's capabilities of feedback generation and understand the differences between EFL teachers and AI technology. This understanding will enable teachers' optimal use of AI to support their provision of feedback.

Previous studies have primarily compared feedback provided by teachers with and without the assistance of machines in order to investigate differences between them (e.g., Link et al., 2022; Wilson & Czik, 2016). Thus far, few (if any) studies have been conducted to compare feedback generated by teachers and by machines and identify differences between these two feedback providers. The reason for this research gap could be the idea that existing AWE systems are not competent enough to compare with teachers. However, with the emergence of advanced AI-powered technologies (e.g., ChatGPT), it might be necessary to reconsider this proposition and compare the feedback of teachers with that of machines. It is important to note here that our goal in making these comparisons (rather than examining machine feedback in isolation from teacher feedback) is not to examine the potential of AI to *replace* human teachers but to explore its possible role in *supporting* teacher feedback (Stevenson & Phakiti, 2014).

1.1 ChatGPT

The advancement of AI has led to the emergence of pre-trained large language models (LLMs) such as BERT (Devlin et al., 2019), which have demonstrated exceptional performance on various downstream tasks. Recently, ChatGPT (OpenAI, 2022), the latest LLM trained on a vast corpus of text data, has gained widespread popularity and attracted over 100 million users within just two months after its release. Compared with its predecessors, such as GPT-1 (Radford et al., 2018), GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020), which may produce untruthful, toxic, and unhelpful content, ChatGPT utilises the reinforcement learning from human feedback method (RLHF; Stiennon et al., 2020) to change its language modelling objective from 'predicting the next token on a webpage from the internet' to 'following users' instructions helpfully and safely' (Ouyang et al., 2022, p. 2). Accordingly, compared with previous LLMs, ChatGPT can generate more truthful and less toxic human-like responses to users' prompts in a conversational manner, acting as a chatbot, and has become a versatile tool with a range of applications, such as composing poetry, commenting on news, debugging code, writing essays, and summarising literature (Taecharungroj, 2023; van Dis et al., 2023).

In the realm of education, researchers have been discussing ChatGPT's implications for teaching and learning (Garcia-Peñalvo, 2023; Rudolph et al., 2023; Zhai, 2022) and exploring its potential use in various subjects, such as journalism and media (Pavlik, 2023), medicine (Gilson et al., 2022; Khan et al., 2023), and engineering education (Qadir, 2022). For example, Kung et al. (2023) found that ChatGPT demonstrated a high level of performance that was either at or very close to the passing threshold for the United States Medical Licensing Examination, suggesting its great potential in assisting learners with medical education. Despite its potential contribution to the education field, ChatGPT has been criticised for its potentially negative effects on student learning, such as facilitating cheating. Some universities and institutions have even banned their students from using ChatGPT for classroom, coursework, and assessment tasks (Yau & Chan, 2023). Such policies seem to be due to our limited knowledge of ChatGPT's potential.

Arguably, because ChatGPT is equipped with the capacity to comprehend language patterns and connections (Rudolph et al., 2023), its prospects for supporting writing-related tasks in language learning, such as correcting grammatical mistakes and rephrasing sentences, are promising. Despite the ongoing heated discussion about whether ChatGPT should be resisted or embraced by teachers and students, few empirical studies have been conducted to examine the role of this newly developed AI tool in the EFL classroom. To the best of our knowledge, there is no empirical evidence showing ChatGPT's potential to support writing teachers' feedback provision. Therefore, we explored how EFL teachers might be able to collaborate with ChatGPT in generating feedback on student essays.

1.2 Research questions

In this exploratory study, we aimed both to investigate ChatGPT's capability to evaluate and provide feedback on EFL students' writing and to understand its potential to support teacher feedback. To reach this goal, the following two research questions were proposed to guide the investigation:

RQ1: What are the differences between ChatGPT- and teacher-generated feedback on EFL students' writing?
RQ2: How do EFL teachers perceive ChatGPT-generated feedback on EFL students' writing?

RQ1 examined ChatGPT's capability of producing feedback on student writing by comparing its feedback performance with that of EFL teachers, thus providing a complete understanding of each feedback provider's strengths and weaknesses. Based on that understanding, the collaboration between ChatGPT and EFL teachers can be properly planned. RQ2 explored ChatGPT's feedback performance from the perspective of EFL teachers because they will be collaborating with ChatGPT, and their perceptions of this collaborator are of great importance. In other words, teachers' perceptions will influence whether an AWE technology is appropriately utilised

for classroom instruction and provide an important source of evidence when examining the efficacy of that technology (Wilson et al., 2021).

2 Methods

2.1 Participants

We recruited five Chinese EFL teachers (one male and four females) using the convenience sampling method. As shown in Table 1, Teachers Y and H, each of whom had 7 years of experience teaching English to undergraduates, were the most senior participants. Teachers X, M, and L had been teaching English to undergraduate students for 5, 5, and 3 years, respectively. Teacher Y held a doctoral degree, and each of the other participants held a master's degree. Teacher L always gave feedback on her students' writing. Teachers Y, H, and X often provided students with writing feedback. Teacher M did not offer feedback very frequently. Teachers Y, M, and L were experienced in using technology for feedback provision, but in different ways – Teachers Y and M asked the students themselves to use AWE tools (e.g., Pigai and Grammarly) to obtain feedback, whereas Teacher L also used AWE tools to assist in her own feedback generation. We obtained the teachers' consent to participate in the study.

2.2 Data collection

Fifty English argumentative essays written by Chinese undergraduate students (24 males and 26 females) were used as the feedback provision targets in this study. The students ranged from 18 to 21 years of age, with a mean age of 19.54 years (SD = 0.48). The students demonstrated proficiency in English, with proficiency levels ranging from B2 to C1 of the Common European Framework of Reference for Languages. The students came from diverse academic backgrounds, including 21 engineering students,

Table 1 Profile of the participating EFL teachers

Teacher	Gender	Age	Educational background	Years of English teaching	Feedback frequency	Technology use for feedback provision
Y	Male	40	Doctoral degree	7	Often	Experienced
H	Female	33	Master's degree	7	Often	Novice
X	Female	32	Master's degree	5	Often	Novice
M	Female	32	Master's degree	5	Sometimes	Experienced
L	Female	26	Master's degree	3	Always	Experienced

'Feedback frequency' was operationalised as the ratio of teacher feedback instances provided on student writing. It was calculated as the number of feedback instances given by the teacher divided by the total number of writing tasks assigned to students. 'Never', 'seldom', 'sometimes', 'often', and 'always' meant that the teacher provided 0, 1–2, 3–5, 6–8, and 9–10 instances of teacher feedback, respectively, among 10 writing assignments in their teaching

17 science students, and 12 business administration students. All of the students were enrolled in an academic English course, and the essays were submitted as an after-class assignment, contributing to 10% of their final course score.

The writing task prompted the students to write an essay of at least 300 words about whether 'college students should base their choice of a field of study on the availability of jobs in that field' (adapted from the Graduate Record Examination's 'Analyze an Issue' task). The students completed the essays as after-class assignments. The 50 essays ranged from 314 to 455 words in length, with an average of 375 words. Following the categorisation of feedback focus commonly adopted in previous studies (e.g., Alshuraidah & Storch, 2019; Guo et al., 2022; Yang et al., 2006), this study examined the differences between ChatGPT and teacher feedback on students' argumentative writing concerning three aspects: *content*, *organisation*, and *language*. Specifically, *content feedback* focused on the quality and development of arguments (e.g., supporting evidence and logical reasoning); *organisation feedback* related to issues such as the absence of topic sentences in paragraphs and the inadequate summarisation of ideas in the conclusion; and *language feedback* pertained to grammatical errors, word choice, and mechanics (e.g., spelling and formatting).

ChatGPT was used to evaluate the 50 essays and generate feedback. Three prompts were created to obtain feedback outputs from ChatGPT: (1) *please provide comments and suggestions on the content of the argumentative essay*; (2) *please provide comments and suggestions on the organisation of the argumentative essay*; and (3) *please provide comments and suggestions on the language of the argumentative essay*. Notably, we only used the first output from ChatGPT in response to each prompt, although ChatGPT had a 'try again' function and could generate multiple different outputs in response to the same prompt. The outputs were collected by a research assistant in the US (ChatGPT was not accessible in China at the time of the study) over a period of 5 days from 14 to 18 December 2022.

At the same time, each participating teacher was assigned to provide feedback on 10 of the 50 essays. To investigate the differences between ChatGPT feedback and teacher feedback, the teachers were given the same prompts as ChatGPT to evaluate the three aspects (i.e., content, organisation, and language) of the essays. To ensure that the teachers would generate feedback in the manner in which they were accustomed in their previous teaching practices, we did not provide them with any other instructions or guidelines. The teachers were required to complete the feedback task within 2 weeks and to note the time they spent commenting on each essay. Overall, the five teachers spent 81 (Teacher Y), 134 (Teacher H), 92 (Teacher X), 305 (Teacher M), and 48 min (Teacher L), respectively, on their comments.

After the teachers' submission of their feedback, the feedback generated by ChatGPT was shared with the corresponding teacher. After receiving the ChatGPT feedback and comparing it with their own feedback, the teachers responded to a questionnaire that asked about (1) their evaluation of ChatGPT feedback quality (on content, organisation, and language) and (2) their perceptions of ChatGPT's potential to support teacher feedback. This part of the study was designed to help us further understand ChatGPT's possible role in supporting teacher feedback. The teachers were first asked to rate their perceived quality of ChatGPT-generated feedback (e.g., *ChatGPT's feedback on the content of student essays is of high quality*) and their perceived usefulness

of ChatGPT for assisting teacher feedback (i.e., *ChatGPT will be useful for supporting my feedback provision*) on a 5-point Likert scale (ranging from 1 = 'strongly disagree' to 5 = 'strongly agree') and then provide reasons for their ratings.

2.3 Data analysis

2.3.1 Data analysis for answering RQ1

To address RQ1, each feedback message generated by ChatGPT and the participating teachers was first parsed into feedback units (i.e., idea units) according to the procedures described in Cho et al. (2006) and Link et al. (2022); a feedback unit was defined as a stand-alone message addressing a single problem or targeting a single feature of the text. To ensure accuracy, all of the feedback messages were parsed into idea units by both authors. All differences were resolved via consensus. This process resulted in identifying 1,284 ChatGPT feedback units and 547 teacher feedback units across the 50 essays (see Appendix Table 8 for details). Notably, among the 1,284 feedback units generated by ChatGPT, 229 units (18%) were found to be off-task, meaning that these comments were not related to the required feedback focus (content, organisation, or language). For example, the comment below is ChatGPT's output in response to the prompt requiring it to provide comments and suggestions on the *organisation* of essay number 3. In fact, ChatGPT was commenting on the *content* of the essay:

> *Overall, the essay effectively argues against the idea that college students should base their choice of a major on the availability of jobs in that field.* [Essay number 3; ChatGPT feedback on organisation].

These 229 feedback units were excluded. In contrast, no off-task comments were found in the teachers' feedback. As a result, 1,055 ChatGPT feedback units and 547 teacher feedback units (see Appendix Table 9 for details) were included in our subsequent analysis.

We categorised the type of feedback provided by ChatGPT and by the teachers. *Feedback type* is a crucial feedback feature that has been extensively investigated in the literature. It relates to the manner in which feedback is presented to the writer. Drawing on the typologies of feedback used in Cho et al. (2006) and Wilson and Czik (2016), we developed a coding scheme for classifying the type of ChatGPT- and teacher-generated feedback (see Appendix Table 10). The coding framework included five feedback types. First, *directive* refers to feedback that directs students to add, remove, or modify text or that takes the form of direct editing. Second, *informative* pertains to feedback that provides students with information to consider when revising but does not direct them to make a specific revision. Third, *query* refers to feedback that asks clarifying questions. Fourth, *praise* relates to positive or encouraging remarks aimed at the author or specific parts of the text. Fifth, *summary* recapitulates the main points of the essay or a portion thereof.

The self-contained feedback segments were coded using the coding framework. Our coding process, in accordance with the approach used by Link et al. (2022), consisted of three phases, which were implemented to ensure the reliability and validity of the data analysis procedure. In the first phase, the two authors independently coded a random 10% of feedback units ($n = 160$) and discussed discrepancies between codes to add credibility to the coding process. In the second phase, we repeated the independent coding using a new randomly selected set of 10% of the data ($n = 160$), and we calculated reliability scores using Cohen's kappa. The reliability scores were as follows: directive (0.87), informative (0.84), query (0.95), praise (0.94), and summary (0.89). These scores indicated a high level of inter-coder reliability. We discussed discrepancies until a final agreement was reached. After these two phases, both of the authors were reliable enough to implement independent coding. However, due to time constraints, the second author was unable to participate in coding the remaining 80% of the data ($n = 1{,}282$). Therefore, the first author coded the remainder of the data, ensuring that the high level of inter-coder reliability achieved in the previous phases was maintained. See Appendices 11, 12, 13, and 14 for the amounts of feedback types generated by ChatGPT and the teachers.

After the coding, we compared the differences in both *feedback amount* and *feedback type* between ChatGPT and the EFL teachers. With regard to feedback amount, it was determined that the normal distribution assumption was not satisfied for the content, organisation, and language feedback generated by either ChatGPT or the teachers. Hence, a Mann–Whitney test was conducted to examine the differences between the two groups.

To compare feedback types, the raw counts of the five feedback types were transformed into proportions by dividing specific counts by the total number of feedback units for each essay in order to control for the variation in the feedback amount and essay length (Wilson & Czik, 2016). For example, the proportion of praise in the content aspect of an essay was calculated as its count divided by the sum of the content feedback units for the essay. The proportions of feedback types did not satisfy the assumption of normal distribution, and a Mann–Whitney test was thus performed to examine the differences between the two groups. The effect sizes of the tests were reported as r.

2.3.2 Data analysis for answering RQ2

To address RQ2, descriptive statistics were used to analyse the EFL teachers' ratings on their perceived quality of ChatGPT feedback and their perceived usefulness of ChatGPT for supporting teacher feedback. Additionally, to further understand the teachers' perceptions of ChatGPT, thematic analysis was conducted to analyse their textual responses to the questions (e.g., *Could you please give reasons for your rating?*) in order to identify recurring themes with regard to their positive and negative perceptions. Specifically, the teachers' written responses were imported into NVivo 11. Following the steps suggested by Braun and Clarke (2006) for conducting thematic analysis, the first author first repeatedly read the data and marked the data with initial codes related to teacher perceptions using *positive perceptions* and *negative perceptions* as higher-level organisation categories. Next, themes were identified inductively

from the initial codes, and then refined and given names. To ensure credibility, the second author reviewed the coding scheme in order to increase its content validity. Revisions were made until the two authors agreed on the scheme. Finally, the first and second authors applied the revised coding scheme to the data separately. The inter-coder agreement was over 95%. All of the discrepancies then underwent further discussion, and a final agreement was reached after modification, as appropriate.

3 Results

3.1 Differences between ChatGPT- and teacher-generated feedback (RQ1)

3.1.1 Differences in feedback amount

Table 2 presents the results of the Mann–Whitney test that compared the amount of feedback generated by ChatGPT with the amount of feedback generated by the teachers on the content, organisation, and language aspects of the student essays. The descriptive results showed that the teachers paid more attention to content-related issues (median = 4; range: 1–14) and language-related issues (median = 3; range: 0–18) in evaluating the student essays and less attention to organisation-related issues (median = 2; range: 1–5). In contrast, ChatGPT provided relatively equal amounts of feedback on content-related issues (median = 7; range: 2–12), organisation-related issues (median = 6; range: 1–13), and language-related issues (median = 7; range: 1–13).

Further, the Mann–Whitney test results indicated that the amount of ChatGPT feedback on content, organisation, and language was significantly higher than the amount of teacher feedback. For instance, regarding the content aspect, the median amount of feedback generated by the teachers was 4 (range: 1–14), whereas that generated by ChatGPT was 7 (range: 2–12). The test statistic was 524.500, the p-value was less than 0.001, and the effect size was -0.503, which suggested a significant difference between the amount of content feedback generated by ChatGPT and that generated by the teachers, with a large effect size in favour of the teachers generating less content feedback. Similar results were obtained in the comparisons of feedback on the other two feedback foci, with large effect sizes for both: organisation (r = -0.676) and language (r = -0.598).

Table 2 Mann–Whitney test results on feedback amount

Feedback aspect	Median (range)		Mann–Whitney	Z	Significance	r
	Teacher	ChatGPT				
Content	4 (1–14)	7 (2–12)	524.500	-5.029	$p < 0.001$ (Teacher < ChatGPT)	-0.503
Organisation	2 (1–5)	6 (1–13)	278.500	-6.757	$p < 0.001$ (Teacher < ChatGPT)	-0.676
Language	3 (0–18)	7 (1–13)	387.000	-5.982	$p < 0.001$ (Teacher < ChatGPT)	-0.598

3.1.2 Differences in feedback type

Table 3 displays the results of the Mann–Whitney test that compared the feedback types generated by the teachers and ChatGPT on the content aspect of the student essays. The descriptive statistics indicated that the teachers tended to provide directive (median = 0.20; range: 0.00–1.00) and informative feedback (median = 0.31; range: 0.00–1.00) when assessing the content of student essays, while ChatGPT was more likely to offer directive (median = 0.43; range: 0.00–1.00) and praise (median = 0.38; range: 0.00–1.00) feedback. Furthermore, the Mann–Whitney test results showed that the teachers provided a significantly greater proportion of content feedback in the form of informative and query than ChatGPT, while ChatGPT generated a significantly greater proportion of directive and praise feedback than the teachers. The effect size for the teachers providing more informative feedback was moderate (r = -0.302), and small for query (r = -0.203). The effect size for ChatGPT giving a greater proportion of directive feedback was moderate (r = -0.366) and small for praise feedback (r = -0.214). No significant difference was found in the proportion of summaries between the teachers and ChatGPT.

Table 4 displays the results of the Mann–Whitney test that compared the types of feedback generated by the teachers and ChatGPT on the organisation aspect. The descriptive statistics showed that the teachers were more likely to provide praise (median = 0.33; range: 0.00–1.00) when commenting on the organisation of student writing, whereas ChatGPT offered a more even distribution of feedback in the form of directive (median = 0.33; range: 0.00–1.00), informative (median = 0.20; range: 0.00–0.83), and praise (median = 0.33; range: 0.00–0.80). Moreover, the Mann–Whitney test results revealed that ChatGPT generated a significantly higher proportion of directive and summary feedback than the teachers. The effect size for its higher proportion of directive feedback was moderate (r = -0.336), and small for summary feedback (r = -0.199). There was no significant difference in the proportion of informative, query, and praise feedback between the teachers and ChatGPT.

Table 3 Mann–Whitney test results on feedback type (feedback focus: content)

Feedback type	Median (range)		Mann–Whitney	Z	Significance	r
	Teacher	ChatGPT				
Directive	0.20 (0.00–1.00)	0.43 (0.00–1.00)	725.00	-3.657	p < 0.001 (Teacher < ChatGPT)	-0.366
Informative	0.31 (0.00–1.00)	0.05 (0.00–0.40)	834.00	-3.020	p = 0.003 (Teacher > ChatGPT)	-0.302
Query	0.00 (0.00–0.25)	0.00 (0.00–0.00)	1150.00	-2.031	p = 0.042 (Teacher > ChatGPT)	-0.203
Praise	0.15 (0.00–1.00)	0.38 (0.00–1.00)	942.50	-2.141	p = 0.032 (Teacher < ChatGPT)	-0.214
Summary	0.00 (0.00–1.00)	0.00 (0.00–0.43)	1221.50	-0.245	p = 0.806	-0.025

Table 4 Mann–Whitney test results on feedback type (feedback focus: organisation)

Feedback type	Median (range)		Mann–Whitney	Z	Significance	r
	Teacher	ChatGPT				
Directive	0.00 (0.00–1.00)	0.33 (0.00–1.00)	773.000	-3.364	p < 0.001 (Teacher < ChatGPT)	-0.336
Informative	0.00 (0.00–1.00)	0.20 (0.00–0.83)	1198.000	-0.372	p = 0.710	-0.037
Query	0.00 (0.00–0.00)	0.00 (0.00–0.00)	1250.000	0.000	p = 1.000	0.000
Praise	0.33 (0.00–1.00)	0.33 (0.00–0.80)	1202.500	-0.333	p = 0.739	-0.033
Summary	0.00 (0.00–0.67)	0.00 (0.00–0.36)	1008.500	-1.991	p = 0.046 (Teacher < ChatGPT)	-0.199

Table 5 displays the results of the Mann–Whitney test that compared the types of feedback generated by ChatGPT and the teachers on the language aspect. The descriptive statistics indicated that the teachers were more likely to provide informative feedback (median = 0.50; range: 0.00–1.00) when evaluating the language of student essays, while ChatGPT favoured directive (median = 0.40; range: 0.00–1.00) and informative feedback (median = 0.40; range: 0.00–0.89). Furthermore, the Mann–Whitney test results demonstrated that the teachers generated a significantly larger proportion of informative and query feedback than ChatGPT, while ChatGPT produced a significantly greater proportion of directive feedback than the teachers. The effect size for the teachers providing more informative feedback was small (r = -0.217), and it was also small for query feedback (r = -0.297). The effect size for ChatGPT providing a higher proportion of directive feedback was moderate (r = -0.538). No significant difference was found in the proportion of praise and summary between the two feedback providers.

Table 5 Mann–Whitney test results on feedback type (feedback focus: language)

Feedback type	Median (range)		Mann–Whitney	Z	Significance	r
	Teacher	ChatGPT				
Directive	0.00 (0.00–1.00)	0.40 (0.00–1.00)	473.000	-5.358	p < 0.001 (Teacher < ChatGPT)	-0.538
Informative	0.50 (0.00–1.00)	0.40 (0.00–0.89)	917.000	-2.164	p = 0.030 (Teacher > ChatGPT)	-0.217
Query	0.00 (0.00–0.75)	0.00 (0.00–0.20)	995.000	-2.951	p = 0.003 (Teacher > ChatGPT)	-0.297
Praise	0.00 (0.00–1.00)	0.14 (0.00–0.43)	1110.000	-0.880	p = 0.379	-0.088
Summary	0.00 (0.00–0.00)	0.00 (0.00–0.08)	1200.500	-0.990	p = 0.322	-0.099

3.2 Teacher perceptions of ChatGPT-generated feedback (RQ2)

Table 6 presents the rating results concerning the five teachers' perceived quality of ChatGPT feedback on the content, organisation, and language aspects of student writing, and their perceived usefulness of ChatGPT for supporting teacher feedback. Four teachers (Teachers M, X, H, and Y) spoke highly of ChatGPT-generated feedback, as indicated by their high ratings (4 or 5 points) of ChatGPT's feedback quality and usefulness. In contrast, Teacher L gave low ratings (only 2 points) to ChatGPT's feedback quality on content and organisation (although she gave a relatively high rating of 4 points to its language feedback) and its usefulness for supporting teacher feedback.

Our thematic analysis of the teachers' written responses to the questionnaire provided further understanding of the reasons behind their ratings. Their perceptions were classified into positive and negative perceptions to help us understand their high and low ratings, respectively. As shown in Table 7, the teachers were found to have evaluated ChatGPT's feedback from two perspectives: (1) the features of ChatGPT's feedback and (2) the relation between ChatGPT's feedback and teacher feedback. Regarding ChatGPT's feedback features, the teachers highlighted ChatGPT's capability of evaluating the content aspect of student writing, which they thought most of the existing AWE systems were not very good at doing. They also noted that ChatGPT often started with positive comments on student essays, followed by pointing out room for improvement. Teacher M believed that this feedback feature would make students accept its feedback more easily and encourage their learning motivation. The teachers also commented that ChatGPT's feedback was detailed and specific, as it not only pointed out students' problems but also recommended solutions to the problems. More importantly, ChatGPT explained the motives or purposes for its feedback. Below is an example:

> *One suggestion for improvement could be to further clarify the structure of the essay by using subheadings or transitional phrases to indicate the change in focus from the first reason to the second reason and from the discussion of the argument to the counterargument. This would make the essay easier to follow and better guide the reader through the writer's thought process.* [Essay number 21; ChatGPT feedback on organisation]

Table 6 Scores on teachers' perceived ChatGPT feedback

Teacher	Perceived quality of ChatGPT feedback			Perceived usefulness of ChatGPT feedback for supporting teacher feedback
	Content	Organisation	Language	
Y	5	5	5	5
H	4	4	5	4
X	4	4	4	5
M	4	4	4	5
L	2	2	4	2

Table 7 Teacher perceptions of ChatGPT-generated feedback on student writing

Positive perceptions			
Theme	Sub-theme	Definition	Example
ChatGPT feedback features	Capability	Teachers think ChatGPT is capable of assessing particular aspects of student essays	ChatGPT can identify problems related to the content aspect of student essays and prompt students to provide further elaboration to support their argument construction. (Teacher H)
	Encouragement	Teachers think ChatGPT's feedback is encouraging	ChatGPT first praises what students have done well and then suggests possible revisions. This will encourage students to write and rewrite. (Teacher M)
	Specificity	Teachers think ChatGPT's feedback is specific and detailed	ChatGPT provides not only revision suggestions but also reasons for these revisions. (Teacher M)
	Flexibility	Teachers think ChatGPT's feedback is flexible	Unlike other tools that give fixed and inflexible comments, ChatGPT's feedback looks flexible. (Teacher Y)
Relation between ChatGPT feedback and teacher feedback	Supplement	Teachers think ChatGPT feedback can supplement teacher feedback	Based on ChatGPT's feedback, I can generate further feedback. (Teacher X)
	Feedback burden	Teachers think using ChatGPT can lessen their feedback burden	With this tool, my workload will be significantly reduced. (Teacher X)
	Reminder	Teachers think ChatGPT can remind teachers to pay attention to different aspects of student writing	My feedback sometimes may not cover every aspect of student essays. ChatGPT's feedback can remind me of what I have missed. (Teacher H)
Negative perceptions			
Theme	Sub-theme	Definition	Example

Table 7 (continued)

ChatGPT feedback features	Length	Teachers think ChatGPT's feedback is lengthy	ChatGPT gave very long feedback, which will make less skilled students feel stressed. (Teacher L)
	Readability	Teachers think ChatGPT's feedback is difficult to read	ChatGPT's feedback is in English. My students may be unable to read and understand it adequately. (Teacher L)
	Relevance	Teachers think ChatGPT's feedback is not relevant	ChatGPT sometimes mixes up its feedback on content and organisation. (Teacher H)
	Locatability	Teachers think ChatGPT's feedback is difficult to locate	ChatGPT cannot directly mark students' language problems in their texts, whereas Grammarly can. This will make it hard for students to find where the problems are. (Teacher X)
	Incapability	Teachers think ChatGPT is incapable of identifying particular problems in student essays	It seems that ChatGPT does not recognise some students' problem of being off-topic in their essays. (Teacher M)
Relation between ChatGPT feedback and teacher feedback	Different criteria	Teachers think ChatGPT's evaluation criteria are different from theirs	I think ChatGPT's categorisation of the three feedback foci is different from mine. (Teacher M)
	Background knowledge	Teachers think ChatGPT has no background knowledge about the class and students	I usually provide feedback according to the teaching content, course progress, and my students' prior performance. ChatGPT does not know these. (Teacher L)
	Accessibility	Teachers think ChatGPT is difficult to access	Currently, ChatGPT is not accessible in China. I cannot use it. (Teacher Y)

As presented in this example, in the first sentence of this comment, ChatGPT provided one suggestion for improving the structure of the essay evaluated; in the second sentence, it explained the purpose of this suggestion, namely to 'make the essay easier to follow and better guide the reader through the writer's thought process'.

Additionally, as Teacher Y stated, ChatGPT was able to provide flexible feedback on student writing, which was unlike other AWE systems that used predetermined feedback templates and offered fixed comments.

The teachers also identified problematic features of ChatGPT's feedback. First, Teacher L claimed that ChatGPT feedback was lengthy, which could increase students' anxiety and decrease their writing motivation. Teacher L mentioned that her students' English language proficiency was low and that ChatGPT's excessive amount of feedback would make them feel overwhelmed and stressed. In addition, she stated that ChatGPT's feedback was in English, which could be difficult for less skilled students to read and understand. This problem could hinder students' uptake of ChatGPT's feedback, even if the quality of that feedback was high. Teacher L's concerns resulted in low scores for the perceived quality of ChatGPT feedback in terms of both content and organisation aspects, along with the perceived usefulness of ChatGPT feedback for supporting teacher feedback (each of these aspects only received a score of 2 points; see Table 6). This concern emphasises how teachers take into account the recipients of ChatGPT's feedback, specifically the students, when evaluating the feedback's value. Third, Teachers M and H noted that ChatGPT sometimes provided feedback that was irrelevant to the required aspect (i.e., content, organisation, or language), which could make students feel confused. Fourth, Teachers M, X, and L mentioned that ChatGPT could not annotate student essays directly, which could waste students' time in locating the problems identified by ChatGPT. Finally, Teacher M claimed that ChatGPT seemed to be incapable of identifying the issue of being off-topic in study writing, which she considered an important criterion for evaluating student essays.

Regarding the second perspective (i.e., the relation between ChatGPT feedback and teacher feedback), the teachers also reported both positive and negative perceptions. First, the teachers foresaw that ChatGPT could supplement their feedback provision. They could provide further feedback based on the feedback generated by ChatGPT. Second, as noted by Teachers X and Y, using ChatGPT could greatly lessen teachers' burden and workload in providing feedback on student writing. Third, Teacher H noted that her own feedback was mainly focused on the content aspect of student essays, whereas ChatGPT generated much more feedback on the organisation and language aspects than she did, thus reminding her to pay more attention to what she had missed. This feature could make teacher feedback more 'comprehensive' (Teacher M).

Although the teachers considered ChatGPT as a helpful collaborator in feedback generation, they also indicated concerns regarding their collaboration with this AI technology. First, Teacher M identified some differences between her and ChatGPT's categorisation of feedback focus. For example, she thought some of ChatGPT's feedback on the content aspect was actually related to the organisation of student essays, according to her understanding. Second, Teacher L stated that feedback

should be tailored according to what the teacher had taught in their class, students' current writing ability and language proficiency, and each student's personality. As ChatGPT had no such knowledge, the collaboration between ChatGPT and teachers might be challenging. This observation underscores Teacher L's concern for her students when she evaluated ChatGPT's potential to provide helpful feedback. Third, Teacher Y mentioned that ChatGPT was not accessible in China, which would prevent him from using it to support his feedback, even though he believed ChatGPT would be a very useful tool.

4 Discussion

4.1 Differences between ChatGPT and teacher feedback

We compared ChatGPT- and teacher-generated feedback, in terms of amount and type, on the content, organisation, and language aspects of EFL students' argumentative essays. Overall, compared with teachers who tended to pay more attention to the content-related and language-related issues in students' writing, ChatGPT displayed a relatively even distribution of its attention to the three aspects (i.e., content, organisation, and language). Importantly, ChatGPT provided a significantly larger amount of feedback than teachers did, and it is important to note that this amount of feedback was created in just a few seconds. In contrast, teachers need much more time than ChatGPT to read and evaluate student essays and provide feedback. For example, Teacher M spent 305 min commenting on the 10 essays assigned to her, with each taking an average of approximately half an hour. These results indicated ChatGPT's efficiency in feedback provision. That is, like any other AWE tool, ChatGPT can produce a sufficient amount of feedback on various aspects of student writing in a very short period. This result was not unexpected, as immediate and large-quantity feedback is an important advantage of automated machines compared with human teachers (Shermis & Burstein, 2013; Warschauer & Grimes, 2008).

Additionally, we further examined the types of feedback generated by ChatGPT and the teachers. Our results suggested that the teachers and ChatGPT showed different tendencies towards feedback type when evaluating different aspects of student writing. The first important difference between them was that ChatGPT tended to offer more directive feedback than the teachers, while the teachers used informative and query feedback more frequently, especially when they commented on the content and language aspects of student essays. That is, ChatGPT directly told the students what needed to be revised, while the teachers preferred to offer feedback more indirectly. The teachers guided the students to shape their own writing either by transmitting ideas, opinions, and information (informative) or by asking them questions and requesting clarifications (query). Different types of feedback will have different impacts on students' writing performance (Nelson & Schunn, 2009). For example, Biber et al. (2011) showed that directive feedback is more effective at improving students' writing quality than informative and query feedback. However, as argued by Cho et al. (2006), 'very directed comments may lead to changes only in the specific draft and not lead to general changes in writing behaviour' (p. 263).

The second notable difference between ChatGPT and teacher feedback was that ChatGPT provided more praise than the teachers when commenting on the content of student essays. From a motivational perspective, praise could be considered as a motivator that increases students' writing or revision activities (Nelson & Schunn, 2009). In particular, as content feedback may lead to substantial revisions to student essays, providing positive remarks could encourage students' uptake of revision suggestions.

The third significant difference between ChatGPT and teacher feedback was that ChatGPT tended to provide more summaries than the teachers when evaluating the organisation aspect. This feature may help students to examine their writing at a global level and facilitate their understanding of the problems with their writing (Cho et al., 2006). Moreover, as suggested by Ferris (1997), receiving summary feedback can promote students' implementation of more feedback, which can benefit writing performance.

It should be noted that, as mentioned earlier, ChatGPT provided off-task feedback, meaning that ChatGPT did not fully follow the categorisation of feedback focus adopted in our study. In contrast, off-task feedback was not found in the feedback provided by the five EFL teachers. Some of ChatGPT's off-task feedback might be due to how we obtained its output. That is, in our study, ChatGPT was prompted to generate separate feedback on the content, organisation, and language of the student essays. When ChatGPT commented on each aspect, it tended to first create a summary of the content for the essay evaluated, and such content summaries were classified as off-task feedback when ChatGPT had been prompted to comment on the other two aspects (i.e., organisation and language). Conceivably, if we merged the three prompts into one (e.g., *Please provide comments and suggestions on the content, organisation, and language of the argumentative essay*) to ask ChatGPT to provide content, organisation, and language feedback all at once, the results would be different, and the off-task feedback problem would be resolved.

4.2 EFL teachers' perceptions of the potential collaborator

Our analysis of the EFL teachers' questionnaire responses revealed both positive and negative teacher perceptions of using ChatGPT to support teacher feedback. On the one hand, the teachers perceived ChatGPT feedback to be useful due to some of its feedback features. Above all, ChatGPT's capability to understand the content of student essays and provide appropriate revision suggestions seemed to make it a more powerful tool for AWE compared with existing AWE tools, whose capabilities are constrained by their dated technical designs. Second, ChatGPT's frequent use of praise for student achievements caught the teachers' attention. This feature was also revealed by our results for RQ1, as presented in Section 3.1.2. Third, ChatGPT adopted flexible language when giving comments, which might make its feedback more acceptable for students. Fourth, ChatGPT provided not only revision suggestions but also the rationale behind those suggestions. Including more details in feedback has been found to be more helpful than offering general comments only (Ferris, 1997). More importantly, ChatGPT's explanations, namely statements explaining the motives for the feedback or

clarifying the purpose of the feedback, could help students to understand the feedback and encourage their uptake, which would be beneficial for their writing performance (Bitchener et al., 2005).

Regarding the relation between ChatGPT and teacher feedback, the teachers believed that using ChatGPT could lessen their feedback burden and reduce their workload. These beliefs will be an important consideration for those who teach large classes. Moreover, the teachers thought that ChatGPT could not only supplement their feedback but also prompt them to pay equal attention to various aspects of student writing, which could contribute to their feedback literacy. In other words, teachers will not only collaborate with but also learn from AI. These benefits may lead to teachers' willingness to use ChatGPT for generating feedback on student writing.

On the other hand, the EFL teachers also noted ChatGPT's limitations in terms of its feedback features, including being lengthy, being difficult for students at a low language proficiency level to read and comprehend, providing irrelevant comments, being difficult to locate, and being relatively incapable of identifying certain writing problems. We argue that some of these problems can be immediately solved by revising the prompt used for obtaining ChatGPT output. For example, teachers may ask ChatGPT to shorten its output and use more readable words or even use the student's first language to write feedback. ChatGPT is able to perform such tasks.

The teachers also indicated concerns related to the relation between ChatGPT and teacher feedback. They found ChatGPT might have adopted evaluation criteria that were different from their own. ChatGPT's lack of knowledge about the class and students could cause inappropriate feedback. Importantly, inaccessibility could prevent teachers in some countries or regions from using ChatGPT. These limitations indicated that although ChatGPT seemed to be powerful in producing feedback on student writing, it could not *replace* teacher feedback. Teachers should play a role in evaluating machine feedback, even if the machine is as powerful as ChatGPT, the most state-of-the-art AI tool. Our findings support the arguments by Bai and Hu (2017) and Foltz et al. (2013), both of which argued against the idea of students using AWE systems without teacher intervention.

These positive and negative teacher perceptions both indicate the rationale for EFL teachers' collaboration with ChatGPT in feedback provision. By combining the strengths of the two feedback providers, their collaboration can contribute to students' development of their writing skills.

5 Implications, limitations, and future directions

Our findings suggest the potential of using LLM-based tools such as ChatGPT for evaluating student writing and supporting teacher feedback. We recommend that EFL teachers integrate ChatGPT feedback and their own feedback on student essays. During the process, teachers may make use of the advantages of ChatGPT feedback while paying attention to its limitations and addressing them using their own strengths. For example, ChatGPT is able to produce a large quantity of feedback in a few seconds, enabling teachers to be more selective about transmitting that feedback. They may examine, select, and adopt ChatGPT's comments as they

see fit for their students. During their selection, teachers may rely on their knowledge of the class and their students. For instance, for students who are less skilled writers, teachers may choose fewer of ChatGPT's comments and use the comments that are the most actionable. Moreover, ChatGPT pays equal attention to different aspects of student writing; importantly, it appears competent in evaluating higher-level issues in student writing. Accordingly, the conventional division of labour, in which machines provide feedback on surface-level issues and teachers offer feedback on deep-level issues, should be changed. Teachers may add further comments based on ChatGPT's comments, if necessary. Additionally, teachers may annotate student essays to clarify the location of the problems pointed out by ChatGPT. Although it seems time efficient to provide students with ChatGPT feedback directly, we suggest that teachers should first carefully examine the feedback generated by ChatGPT, modify it if necessary, and incorporate their own feedback into it before providing the final feedback to their students.

As one of the first attempts to examine ChatGPT's potential in language education settings, the present study has some noteworthy limitations that may open up avenues for future research. First, we used argumentative essays as our feedback targets. Arguably, ChatGPT and teachers' feedback on different writing genres may vary (Peterson et al., 2004). As a result, it is necessary to examine ChatGPT's feedback on other genres, such as narrative writing and expository writing, and compare its feedback with that provided by teachers to obtain a complete picture of ChatGPT's feedback ability and to seek the optimal teacher–ChatGPT collaboration mode. Moreover, we only compared the quantity and type of feedback provided by ChatGPT and by teachers. Future studies may further examine and compare their feedback quality, such as accuracy (do their comments correctly diagnose a problem or make reasonable suggestions for revision?), by involving experts.

Second, our study primarily investigated teachers' perceptions of ChatGPT feedback on student essays. However, we recommend that future research extend the scope of this study to include students' perspectives and their incorporation of feedback from both ChatGPT and teachers. To better understand the impact of these feedback sources on students' writing development, it is also important to analyse students' actual revisions in response to the feedback provided, which may reflect their uptake of the feedback (Cho et al., 2006). Additionally, as suggested by Teacher L, while the quality of ChatGPT feedback is important, its effectiveness ultimately depends on students' ability to comprehend and apply that feedback to improve their writing. It would be beneficial to examine how students' language proficiency levels and writing abilities influence their perceptions and utilisation of ChatGPT feedback. We recommend that future studies take these factors into account to gain further insights into how ChatGPT feedback can be optimised for both skilled and less-skilled student writers.

Third, this study only involved five Chinese EFL teachers. There was only one male teacher, and all five were relatively young (under 40 years old). Both gender and age are factors that can affect teachers' perception and use of technology. Therefore, future research could involve EFL teachers with diverse backgrounds so as to generalise the findings of our study.

Fourth, because of the access issue, the participating teachers were unable to use ChatGPT to generate feedback on student essays by themselves; instead, we asked a

research assistant in the US to obtain ChatGPT feedback for the teachers. In those countries or regions where ChatGPT is accessible, researchers may ask teachers to use ChatGPT directly and investigate the strategies they adopt for interacting with ChatGPT (e.g., the prompts that they use to obtain outputs) to produce feedback on student essays. Such studies will provide valuable insights into the collaboration between teachers and ChatGPT.

Fifth, we used specific prompts to elicit ChatGPT's feedback on student essays in terms of content, organisation, and language. Notably, ChatGPT is sensitive to tweaks to the input phrasing (OpenAI, 2022). That is, if its prompts change, ChatGPT will generate different outputs, and the difference between the outputs for different prompts can be significant. In the future, researchers may pay attention to the importance of *prompt programming* (Reynolds & McDonell, 2021) and attempt diverse inputs and compare their output results to identify effective ways to use ChatGPT for performing the desired task of feedback generation. For example, as the teachers in this study commented, ChatGPT may need background information about students, such as their personality and language proficiency, to provide more personalised feedback on their writing than it could provide otherwise. Researchers may attempt prompts integrated with such information when generating ChatGPT output and examine their effectiveness. Additionally, we only used the first output generated by ChatGPT in response to each prompt, although users can attempt the same prompt multiple times to obtain different outputs. It would be interesting to compare ChatGPT's multiple feedback provisions on the same student essay.

Sixth, the writers of the 50 essays evaluated in this study were not the students of the five EFL teachers who participated. Arguably, if teachers evaluate their own students' writing, they can use their knowledge about the students' backgrounds, such as the students' personalities, language proficiency levels, and prior writing performance, which can help the teachers provide more tailored feedback to individual students. This is one of teachers' advantages compared with ChatGPT's advantages. Further, another possible advantage enjoyed by teachers might be that they, unlike ChatGPT, can provide oral feedback. As suggested by Neuwirth et al. (1994) and Taylor and Hoedt (1966), speaking could increase the fluency of feedback providers' comments and facilitate the inclusion of mitigating language. Students might respond more favourably to spoken comments and consider such comments as having more integrity and being more likeable than written comments. Hence, future research may take these factors into consideration in comparing teacher feedback and ChatGPT feedback and further explore possible collaboration methods.

Finally, it is worth noting that many LLM-based tools have been developed since the launch of ChatGPT. However, our focus was on ChatGPT due to its widespread recognition and prominence in the field. It would be interesting for future research to conduct a comparative analysis between the various LLM-based tools to identify their respective strengths and weaknesses in supporting teacher feedback on student writing. Such an analysis would be beneficial to writing educators and researchers. Furthermore, since the writing of this paper, more advanced LLM-based tools, such as GPT-4, have become available. Thus, there is an opportunity for future studies to build upon our research and conduct more comprehensive investigations in this field from a broader perspective.

Appendix A

Table 8 Amount of feedback generated by ChatGPT and teachers (including ChatGPT's off-task feedback)

Student essay no	ChatGPT feedback					Teacher feedback					
	Generation date	Feedback amount				Teacher	Time used	Feedback amount			
		Content	Organisation	Language	Total			Content	Organisation	Language	Total
1–10	14 December 2022	79 (34%)	74 (32%)	76 (33%)	229 (100%)	Y	81 min	20 (29%)	20 (29%)	29 (42%)	69 (100%)
11–20	15 December 2022	84 (35%)	73 (30%)	86 (35%)	243 (100%)	H	134 min	75 (47%)	41 (26%)	42 (27%)	158 (100%)
21–30	16 December 2022	86 (32%)	85 (31%)	101 (37%)	272 (100%)	X	92 min	24 (32%)	26 (34%)	26 (34%)	76 (100%)
31–40	17 December 2022	83 (32%)	77 (30%)	102 (39%)	262 (100%)	M	305 min	51 (34%)	22 (15%)	76 (51%)	149 (100%)
41–50	18 December 2022	88 (32%)	99 (36%)	91 (33%)	278 (100%)	L	48 min	50 (53%)	18 (19%)	27 (28%)	95 (100%)
Total		420 (33%)	408 (32%)	456 (36%)	1,284 (100%)	Total	660 min	220 (40%)	127 (23%)	200 (37%)	547 (100%)

Appendix B

Table 9 Amount of feedback generated by ChatGPT and teachers (excluding ChatGPT's off-task feedback)

Student essay no	ChatGPT feedback					Teacher feedback					
	Generation date	Feedback amount				Teacher	Time used	Feedback amount			
		Content	Organisation	Language	Total			Content	Organisation	Language	Total
1–10	14 December 2022	70 (37%)	62 (32%)	59 (31%)	191 (100%)	Y	81 min	20 (29%)	20 (29%)	29 (42%)	69 (100%)
11–20	15 December 2022	62 (33%)	53 (28%)	71 (38%)	186 (100%)	H	134 min	75 (47%)	41 (26%)	42 (27%)	158 (100%)
21–30	16 December 2022	69 (35%)	54 (27%)	76 (38%)	199 (100%)	X	92 min	24 (32%)	26 (34%)	26 (34%)	76 (100%)
31–40	17 December 2022	72 (31%)	65 (28%)	95 (41%)	232 (100%)	M	305 min	51 (34%)	22 (15%)	76 (51%)	149 (100%)
41–50	18 December 2022	79 (32%)	86 (35%)	82 (33%)	247 (100%)	L	48 min	50 (53%)	18 (19%)	27 (28%)	95 (100%)
Total		352 (33%)	320 (30%)	383 (36%)	1,055 (100%)	Total	660 min	220 (40%)	127 (23%)	200 (37%)	547 (100%)

Appendix C

Table 10 Coding framework for feedback types

Feedback type	Definition	Example from participating teachers	Example from ChatGPT
Directive	Feedback that directs students to add, remove, or modify text or that takes the form of direct editing	*It will be better if you write one more paragraph at the end as the final conclusion in which you recapitulate your points*	*In the first sentence, 'It is true that availability of jobs often plays an important role in students choosing the fields of study' could be revised to 'The availability of jobs often plays a significant role in students' choices of field of study'. This change makes the sentence more concise and eliminates the unnecessary use of 'it is true that'*
Informative	Feedback that provides information for the student to consider when revising but does not direct the student to make a specific revision	*It is a bit difficult to understand this long and complex sentence*	*In terms of organisation, the essay could benefit from a more logical and structured progression of ideas*
Query	Feedback that asks clarifying questions	*Which viewpoint do you agree with?*	*Do you mean that these majors have few job openings or that they pay poorly?*
Praise	Feedback that provides positive or encouraging remarks aimed at the author or specific parts of the text	*You have done a good job of providing contextual information to frame the topic in the beginning paragraph*	*The conclusion effectively summarises the main points of the essay and reiterates the thesis*
Summary	Recapitulates the main points of the essay or a portion of the essay	*The author presents his/her arguments with examples in the body parts and restates the thesis statement in the conclusion paragraph*	*The writer argues that following one's interests leads to more success in both academics and career, and that the rapidly changing job market makes it difficult to predict which field will be most in demand in the future*

Adapted from Cho et al. (2006) and Wilson and Czik (2016)

Appendix D

Table 11 Type of feedback generated by ChatGPT and teachers (in general)

Student essay no	Feedback provider	Number of pieces of six types of feedback					Total
		Directive	Informative	Query	Praise	Summary	
1–10	ChatGPT	84 (44%)	36 (19%)	0 (0%)	56 (29%)	15 (8%)	191 (100%)
11–20	ChatGPT	61 (33%)	64 (34%)	0 (0%)	49 (26%)	12 (6%)	186 (100%)
21–30	ChatGPT	95 (48%)	46 (23%)	2 (1%)	43 (22%)	13 (7%)	199 (100%)
31–40	ChatGPT	109 (47%)	57 (25%)	0 (0%)	58 (25%)	8 (3%)	232 (100%)
41–50	ChatGPT	97 (39%)	64 (26%)	0 (0%)	73 (30%)	13 (5%)	247 (100%)
Total		446 (42%)	267 (25%)	2 (0%)	279 (26%)	61 (6%)	1,055 (100%)
Student essay no	Feedback provider	Number of pieces of six types of feedback					Total
		Directive	Informative	Query	Praise	Summary	
1–10	Teacher Y	22 (32%)	15 (22%)	12 (17%)	16 (23%)	4 (6%)	69 (100%)
11–20	Teacher H	24 (15%)	80 (51%)	1 (1%)	51 (32%)	2 (1%)	158 (100%)
21–30	Teacher X	3 (4%)	29 (38%)	0 (0%)	29 (38%)	15 (20%)	76 (100%)
31–40	Teacher M	70 (47%)	67 (45%)	1 (1%)	5 (3%)	6 (4%)	149 (100%)
41–50	Teacher L	27 (28%)	35 (37%)	6 (6%)	25 (26%)	2 (2%)	95 (100%)
Total		146 (27%)	226 (41%)	20 (4%)	126 (23%)	29 (5%)	547 (100%)

Appendix E

Table 12 Type of feedback generated by ChatGPT and teachers (on content)

Student essay no	Feedback provider	Number of pieces of feedback on content					Total
		Directive	Informative	Query	Praise	Summary	
1–10	ChatGPT	31 (44%)	5 (7%)	0 (0%)	26 (37%)	8 (11%)	70 (100%)
11–20	ChatGPT	20 (32%)	5 (8%)	0 (0%)	31 (50%)	6 (10%)	62 (100%)
21–30	ChatGPT	33 (48%)	9 (13%)	0 (0%)	22 (32%)	5 (7%)	69 (100%)
31–40	ChatGPT	37 (51%)	9 (13%)	0 (0%)	26 (36%)	0 (0%)	72 (100%)
41–50	ChatGPT	32 (41%)	13 (16%)	0 (0%)	28 (35%)	6 (8%)	79 (100%)
Total		153 (43%)	41 (12%)	0 (0%)	133 (38%)	25 (7%)	352 (100%)
Student essay no	Feedback provider	Number of pieces of feedback on content					Total
		Directive	Informative	Query	Praise	Summary	
1–10	Teacher Y	8 (40%)	2 (10%)	0 (0%)	9 (45%)	1 (5%)	20 (100%)
11–20	Teacher H	20 (27%)	29 (39%)	1 (1%)	25 (33%)	0 (0%)	75 (100%)
21–30	Teacher X	0 (0%)	7 (29%)	0 (0%)	8 (33%)	9 (38%)	24 (100%)
31–40	Teacher M	13 (25%)	30 (59%)	1 (2%)	2 (4%)	5 (10%)	51 (100%)
41–50	Teacher L	15 (30%)	14 (28%)	3 (6%)	16 (32%)	2 (4%)	50 (100%)
Total		56 (25%)	82 (37%)	5 (2%)	60 (27%)	17 (8%)	220 (100%)

Appendix F

Table 13 Type of feedback generated by ChatGPT and teachers (on organisation)

Student essay no	Feedback provider	Number of pieces of feedback on organisation					Total
		Directive	Informative	Query	Praise	Summary	
1–10	ChatGPT	25 (40%)	8 (13%)	0 (0%)	22 (35%)	7 (11%)	62 (100%)
11–20	ChatGPT	16 (30%)	18 (34%)	0 (0%)	13 (25%)	6 (11%)	53 (100%)
21–30	ChatGPT	23 (43%)	10 (19%)	0 (0%)	14 (26%)	7 (13%)	54 (100%)
31–40	ChatGPT	25 (38%)	11 (17%)	0 (0%)	21 (32%)	8 (12%)	65 (100%)
41–50	ChatGPT	29 (34%)	23 (27%)	0 (0%)	27 (31%)	7 (8%)	86 (100%)
Total		118 (37%)	70 (22%)	0 (0%)	97 (30%)	35 (11%)	320 (100%)
Student essay no	Feedback provider	Number of pieces of feedback on organisation					Total
		Directive	Informative	Query	Praise	Summary	
1–10	Teacher Y	11 (55%)	0 (0%)	0 (0%)	6 (30%)	3 (15%)	20 (100%)
11–20	Teacher H	1 (2%)	18 (44%)	0 (0%)	20 (49%)	2 (5%)	41 (100%)
21–30	Teacher X	3 (12%)	11 (42%)	0 (0%)	6 (23%)	6 (23%)	26 (100%)
31–40	Teacher M	7 (32%)	11 (50%)	0 (0%)	3 (14%)	1 (5%)	22 (100%)
41–50	Teacher L	9 (50%)	2 (11%)	0 (0%)	7 (39%)	0 (0%)	18 (100%)
Total		31 (24%)	42 (33%)	0 (0%)	42 (33%)	12 (9%)	127 (100%)

Appendix G

Table 14 Type of feedback generated by ChatGPT and teachers (on language)

Student essay no	Feedback provider	Number of pieces of feedback on language					Total
		Directive	Informative	Query	Praise	Summary	
1–10	ChatGPT	28 (47%)	23 (39%)	0 (0%)	8 (14%)	0 (0%)	59 (100%)
11–20	ChatGPT	25 (35%)	41 (58%)	0 (0%)	5 (7%)	0 (0%)	71 (100%)
21–30	ChatGPT	39 (51%)	27 (36%)	2 (3%)	7 (9%)	1 (1%)	76 (100%)
31–40	ChatGPT	47 (49%)	37 (39%)	0 (0%)	11 (12%)	0 (0%)	95 (100%)
41–50	ChatGPT	36 (44%)	28 (34%)	0 (0%)	18 (22%)	0 (0%)	82 (100%)
Total		175 (46%)	156 (41%)	2 (1%)	49 (13%)	1 (%)	383 (100%)
Student essay no	Feedback provider	Number of pieces of feedback on language					Total
		Directive	Informative	Query	Praise	Summary	
1–10	Teacher Y	3 (10%)	13 (45%)	12 (41%)	1 (3%)	0 (0%)	29 (100%)
11–20	Teacher H	3 (7%)	33 (79%)	0 (0%)	6 (14%)	0 (0%)	42 (100%)
21–30	Teacher X	0 (0%)	11 (42%)	0 (0%)	15 (58%)	0 (0%)	26 (100%)
31–40	Teacher M	50 (66%)	26 (34%)	0 (0%)	0 (0%)	0 (0%)	76 (100%)
41–50	Teacher L	3 (11%)	19 (70%)	3 (11%)	2 (7%)	0 (0%)	27 (100%)
Total		59 (30%)	102 (51%)	15 (8%)	24 (12%)	0 (0%)	200 (100%)

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability The data that support the findings of this study are available from the first author, Kai Guo, upon reasonable request.

Declarations

Conflict of interest None.

References

Alshuraidah, A., & Storch, N. (2019). Investigating a collaborative approach to peer feedback. *ELT Journal, 73*(2), 166–174. https://doi.org/10.1093/elt/ccy057

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology, 37*, 67–81. https://doi.org/10.1080/01443410.2016.1223275

Biber, D., Nekrasova, T., & Horn, B. (2011). *The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. TOEFL iBT™ research report.* Educational Testing Service.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing, 14*, 191–205. https://doi.org/10.1016/j.jslw.2005.08.001

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems* (pp. 1877–1901). Curran Associates, Inc.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine, 25*(3), 27–36. https://doi.org/10.1609/aimag.v25i3.1774

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on writing: Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication, 23*(3), 260–294. https://doi.org/10.1177/0741088306289261

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 4171–4186). Association for Computational Linguistics.

Dujinhower, H., Prins, F. J., & Stokking, K. M. (2010). Progress feedback effects on students' writing mastery goal, self-efficacy beliefs, and performance. *Educational Research and Evaluation, 16*, 53–74. https://doi.org/10.1080/13803611003711393

Ferris, D. R. (1997). The influence of teacher commentary on student revision. *TESOL Quarterly, 31*(2), 315–339. https://doi.org/10.2307/3588049

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the intelligent essay assessor. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 68–88). Routledge.

Fu, Q. K., Zou, D., Xie, H., & Cheng, G. (2022). A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*, 1–43. https://doi.org/10.1080/09588221.2022.2033787

Garcia-Peñalvo, F. J. (2023). The perception of artificial intelligence in educational contexts after the launch of chatgpt: Disruption or panic? *Education in the Knowledge Society*. https://doi.org/10.14201/eks.31279

Geng, J., & Razali, A. B. (2020). Tapping the potential of Pigai automated writing evaluation (AWE) program to give feedback on EFL writing. *Universal Journal of Educational Research*, *8*(12B), 8334–8343. https://doi.org/10.13189/ujer.2020.082638

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6), 1–44.

Gilson, A., Safranek, C., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2022). How well does ChatGPT do when taking the medical licensing exams? The implications of large language models for medical education and knowledge assessment. *medRxiv*, 1–9. https://doi.org/10.1101/2022.12.23.22283901

Guo, K., Chen, X., & Qiao, S. (2022). Exploring a collaborative approach to peer feedback in EFL writing: How do students participate? *RELC Journal*. 1–15. https://doi.org/10.1177/00336882221143192

Hearst, M. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and Their Applications, 15*(5), 22–37. https://doi.org/10.1109/5254.889104

Hyland, K. (1990). Providing productive feedback. *ELT Journal, 44*(4), 279–285. https://doi.org/10.1093/elt/44.4.279

Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research, 42*(2), 173–196. https://doi.org/10.2190/EC.42.2.c

Khan, R. A., Jawaid, M., Khan, A. R., & Sajjad, M. (2023). ChatGPT - Reshaping medical education and clinical management. *Pakistan Journal of Medical Sciences*, *39*(2), 605–607. https://doi.org/10.12669/pjms.39.2.7653

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

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198.

Lee, I. (2014). Feedback in writing: Issues and challenges. *Assessing Writing, 19*, 1–5. https://doi.org/10.1016/j.asw.2013.11.009

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing, 27*, 1–18. https://doi.org/10.1016/j.jslw.2014.10.004

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

McMartin-Miller, C. (2014). How much feedback is enough?: Instructor practices and student attitudes toward error treatment in second language writing. *Assessing Writing, 19*, 24–35. https://doi.org/10.1016/j.asw.2013.11.003

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science, 37*(4), 375–401. https://doi.org/10.1007/s11251-008-9053-x

Neuwirth, C. M., Chandook, R., Charney, D., Wojahn, P., & Kim, L. (1994). Distributed collaborative writing: A comparison of spoken and written modalities for reviewing and revising documents. *Proceedings of the Computer-Human Interaction '94 Conference, April 24–28, 1994, Boston Massachusetts* (pp. 51–57). Association for Computing Machinery.

OpenAI. (2022). *ChatGPT: Optimizing language models for dialogue*. Retrieved on 7 January 2023 from https://openai.com/blog/chatgpt/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, *78*(1). https://doi.org/10.1177/10776958221149577

Peterson, S., Childs, R., & Kennedy, K. (2004). Written feedback and scoring of sixth-grade girls' and boys' narrative and persuasive writing. *Assessing Writing, 9*(2), 160–180. https://doi.org/10.1016/j.asw.2004.07.002

Qadir, J. (2022). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. *TechRxiv. Preprint*. https://doi.org/10.36227/techrxiv.21789434.v1

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. Retrieved on 15 January 2023 from https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9. Retrieved 15 January 2023 from https://life-extension.github.io/2020/05/27/GPT%E6%8A%80%E6%9C%AF%E5%88%9D%E6%8E%A2/language-models.pdf

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning, 31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm (arXiv:2102.07350). *arXiv*. https://doi.org/10.48550/arXiv.2102.07350

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1). https://doi.org/10.37074/jalt.2023.6.1.9

Shermis, M. D., & Burstein, J. C. (2013). *Handbook of automated essay evaluation*. Routledge.

Stevenson, M. (2016). A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition, 42*, 1–16. https://doi.org/10.1016/j.compcom.2016.05.001

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2020). Learning to summarize from human feedback. *arXiv* preprint arXiv:2009.01325.

Taecharungroj, V. (2023). "What can ChatGPT do?" Analysing early reactions to the innovative AI chatbot on Twitter. *Big Data and Cognitive Computing, 7*(1), 35. https://doi.org/10.3390/bdcc7010035

Taylor, W., & Hoedt, K. (1966). The effect of praise upon the quality and quantity of creative writing. *Journal of Educational Research, 60*, 80–83. https://doi.org/10.1080/00220671.1966.10883440

van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature, 614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

Wambsganss, T., Janson, A., & Leimeister, J. M. (2022). Enhancing argumentative writing with automated feedback and social comparison nudging. *Computers & Education*, *191*, 104644. https://doi.org/10.1016/j.compedu.2022.104644

Wang, E. L., Matsumura, L. C., Correnti, R., Litman, D., Zhang, H., Howe, E., ... & Quintana, R. (2020). eRevis (ing): Students' revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, *44*, 100449. https://doi.org/10.1016/j.asw.2020.100449

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3*, 22–36. https://doi.org/10.1080/15544800701771580

Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education*, *168*, 104208. https://doi.org/10.1016/j.compedu.2021.104208

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100*, 94–109. https://doi.org/10.1016/j.compedu.2016.05.004

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179–200. https://doi.org/10.1016/j.jslw.2006.09.004

Yau, C., & Chan, K. (2023). University of Hong Kong temporarily bans students from using ChatGPT. *South China Morning Post*. Retrieved on 17 February from https://www.scmp.com/news/hong-kong/education/article/3210650/university-hong-kong-temporarily-bans-students-using-chatgpt-other-ai-based-tools-coursework

Zhai, X. (2022). ChatGPT user experience: Implications for education. Available at SSRN4312418.

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668. https://doi.org/10.1016/j.compedu.2019.103668

Authors and Affiliations

Kai Guo[1] · Deliang Wang[1]

✉ Deliang Wang
wdeliang@connect.hku.hk

Kai Guo
kaiguo@connect.hku.hk

[1] Faculty of Education, The University of Hong Kong, Hong Kong, China