International Journal of Assessment Tools in Education
2026, Vol. 13, No. 1, 248–269
https://doi.org/10.21449/ijate.1678002
journal homepage: https://dergipark.org.tr/en/pub/ijate *Research Article*

Statistical and qualitative analysis of ChatGPT and human raters in preservice teachers' writing assessment

Bahadır Gülden[1], **Huzeyfe Bilge**[2*], **Pınar Kanık Uysal**[3]

[1]Bayburt University, Faculty of Education, Department of Turkish Language Education, Bayburt, Türkiye
[2]Kafkas University, Faculty of Education, Department of Turkish Language Education, Kars, Türkiye
[3]Ordu University, Faculty of Education, Department of Turkish Language Education, Ordu, Türkiye

ARTICLE HISTORY
Received: Apr. 17, 2025
Accepted: Nov. 9, 2025

KEYWORDS
Artificial Intelligence,
ChatGPT,
Writing feedback,
Human-raters.

Abstract: Teachers spend a significant amount of time providing feedback. This study compared expert and ChatGPT assessments and feedback on written texts to determine the suitability of AI for writing skill assessments that are time-consuming to assess and provide feedback. Three experts and ChatGPT graded 14 Turkish undergraduate students' assignments using rubric that included content, language use, vocabulary, organization, and mechanics, and justified their decisions. The study involved document review and triangulation, a qualitative design. In addition, an intraclass correlation coefficient was used to assess the consistency of the ChatGPT and the experts' scores. All feedback was qualitatively analyzed to identify the strengths and weaknesses of the experts and their similarities with ChatGPT. Experts and ChatGPT had moderate to weak consistency in the writing subscales, while good reliability was found in the total score. Experts excelled in 'explanatory feedback', 'interpretation' and 'experience', while ChatGPT excelled in 'automation and continuity' and 'data processing capacity'. Experts' weaknesses included 'limited time and energy' and 'comparison bias', while ChatGPT's weaknesses were 'ambiguous expressions' and 'repetition'. The study also found that experts and ChatGPT preferred to provide constructive and supportive feedback.

INTRODUCTION

Assessment is a crucial component of language teaching, as it provides valuable information about learners' progress and guides instructional decisions. For assessments to be effective, they should be both feasible and reliable (McMillan, 2018). In language classes, developing writing skills often requires assessment formats that allow for richer, more authentic responses, most notably, open-ended questions. These questions can offer advantages over closed-ended formats by enabling students to demonstrate higher-order thinking and language production skills. However, despite their potential, open-ended questions make up only 24% of language teachers' question sets (Kanık Uysal *et al*., 2022), partly because they require more time to prepare, administer, and evaluate. Open-ended writing assessments can take 15-20 minutes per student (Wang *et al*., 2024), which may require teachers to work outside of school hours. Providing feedback to students is crucial for their learning. However, teachers struggle to

*Corresponding author at: Huzeyfe BİLGE hbilge2@gmail.com Kafkas University, Faculty of Education, Department of Turkish Language Education, Kars, Türkiye

e-ISSN: 2148-7456

manage such a workload due to time constraints. According to a global study, teachers spend 25 hours a week on non-student responsibilities, including assessment and feedback (Bryant *et al*., 2020). For teachers who even struggle to follow the curriculum (Akaya & Kurtuluş, 2011), it is difficult to assess writing skills and provide feedback using open-ended questions. This can lead to students being deprived of feedback that is vital for their learning. In addition to the time-consuming evaluation of open-ended questions, teachers also give scores to irrelevant and incorrect answers (Yavuz & Bilgeç, 2016) and make different evaluations for the same questions (Özen, 2020). Therefore, there are problems regarding the reliability of the scores given by teachers to open-ended questions. As a result, there is a need for a solution that addresses teachers' time constraints, enables the use of open-ended assessments to assess writing skills, and ensures consistent scoring of questions. Utilizing artificial intelligence (AI) tools to meet these needs can effectively solve the challenges faced.

ChatGPT can be used for personalized feedback as well as assessment (Baidoo-Anu & Owusu Ansah, 2023; Rahman & Watanobe, 2023). Teachers are already using ChatGPT in the assessment of writing skills, which requires a significant amount of time (Wang *et al*., 2024), because AI has significant potential to save teachers' time (Sok & Heng, 2023). However, research on the effectiveness of AI tools such as ChatGPT in assessing open-ended questions in various languages has yielded different results (Jackaria *et al*., 2024; Klyshbekova & Abbott, 2024). In addition to the time-consuming evaluation of open-ended questions, teachers also give scores to irrelevant and incorrect answers (Yavuz & Bilgeç, 2016) and make different evaluations for the same questions (Özen, 2020). Therefore, there are problems regarding the reliability of the scores given by teachers to open-ended questions. As a result, there is a need for a solution that addresses teachers' time constraints, enables the use of open-ended assessments to assess writing skills, and ensures consistent scoring of questions. Utilizing artificial intelligence (AI) tools to meet these needs can effectively solve the challenges faced.

ChatGPT can be used for personalized feedback as well as assessment (Baidoo-Anu & Owusu Ansah, 2023; Rahman & Watanobe, 2023). Teachers are already using ChatGPT in the assessment of writing skills, which requires a significant amount of time (Wang *et al*., 2024), because AI has significant potential to save teachers' time (Sok & Heng, 2023). However, research on the effectiveness of AI tools such as ChatGPT in assessing open-ended questions in various languages has yielded different results (Jackaria *et al*., 2024; Klyshbekova & Abbott, 2024).

Recently, the consistency of ChatGPT's text evaluations with expert assessments and its feedback mechanisms has been subject to heightened scrutiny. These studies have analyzed the feedback types that ChatGPT offers on foreign language learners' texts and juxtaposed them with expert evaluations (Guo & Wang, 2024), assessed the quality of feedback from ChatGPT versus experts (Steiss *et al*., 2024), and investigated the correlational relationship and statistical discrepancies between ChatGPT and expert assessments (Bui & Barrot, 2025; Uyar & Büyükahıska, 2025). However, these studies have various limitations. The first of these is that the studies generally evaluate English texts. Our knowledge on the comparison of ChatGPT with human raters for texts written in different languages is limited. It is important to address this limitation because ChatGPT performs differently in different languages (OpenAI, 2023) The second limitation of the studies is that quantitative designs are generally preferred. More in-depth knowledge of the scoring difference or similarity between human raters and ChatGPT could deepen our knowledge on the use of artificial intelligence applications in the evaluation of written texts. To our knowledge, no mixed methods study concurrently evaluates the efficacy of ChatGPT in assessing Turkish written texts and its merits and demerits relative to experts. It is essential to analyze ChatGPT's alignment with human evaluators in assessing written texts in various languages, as well as to elucidate its merits and shortcomings relative to human raters, to explore the reliability and advantages of employing AI. This study investigated the degree of alignment between ChatGPT (3.5 [free version]) scores for articles authored by

undergraduate Turkish language teaching students and expert evaluations. Moreover, the benefits and drawbacks of ChatGPT and human evaluators in the feedback process were qualitatively examined. The research questions were structured as follows:

1. What is the level of reliability between the scores given by the experts and ChatGPT to the written texts?
2. What are the advantages and disadvantages of experts and ChatGPT in giving feedback to written texts?

Literature

Technology has become an integral part of learning, especially since the pandemic (Klyshbekova & Abbott, 2024). With the developments in AI, the process of integrating technology and education has accelerated. ChatGPT is the best-known AI tool and it is a public platform that can provide information, support discussions and help with almost anything related to teaching (Busch & Hausvik, 2023). ChatGPT is one of the most advanced chatbots (Vargas-Murillo *et al*., 2023) and, according to Busch and Hausvik (2023), has many human-like capabilities. However, it still has significant weaknesses.

The main flaws of ChatGPT are that, unlike humans, it cannot control its responses (Kortemeyer, 2023), the quality of its responses is directly proportional to the quality of the inputs and prompts (Qadir, 2023), and the quality of the responses varies according to the language (OpenAI, 2023). Therefore, relying solely on ChatGPT can be risky for educational stakeholders.

AI and education

Although positive and negative reactions to AI have divided the educational community (Rahman & Watanobe, 2023), expectations for AI in education are high (Altundal, 2024). Due to the expected increase in the use of AI and its impact on education (Jackaria *et al*., 2024), it is crucial to identify ways in which AI can enhance teaching and learning (Klyshbekova & Abbott, 2024). The use of AI can cause major paradigm shifts in education as well as changes in the habits of teachers and students. For example, researchers state that teachers can now use new types of assessments, such as asking students to evaluate ChatGPT's responses to posed problems, interpret randomized graphs or circuit diagrams, or create graphical solutions (Kortemeyer, 2023). Therefore, the impact of AI on education needs to be carefully studied and monitored.

In many areas, including lesson planning and question generation, educational stakeholders use ChatGPT to save time (Fuchs, 2023; Kiryakova & Angelova, 2023; Qadir, 2023; Sok & Heng, 2023). However, users should be cautious when using ChatGPT, as it may occasionally provide inaccurate information and therefore should not be considered a fully reliable source (Han *et al*., 2024; Rahman & Watanobe, 2023). For example, ChatGPT has been found to provide 90% false information when compared to database information (Safdar *et al*, 2024). Despite all these concerns, the current state of AI should be considered as a preview of what it can do in the future (Qadir, 2023) and steps should be taken towards more perfect forms of it in the future. These steps include issues related to evaluation methods (Baidoo-Anu & Owusu Ansah, 2023).

AI and writing skill

The writing skills of ChatGPT have been the subject of numerous studies. ChatGPT can sometimes write better than humans and it is difficult to distinguish these texts from human-generated texts (Busch & Hausvik, 2023; Herbold *et al*., 2023; Waltzer *et al*. 2023; Yan, 2023). Moreover, the vocabulary of ChatGPT is comparable to human-generated texts (Bilge, 2024). However, it has problems with figurative meanings and text difficulty (Oğuz, 2024; Tarakçı & Tarakçı, 2024). Whatever its flaws, it is critical to determine how reliable the ChatGPT is in "assessing writing skills". Given ChatGPT's demonstrated ability to produce human-like

written content, its potential role in evaluating such content—particularly in educational contexts—warrants closer examination.

Writing is a skill that should be assessed through open-ended questions, which requires substantial time for both evaluation and feedback. Improving students' writing ability necessitates providing detailed feedback on their compositions (Graham *et al*., 2015). However, providing such feedback carries a considerable time cost for teachers. In a class of 20 students, it is estimated that a teacher would need 300–400 minutes to provide feedback on all students' written work (Wang *et al*., 2024). This demand represents an additional workload that frequently extends beyond teachers' contractual hours and places considerable strain on them. An international survey revealed that teachers spend approximately 50% of their working hours on assessment, evaluation, and feedback (Bryant *et al*., 2020), highlighting a global challenge related to teachers' workload. This persistent burden has, for decades, hindered teachers from dedicating sufficient time to improving students' writing skills (Applebee & Langer, 2011). Beyond the issue of time, concerns also remain about the reliability of teachers' scoring of written texts. Research indicates that teachers sometimes award points for incorrect answers to open-ended questions (Yavuz & Bilgeç, 2016) and assign different scores to similar answers (Özen, 2020). These challenges underscore the need for solutions that can both maximize educators' time efficiency and enhance the reliability of scoring open-ended questions.

AI applications can increase the inclusion of open-ended questions in classroom assessments (Lo, 2023). Researchers recommend rethinking traditional assessments and considering the use of AI in educational evaluation (de Winter, 2023; Klyshbekova & Abbott, 2024). In the literature, there have been studies on assessment and evaluation with open-ended questions, especially using ChatGPT. Klyshbekova and Abbott (2024) found that ChatGPT gave varying scores to its own text depending on the prompt, highlighting reliability issues. Moreover, in the same study, significant differences in ratings of experts and ChatGPT also emerged as a major problem. Nevertheless, the accuracy of ChatGPT, especially in open-ended assessments, suggests that artificial intelligence can play an important role in the evaluation of such questions.

AI and feedback

Providing feedback on students' written texts is crucial for the development of writing skills (Graham *et al*., 2015) as it enables students to understand the expectations of teachers. But it brings a significant workload for educators. Teachers spend 15-20 minutes per student on giving feedback (Parker *et al*., 2023); nevertheless, technology can alleviate this workload (Kiryakova & Angelova, 2023). Technology has played a role in assessing written assignments and providing feedback to students for years. However, these technologies often focus on a singular dimension such as grammar. ChatGPT stands out with its multidimensional functions compared to previous technologies (Wang *et al*., 2024; Yan, 2023), therefore, it has a greater potential than previous technologies for feedback on written texts.

METHOD

Research Model

This study was conducted in a mixed research design combining qualitative and quantitative methods. The mixed research design used in this study is based on Creswell's (2022) inclusive design. The purpose of this design is to analyze quantitative and qualitative data separately and combine the results. In this study, articles submitted by undergraduate students were used as documents. The students did not write these texts specifically for the study. Instead, the students had previously submitted these texts as homework for the course. The scores given to the students' texts were analyzed for consistency both among the experts and between the experts and ChatGPT. Furthermore, the rationales of the experts and ChatGPT for their scoring decisions were requested to identify qualitative differences or similarities between the scores

given by the experts and ChatGPT. The research results were structured by analyzing quantitative and qualitative data separately. Thus, we tried to see the problem from various perspectives.

Sample of Study

In the study, 14 students enrolled in the Turkish language teaching undergraduate program were defined as the sample of study. The written texts they had previously presented in the Writing Education course were used for assessment. The students are studying in the Turkish Language Teaching undergraduate program at Bayburt University's Faculty of Education and are taking Writing Education, one of the program's courses. This course provides theoretical instruction to teach the cognitive planning of writing, followed by practical writing exercises. Students structured their texts according to the subcomponents of the Flower & Hayes Writing Model (Flower & Hayes, 1981). Three experts in the field of writing evaluated these texts. Criterion sampling was used to select the texts and the following selection criteria were determined:

- Texts were written by undergraduate students learning Turkish.
- The authors of the texts should have completed Turkish Language I-II and Writing Education courses.

The experts were selected from those who have taught courses and conducted research in the field of writing education. Each expert has experience in holistic and analytical assessment and evaluation, as well as writing education. In addition, all three experts who participated in the text assessment have a PhD degree in language teaching. These texts were evaluated by three experts in the field. In order to increase the reliability and validity of the results obtained from human evaluation and to see the internal consistency among human raters, the texts were scored by three separate experts in the field.

Assessment Tools

An analytical rubric was used to evaluate texts. The texts were analyzed in the areas of content, organization, vocabulary, language use and mechanics using Jacobs *et al.*'s (1981, as cited in Weigle, 2002) analytical rubric, ranging from 34 to 100 points. Content scores were categorized as follows: 13–16 as "very poor," 17–21 as "fair to poor," 22–26 as "good to average," and 27–30 as "excellent to very good." Organization and Vocabulary were each evaluated on a scale of 7–20, with 7–9 considered "very poor," 10–13 "fair to poor," 14–17 "good to average," and 18–20 "excellent to very good." Language Use scores ranged from 5–25, with 5–10 rated as "very poor," 11–17 as "fair to poor," 18–21 as "good to average," and 22–25 as "excellent to very good." Mechanics were scored from 2–5, with 2 representing "very poor," 3 "fair to poor," 4 "good to average," and 5 "excellent to very good." As cited in Picken (1988), Jacobs *et al.* (1981) evaluated the performance of two student groups (sample sizes not reported) who completed writing tests consisting of two different tasks to examine test-retest reliability. They reported a correlation coefficient of .72 between the mean scores of the two tasks, which they described as effectively representing a test-retest reliability coefficient. A Turkish translation of the scale is also available (Ülper, 2019). Since the texts examined by AI were written in Turkish, the Turkish translation of the scale was shared with AI. The rubric, which consists of five criteria, summarizes the evaluation standards and scores as follows:

Content relates to the level of clarity of information on the topic, the level of development of the thesis, and the level of detail on the topic and ranges from 13-30 points. Organization relates to the fluency of the narrative, the degree of intellectual support and clarity, logical sequencing and coherence, and ranges from 7 to 20 points. Vocabulary refers to the selection and use of original and effective words and phrases and the fluency of the types of words used, ranging from 7 to 20 points. Language use ranges from 5 to 20 points, considering original and developed sentence structure, the degree of expression disorders and adherence to the rules of sentence structure. Mechanics focuses on spelling, punctuation and paragraphing on a 2–5-point

scale. This rubric allows for the evaluation of both content and style features of a text. Therefore, this rubric was selected as better aligned with the study's aim by enabling both experts and ChatGPT to conduct a more detailed analysis of form and content.

Data Collection Process

The first author, who is also the instructor of the course, obtained the assignments written by Turkish language teaching undergraduate students on reading skills. The author asked the students to write texts about "reading" within the scope of the course. The author chose the topic of "reading" because it is an important component of Turkish language teaching. Three domain experts scored the texts and provided feedback using Jacobs *et al.*'s (1981 as cited in Weigle, 2002) analytical rubric. Similarly, ChatGPT (3,5) scored the texts and provided feedback using the same rubric. Within the scope of the study, the zero-shot method was used for ChatGPT's scoring. Accordingly, ChatGPT was not trained in any way. Although it is expected to give more qualified results if it is trained, it is considered that many teachers will use ChatGPT more frequently with the zero-shot method since they will not know this technical detail.

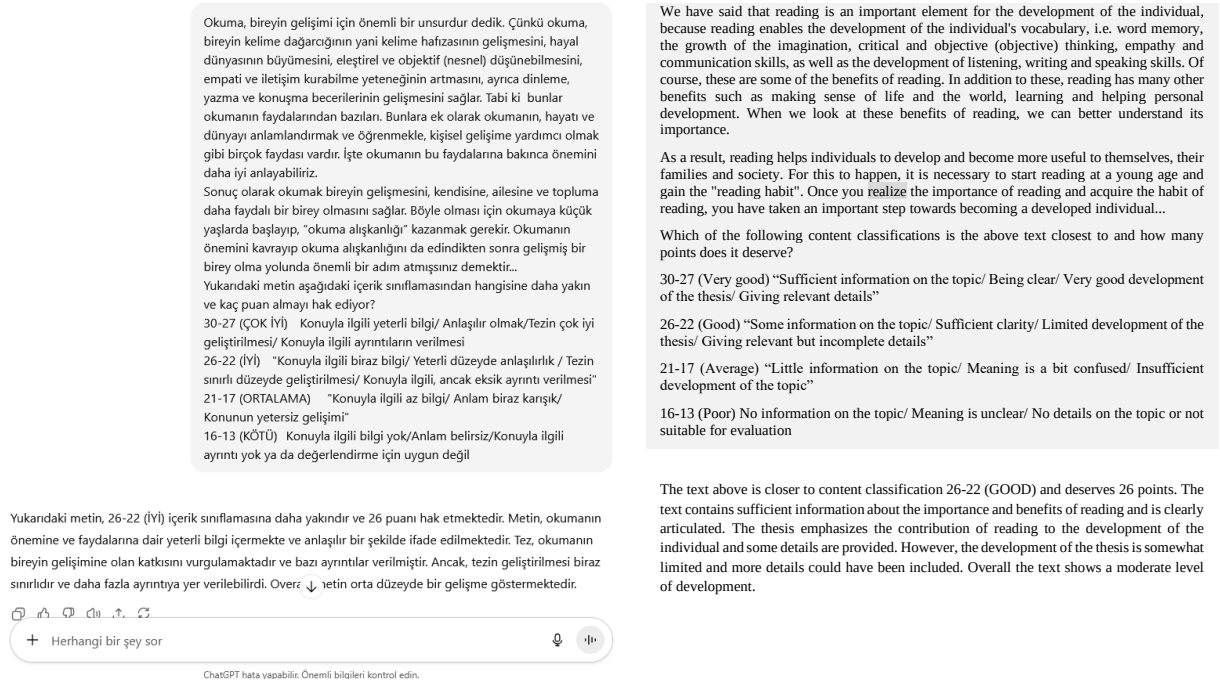

Figure 1. *An example of scoring and evaluation of the language use sub-dimensions of ChatGPT's analytical scoring rubric. a) the Turkish version, b) the Englis version*

ChatGPT's evaluation process: Before using it as a scorer, ChatGPT was introduced to each sub-dimension of the rubric (content, organization, vocabulary, language use, mechanics) and the standards of these sub-dimensions. ChatGPT was also asked to explain the rationale behind the score it assigned to the given text. Before the assessment and evaluation process, ChatGPT was asked to consider itself as a language expert and to evaluate and provide feedback accordingly. In addition, the topic and main idea of the texts were explained to ChatGPT before the assessment, and it was asked to take this information into account when scoring the text. After providing this information, ChatGPT was asked "Can you score the text I will share with you according to the content criteria I have given above?", "Can you explain why you think the text deserves this score and justify it with its positive and negative aspects?". Each text was evaluated under a single criterion in turn, and after the evaluation of one criterion was completed for all texts, the other criterion was started. For each sub-dimension of the rubric,

the texts were shared with ChatGPT repeatedly. Figure 1 shows an example of ChatGPT's scoring and evaluation of the language use sub-dimensions of the analytic rubric.

Expert scoring process: The same evaluation process was carried out with experts, and they were asked to score each text according to Jacobs *et al.*'s (1981 as cited in Weigle, 2002) analytical rubric and to justify their scores according to the positive and negative aspects of the texts. The field experts who scored and evaluated the texts were informed about the purpose of the study and the data collection tool, but they were not told how many raters participated or who they were. Thus, it was ensured that each field expert independently evaluated the student texts in question.

Data Analysis

Descriptive statistics were employed to analyze the quantitative data by calculating the average of the experts' ratings. In addition, the intraclass correlation coefficient (ICC(3,k), two-way mixed-effects model, absolute agreement) was used to examine the reliability between the ratings. A total of 14 papers were rated by the experts. All analysis was conducted using SPSS. ICC values were interpreted according to the thresholds suggested by Koo and Li (2016), where < 0.50 indicates poor, 0.50–0.75 moderate, 0.75–0.90 good, and > 0.90 excellent reliability. Content analysis was preferred to analyze qualitative data, and the process was carried out in line with the steps of thematic analysis (Braun & Clarke, 2006). In this context, codes represented the smallest meaningful units derived from the data. Similar codes were grouped together to form categories, which in turn were organized into broader themes that captured higher-order patterns relevant to the research questions. When necessary, sub-dimensions were identified to provide a more detailed representation of the themes. This hierarchical structure ensured a clear distinction between analytic units (Braun & Clarke, 2006; Miles & Huberman, 1994). First, the data were analyzed as a whole by recording the expert and ChatGPT feedback separately and in sequence for each text. For example, for Text 1, the scores and feedback of ChatGPT, Expert 1, Expert 2, and Expert 3 were recorded in sequence, and the scores and feedback of each expert on the same text were evaluated from a holistic perspective. After several readings of this file, two experts in the field of Turkish education undertook independent coding of the data. The coding process was conducted by the two authors of this study, both of whom are academics specialized in reading and writing skills within the field of Turkish language education. One of the coders also teaches a graduate-level course on qualitative research methods. Themes were derived through a systematic process. After independent coding, the two coders compared their code lists and discussed similarities and differences. Codes that shared conceptual proximity were clustered into categories, and through iterative discussion and negotiated agreement, these categories were further refined into overarching themes. This process ensured that the themes were not imposed a priori but were inductively developed from the data. Consensus was reached through joint meetings in which all disagreements were debated until full agreement was established. After the thematic framework was established, an additional Turkish language education expert, who was not involved in the coding process, served as an external auditor. This expert reviewed the codes, categories, and themes to assess their appropriateness and the extent to which they reflected the data. Inter-coder reliability was not calculated quantitatively; instead, reliability was ensured through iterative discussions and negotiated agreement between the two coders.

To ensure the validity and reliability of the results obtained from this study, the validity strategies of expert review, participant confirmation and detailed explanation, and the coding reliability strategy recommended by Cresswell (2021) were used. With expert review, the reliability and validity of the findings obtained from the data analysis were examined by subjecting them to the evaluation of an independent expert. The experts who participated in the evaluation of the written texts approved the themes and categories obtained from the data analysis. In the findings, detailed explanations about the participants, data collection process

and analysis process are presented. Coding reliability was ensured by involving more than one coder in the data coding process, checking their consistency and discussing, resolving and reconciling any disagreements.

In this section of the study, quantitative and qualitative findings are presented. Table 1 presents the intraclass correlation coefficients (ICCs) for the scoring reliability between the experts and between the experts and ChatGPT, along with the corresponding reliability levels. Table 2 shows the average scores given by the experts and the scores given by ChatGPT. Then, the themes and categories obtained from the feedback of the experts and ChatGPT on the written texts are presented.

RESULTS

Quantitative Findings

Table 1 shows the intraclass correlation coefficients among the experts and between experts and ChatGPT.

Table 1. *Intraclass correlation coefficients for reliability.*

Criteria	Between Experts	Reliability	Between experts and ChatGTP	Reliability
Content	.665	Fair	.578	Fair
Organization	.665	Fair	.248	Weak
Vocabulary	.649	Fair	.568	Fair
Language Use	.214	Weak	.271	Weak
Mechanics	.389	Weak	.346	Weak
Total	.686	Fair	.773	Good

According to Table 1, the intraclass correlations between experts are moderate (r = .665) in content, organization (r = .665) and vocabulary (r = .649), but weak in language use (r = .214) and mechanics (r = .389). In the agreement between experts and ChatGPT, there is a moderate correlation in content (r = .578) and vocabulary (r = .568) criteria, while there is a weak correlation in organization (r = .248), language use (r = .271) and mechanics (r = .346). In terms of total scores, the experts' scores have a moderate level of reliability (r = .686), while there is a good level of reliability between ChatGPT and the experts' scores (r = .773). Table 2 shows the scores given to the texts by the experts and ChatGPT.

Table 2. *Scores given to texts by experts and ChatGPT.*

	Scorer	T-1	T-2	T-3	T-4	T-5	T-6	T-7	T-8	T-9	T-10	T-11	T-12	T-13	T-14
Content	M.E.S.*	24.6	21	23.3	22.6	27.3	24.3	25	25.6	22.6	25.3	22.6	24.3	21	25
	ChatGPT	27	24	21	25	26	20	26	30	24	27	23	26	28	24
Organization	M.E.S.*	17.3	15.3	16	14.3	18	17.6	18	18	17.3	18.6	17.6	18.6	15.6	18.3
	ChatGPT	18	17	15	16	17	20	20	14	14	15	18	14	20	14
Vocabulary	M.E.S.*	18	15	15.3	15	17	17.6	17	17.6	17.3	18	18	17.6	16.6	17.3
	ChatGPT	17	17	16	16	17	16	17	16	16	19	15	16	16	16
Language Use	M.E.S.*	20	19	19.6	20.3	21	19.6	21	20.3	20.3	21.6	20	22.3	19	19.6
	ChatGPT	21	18	21	22	20	20	20	25	18	18	21	18	23	23
Mechanics	M.E.S.*	4.3	3.6	3.3	2.6	4	3.6	4	3.6	3.6	3.6	3.3	4.3	3.6	3.6
	ChatGPT	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Total	M.E.S.*	83.3	74	77.6	75	87.3	83	85	85.3	81.3	87.3	81.6	85.6	76	85
	ChatGPT	90	80	77	86	84	80	87	89	76	83	81	78	91	83

*Means of expert scores

Content: According to the averages of the evaluations made by the experts, one text (T-5) was evaluated as "excellent to very good", 11 texts as "good to average" and two texts as "fair to poor". ChatGPT rated four texts as "excellent to very good", eight texts as "good to average" and two texts as "fair to poor". The number of texts rated with the same quality by experts and ChatGPT is six (T-4, T-7, T-9, T-11, T-12, and T-14). The score difference between experts and ChatGPT ranged from 0.4 to 7 out of 30, with an average of 2.5 points.

Organization: In the organization, experts rated six texts as "excellent to very good" organized and eight texts as "good to average" organized. ChatGPT rated nine texts as "good to average" and five as "excellent to very good". The number of texts rated with the same quality by experts and ChatGPT is five (T-2, T-3, T-4, T-7, and T-9). The difference ranged from 0.4 to 4.4 out of 20, averaging 2.3 points.

In the vocabulary criterion, experts rated 11 texts as "good to average" and three texts as "excellent to very good". ChatGPT rated one text as "excellent to very good" and 13 texts as "good to average". The number of texts rated with the same quality by experts and ChatGPT is twelve (T-2, T-3, T-4, T-5, T-6, T-7, T-8, T-9, T-10, T-12, T-13, and T-14). Vocabulary showed the smallest difference, ranging from 0 to 3 out of 20, with a 1.2-point average.

Language Use: The experts rated 13 texts as "good to average" and one text as "excellent to very good", ChatGPT rated 10 texts as "good to average" and four texts as "excellent to very good". The number of texts rated with the same quality by experts and ChatGPT is nine (T-1, T-2, T-3, T-5, T-6, T-7, T-9, T-10, T-11). The range was 0.4 to 4.7 out of 25.

Mechanics: According to the mean scores of the experts and ChatGPT, no text in the mechanics sub-sample was categorized as "excellent to very good". Experts rated four texts as "good to average", nine texts as "fair to poor", and one text as very poor. ChatGPT assigned all texts to the "good to average" category. The range was 1.4 out of 5, with an average of 0.4.

Total: There are eight texts with a difference of 0-5 points between the average total scores of ChatGPT and experts. The score difference in the remaining six texts ranges between 5.3 and 15 points. Total score differences ranged from 0.6 to 15 points, with an average of 5.1.". The number of texts rated with the same quality by experts and ChatGPT is four (T-1, T-5, T-7, T-12).

Qualitative Findings

Figure 2 shows the themes, categories and subcategories obtained from the feedback given to the written texts by the experts and ChatGPT. In the study, three themes, five sub-themes and ten categories were derived because of the analysis of the written feedback data.

The data were analyzed under three main themes: "*strengths*", "*weaknesses*", and "*commonalities*". The themes of strengths and weaknesses included two sub-themes each (e.g., "*strengths of experts*", "*strengths of artificial intelligence*"; "*weaknesses of experts*", "*weaknesses of artificial intelligence*", which were further divided into categories and sub-categories. The theme of commonalities included a single sub-theme "commonalities between experts and artificial intelligence". The relevant categories and direct quotations are as follows.

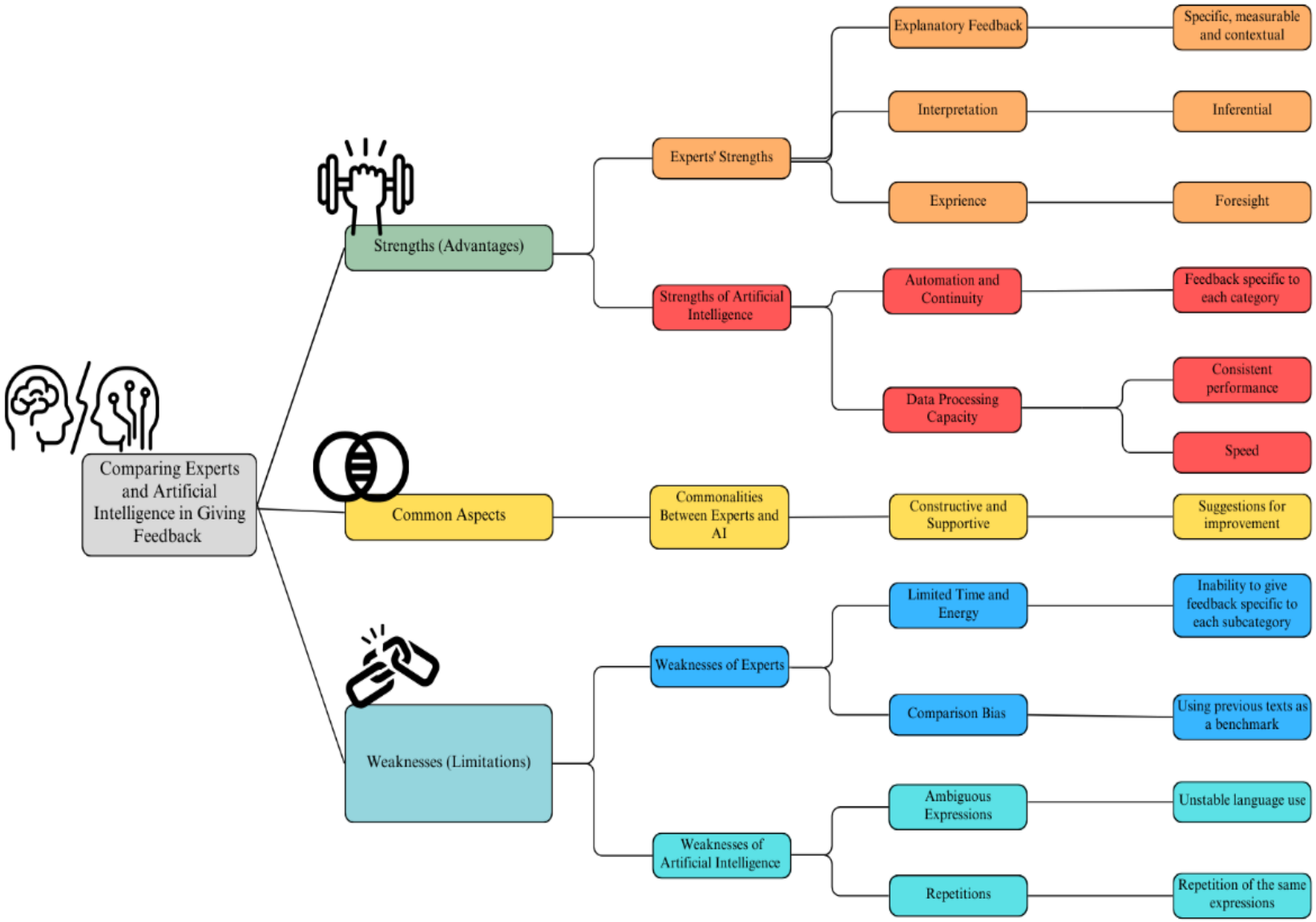

Figure 2. *Comparison of artificial intelligence and experts in feedback.*

Strengths of experts

The relevant categories and direct quotations are as follows.

Explanatory Feedback: The advantages of experts in the feedback process and their superiority over AI were in giving descriptive, specific and measurable feedback. For example, while Expert 1 gave feedback for the same text with more specific and direct statements such as "There is an error in the spelling of the conjunction "de [also], tabii-tabi [natural-subject to]" and the word "hikaye [story]", there is an error in the use of the conjunction 've [and]' and the semicolon, the use of comma after the adverbial verb is incorrect, but there are not seem to be a spelling error", ChatGPT used general expression patterns such as: "In general, correct spelling, punctuation and paragraphing are used in the text, but there are some minor errors and ambiguities in meaning". Again, in the language use category of a different text, Expert 2 provided the following direct quotation: "Thus, he can even come up with an invention based on something he has read' – expression disorder due to the use of the word in the wrong place", clearly indicating the location of the error. In contrast, ChatGPT commented more generally: "While the text has original and developed sentence construction skills, there are some simple expression disorders."

Interpretation: Another strength of the experts was identified as making associations between the opinions presented in the text, comments and predictions about the content of the text. For example, Expert 2 gave synthesis level feedback pointing out the incompatibility of the title and the text by saying "The benefits of reading are discussed, but the title does not cover the whole text because it refers to the world of imagination." ChatGPT gave feedback as follows: "Topics such as the importance of reading, the development of imagination and thinking skills are discussed in a clear and detailed way. Also, the importance of teaching children the habit of reading books is emphasized. In general, the text contains a strong argument and supporting details.". Thus, it failed to draw attention to text-title coherence.

Experience: Another strength of the experts is their predictions based on their experiences regarding written expression and reading comprehension skills. Expert 3's statement "a reader who only considers the text will perceive the importance of reading only as an act of psychological contribution." is a feedback based on the interpretation of how the text will be perceived by the reader and how it will affect the reader's comprehension. ChatGPT, on the other hand, made an explanation for this text as "The text deals with the effect of reading on human psychology and provides sufficient information on the subject". In addition, another advantage of experts is that they can give feedback on metaphors and implicit meaning. For example, Expert 2 mentioned the use of metaphors by saying "One of the first points that stands out in the text is the metaphorical use of words and the concretizing quality that this adds to the narrative". In order to do this, a comprehension skill that goes beyond the literal meaning of words is required. Again, Expert 1 mentioned that there is prejudice in the text by saying "Some thoughts are approached with prejudice without being fully expressed and detailed." ChatGPT, on the other hand, did not mention any bias in the information and ideas presented in the texts.

Strengths of artificial intelligence

Automation and continuity: One of the most superior aspects of ChatGPT was that it provided individual explanations on almost all aspects of the sub-dimensions of written expression in each feedback. For example, for Text 1, ChatGPT gave 542 words of feedback for all 5 subdimensions, while Experts 1, 2, and 3 provided 75, 53, and 58 words in 4, 5, and 3 categories, respectively. Experts and ChatGPT gave the same score to the "mechanics" sub-dimension of Text 10, but only ChatGPT gave detailed explanations while giving feedback: "The mechanical classification of the text seems closer to the category 4 (Good) and deserves a score of 4. There are some punctuation errors, some irregularities in sentence structure and repetitions in the text. However, the text is generally clearly written and the ideas are clearly expressed. The language

used in the text is very vivid and impressive and emphasizes the importance of reading and learning. The organized paragraph structure of the text helped to express different ideas separately". In the same sub-dimension, Expert 1 only gave a score of "4", Expert 2 stated the words with spelling mistakes as "tabi [subject to], hakim [judge], imkan [possibility]" and Expert 3 wrote an explanation as "there are a few spelling and punctuation mistakes".

Data processing capacity: As the number of texts increased, the number of explanations of experts decreased, they got tired and skipped some categories without giving feedback. ChatGPT, on the other hand, continued to provide feedback continuously and quickly with its automation feature and data processing capacity. The inability of the experts to provide feedback specific to each subcategory is explained by the limited time and energy capacity of human beings.

Weaknesses of experts

Limited Time and Energy: As the number of texts increased and the time allocated for the evaluation became longer, the experts became tired and distracted. This tendency was evidenced by the fact that, while the initial texts contained more detailed written feedback, the later texts included shorter and less elaborate comments, indicating a gradual decrease in feedback length across the evaluation process. Consequently, although each text was supposed to be evaluated across five subcategories, the experts often provided feedback on only three or four subcategories and shortened their explanations as the process progressed.

Comparison Bias: Another weakness of the experts was their tendency to use previous texts as a criterion as the texts progressed. For example: "It is definitely a more successful text in terms of content compared to other texts. It has relatively more details" (Expert-1), "It is a text with relatively more developed sentence structures compared to the sentences in other texts" (Expert-2), "When we look at the general vocabulary and volume of the text, it is noteworthy that both the volume and the preferred words have different meanings and are used correctly compared to other texts" (Expert-1). These explanations showed that the experts made the mistake of taking the previous text or texts as a criterion, deviating from the need to evaluate each text on its own in accordance with the scoring criteria. ChatGPT, on the other hand, evaluated each text independently and performed its scoring according to the criteria.

Weaknesses of artificial intelligence

Ambiguous Statements: One of the weaknesses of ChatGPT is the language and ambiguous expressions it uses in some of its evaluations. While the experts made a definite judgment when there was an error and expressed their feedback in a precise language as "The number of spelling mistakes is high and the spelling rules are not followed" (Expert 3), ChatGPT did not make a definite judgment on whether there was an error or not by saying "In the text... there may be errors in word type and usage in some places".

Repetition: One of the limitations of the ChatGPT is that it often repeats itself and in some of its feedback, it quotes from the written text and writes the same sentences over and over again. For example, in a text in which it evaluated the content dimension, ChatGPT said: "The text emphasizes that theater is a wide-ranging field and that students are primarily taught about theater. The text states that the text also explains other different play branches of theater such as comedy, tragedy, drama, meddah, theatrical village play, theatre-in-the-round.". Likewise, ChatGPT, while evaluating the vocabulary dimension in the same text, repeats again: "There is no ambiguity in the text. In addition, it is stated that students should do research to learn about theater and other branches of theater." Moreover, ChatGPT repeated some sentence patterns quite often. To support this observation, a frequency analysis of the AI-generated feedback indicated that several sentence patterns were repeatedly used across different texts. For instance, the phrase "There is almost no expression disorder in the text" appeared in 11 out of 15 texts,

while similar templates such as "There may be errors in word type and usage …" or "There is no ambiguity in the text" also recurred multiple times (see Appendix 1).

Commonalities between experts and ChatGPT

Experts and ChatGPT have strengths and weaknesses compared to each other, but they also have commonalities. Constructive and Supportive Tone. Both experts and ChatGPT tended to frame their feedback in a constructive and supportive manner, aiming to motivate the writers. For instance, Expert 1 commented, "A brief but concise opinion piece was created by making associations about the functions of reading. I consider it a very good text because it fully reflects the thesis", while ChatGPT similarly highlighted, "The thesis presented in the text is strongly developed from the beginning. Paragraphs and sentences are presented in an organized flow that supports the thesis."

Content-Oriented Feedback. Both also provided suggestions for improvement that addressed the substance of the text. For example, ChatGPT noted that, "Although the text includes statements about the benefits of reading books, stronger support could have been provided by using examples and evidence," reflecting a shared tendency to encourage more elaboration and justification.

Attention to Structure and Sentence Complexity. Another commonality was the focus on sentence structure. ChatGPT suggested that, "The text could get a higher score if some improvements could be made, using a richer vocabulary and more sophisticated sentence structures," while Expert 2 remarked that the text was, "composed of simple sentences, but there is no expression disorder." Both thus attended to the simplicity or complexity of sentences as part of their evaluative lens.

These overlaps may partly stem from the design of the feedback prompt, which directed both experts and ChatGPT to evaluate texts according to predefined subcategories. They may also be influenced by the nature of the writing task itself, which emphasized thesis development, content elaboration, and linguistic accuracy. At the same time, the similarities arguably reflect shared pedagogical principles of academic writing, such as clarity, coherence, and the encouragement of idea development. Nevertheless, it is important to recognize that such commonalities might not necessarily indicate a deep alignment in evaluative criteria. In some cases, they may represent surface-level similarities in phrasing or tone rather than identical underlying judgments.

DISCUSSION and CONCLUSION

In this study, three experts and ChatGPT evaluated the written texts of undergraduate Turkish language teaching students according to the criteria of content, organization, vocabulary, language use and mechanics. Intraclass correlation coefficient was used to examine the reliability of the raters both between the experts and between ChatGPT and the experts. A qualitative analysis of the responses to the written texts by the experts and ChatGPT was also conducted to reveal the strengths and weaknesses of the experts and ChatGPT.

Discussion of Quantitative Findings

There was a weak correlation between the expert scores in the language use and mechanics subscales, but a moderate correlation in content, organization, vocabulary and total score. ChatGPT and expert scores are weak in the areas of organization, language use and mechanics. The scores were only moderately similar in the content and vocabulary subscales. In terms of total scores, there was a good reliability between ChatGPT and expert raters. It is worth noting that the sub-dimensions with the lowest reliability among experts (language use and mechanics) also have low reliability with ChatGPT, while most of the other dimensions with moderate reliability have moderate correlation with ChatGPT. Overall, ChatGPT and experts have acceptable reliability in total scores.

Across the five sub-dimensions and 14 texts, yielding a total of 70 evaluations (5 sub-dimensions × 14 texts), ChatGPT placed the texts in a higher category compared to the human raters on 23 occasions, whereas human raters did so 11 times. In 36 instances, both ChatGPT and the human raters assigned the texts to the same category, indicating partial convergence in their evaluations. In total scores, ChatGPT gave higher scores than experts in 6 texts and lower scores in 8.

While open-ended questions offer many advantages, inter-rater reliability and subjective judgments of raters are two major disadvantages. Although the use of rubrics can reduce subjective judgments, it does not guarantee complete objectivity (McMillan, 2018). The aim of measurement and evaluation processes in open-ended questions is to achieve a high degree of consistency and reliability. This study, as in previous studies (Özen, 2020; Yavuz & Bilgeç, 2016), reveals that moderate to weak consistency among human raters is common. The main reason for inconsistencies between assessors in open-ended questions is insufficient equipment in the relevant field (Baidoo-Anu & Owusu Ansah, 2023). Nevertheless, it is important to include open-ended questions in educational courses. Kanık Uysal *et al.* (2022) advocate the implementation of open-ended questions, especially in language teaching courses.

The use of ChatGPT for scoring written texts can offer significant opportunities, as the sufficient reliability between its scores and those of experts can provide supportive feedback to students and reduce the time teachers spend on time-consuming assessment processes. However, studies have shown that technology-mediated ratings show low (significant or insignificant) correlation with human raters' ratings. For example, the correlation between AES systems' evaluations of written texts and human raters' evaluations of written texts has been found to be insignificant or significant but not high (Huang, 2014). Similarly, assessments of ChatGPT have been found to show low correlation with assessments of human raters (Bui & Barrot, 2025). For now, it seems difficult to expect a high correlation between ChatGPT and human raters in scoring open-ended written texts.

Additionally, the study revealed that ChatGPT and the experts exhibited different levels of reliability in different dimensions in the assessment of writing skills, which is in line with the findings of previous studies. Jackaria *et al*. (2024) found that experts had poor reliability in the language dimension, moderate reliability in the content, organization and mechanics dimensions, and good reliability in the total score. Moreover, there was good reliability between ChatGPT and the experts on the organization dimension and moderate reliability on the other dimensions. In addition, there was no significant difference between the scores of ChatGPT and experts. Jackaria *et al*. (2024) reported that ChatGPT ratings still need to be checked by human raters. However, poor reliability in the mechanics sub-dimension in both our study and previous studies is unexpected. This is because automatic assessment tools have been used for years to assess the mechanical aspect of writing (Huang, 2014; Wang *et al*., 2024). Therefore, higher reliability with experts could have been expected in this sub-dimension, which is expected to have already shown progress, and which is in line with the nature of ChatGPT as it is a more technical area compared to content and editing. It is important to reveal the possible reasons for this situation in future studies.

In writing skills, it is usual to find varying reliability between ChatGPT and experts in different sub-dimensions of writing because ChatGPT exhibits varying performances in areas such as accuracy, depth, and relevance while producing texts (Busch & Hausvik, 2023). This variability in different areas of writing is also seen in the writing evaluation process. ChatGPT exhibits skills that vary from domain to domain (Busch & Hausvik, 2023) and produces texts that are sometimes comparable to humans in writing skills and sometimes weaker than humans (Klyshbekova & Abbott, 2024; Waltzer *et al*., 2023; Yan, 2023). From this point of view, it is normal to have varying reliabilities in the sub-dimensions of writing.

The distinction between the leniency or strictness of ChatGPT and human raters remains unresolved. Jackaria *et al*. (2024) discovered that ChatGPT assigned more permissive scores. Uyar and Büyükahıska (2025) similarly discovered that human evaluators regularly assigned higher scores, with the difference being statistically significant. Analogous to ChatGPT, prior research has demonstrated that technologies such as AES yield superior evaluations compared to human assessors (Huang, 2014). Nonetheless, Parker *et al.* (2023) discovered that ChatGPT is more stringent. In our study, ChatGPT typically assigned higher ratings, although it was occasionally more stringent. Consequently, it is challenging to ascertain whether ChatGPT is more stringent or permissive than human evaluators; yet it can be asserted that it typically assigns more lenient scores.

Discussion of Qualitative Findings

In the study, experts and ChatGPT were asked to justify the participants' ratings of the texts. In this feedback, differences in strengths and weaknesses between experts and ChatGPT were identified. Compared to ChatGPT, experts provided more descriptive, specific and quantifiable feedback. Experts provided important evaluations such as the incorrect use of a conjunction, misspelling or misuse of a word, the number of spelling mistakes in the text, and good aspects of the text. This kind of feedback occurred a few times in ChatGPT but not very often. ChatGPT mostly used general statements in its feedback and did not give much specific feedback. This is a flaw and a limitation of the AI chatbot. Non-specific feedback can benefit high achievers, but low achievers need more specific feedback (McMillan, 2018). As a result, the rationales provided in ChatGPT's assessment may be ineffective for low-achieving students. However, it is worth noting that experts do not always provide specific and detailed feedback. Experts provided more specific feedback, but this was not consistent.

Another dimension in which the experts outperformed ChatGPT was their ability to make metacognition-based comments by associating the views presented in the text. Such associations were not found in ChatGPT. In conclusion, the experts' metacognition skills and their ability to associate text elements distinguish them from ChatGPT. ChatGPT lacks metacognition, cannot think critically or creatively, and cannot reason (Klyshbekova & Abbott, 2024; Kortemeyer, 2023; Mondal & Mondal, 2023). AI models like ChatGPT cannot understand anything (Mahowald, 2023) but can calculate probabilities using mathematical calculus. This weakness highlights the most important difference between language processing models and humans.

Another strength of experts is their practical knowledge of written expression and reading comprehension, as well as their understanding of the implicit meanings of text. Since ChatGPT cannot think like a human, it cannot provide feedback on such issues. Teachers' experiences can also help explain the findings of our study. However, ChatGPT lacks this feature. As a result, as Wang *et al.* (2024) point out, while experts provide feedback based on their experience, ChatGPT can only provide feedback based on the text presented to it.

Experts have a linguistic advantage over ChatGPT. Compared with human experts, AI systems such as ChatGPT may face certain linguistic limitations. In some cases, its assessments included ambiguous or hedged statements, which could make the feedback less clear for recipients and may therefore be regarded as a potential weakness. This observation appears consistent with research on large language models (LLMs), which indicates that they generate probabilistic rather than categorical outputs and may encounter difficulties in reliably expressing or calibrating their confidence (Kadavath *et al*., 2022; Steyvers *et al*., 2025). Similar concerns have been noted in the field of AI explainability, where limited transparency can occasionally result in vague or less actionable explanations, thereby constraining the usefulness of model outputs (Doshi-Velez & Kim, 2017; Zhao *et al*., 2024).

In educational contexts, studies comparing AI-generated and human feedback suggest a mixed picture. On the one hand, LLM feedback may provide encouragement and supportive

comments; on the other hand, it is at times reported to be less precise or less actionable than expert feedback (Shi & Aryadoust, 2024; Steiss *et al*., 2024). For instance, Shi and Aryadoust's (2024) systematic review of automated written feedback (AWF) systems identified positive, negative, and mixed findings across studies, pointing to variable performance depending on the educational setting. Likewise, Steiss *et al*. (2024) observed that human raters tended to produce higher-quality feedback than ChatGPT in most dimensions, such as clarity, accuracy, and supportive language. Taken together, these insights suggest that the tendency of ChatGPT to use hedged or ambiguous language should be acknowledged as a possible limitation when considering its reliability in educational feedback practices.

One notable weakness of AI compared to human experts is its tendency to be repetitive. In some cases, it directly quoted from the student's text and repeated the same sentences multiple times. Such recycling of templates is consistent with what has been described in the literature as template-based generation in large language models, where outputs rely on pre-trained linguistic patterns rather than entirely novel formulations (Ji *et al*., 2023). This repetition may reduce the impact of the feedback, potentially causing desensitization in students. Despite such limitations, AI also offers significant advantages, including automation, consistency, and high data-processing capacity. One of the most important strengths of AI compared to experts is its ability to address nearly all aspects of sub-dimensions in written texts individually. For instance, in the "organization" sub-dimension—covering elements such as fluency, clarity of ideas, logical sequencing, and coherence—AI provides specific feedback on each component. In contrast, experts often assigned scores without elaborating on the reasoning, and when explanations were given, they were generally less detailed than those generated by ChatGPT. This difference may stem from human limitations, such as fatigue or reduced focus over time. AI, by contrast, operates continuously and consistently, unaffected by such constraints. While some earlier studies reported that automated systems offer less and lower-quality feedback than human raters (Dikli & Bleyle, 2014), recent findings suggest that AI can outperform humans in terms of feedback quantity (Wang *et al*., 2024). Although experts can offer more comprehensive feedback in both quality and depth, factors like time pressure and mental fatigue may hinder this potential in practice. When artificial intelligence and human experts are compared, it is evident that both have strengths and weaknesses, yet they also share certain similarities. Both offer constructive and supportive feedback, providing authors with suggestions to enhance their written expression. According to the literature, ChatGPT can perform at a level comparable to humans in certain contexts. For instance, ChatGPT can produce more complex sentences than humans (Herbold *et al*., 2023), make similar judgments (Mahowald, 2023), and process language in a human-like manner (Cai *et al*., 2024). Therefore, while AI may not rival experts in every domain, it can achieve comparable success in specific areas. In conclusion, AI applications are expected to bring significant changes to education. Despite ongoing concerns, their practicality for both teaching and learning is likely to increase interest in their use. Research on the potential of ChatGPT for reading assignments, exams, and providing feedback—tasks that typically demand a substantial amount of teachers' time (Bryant *et al*., 2020; Parker *et al*., 2023) — has produced mixed results.

Although AI tools approach quantitative reliability, the quality of their feedback remains lower (Wang *et al*., 2024). This is partly because AI can fabricate justifications (Guo *et al*., 2023) and is unable to assess the accuracy of its own responses (Han *et al*., 2024). If ChatGPT's feedback goes beyond scoring to include interpretation of messages, implicit meanings, or metacognitive elements, even greater caution is required. The findings of this study are consistent with previous research (Fuchs, 2023; Klyshbekova & Abbott, 2024; Kortemeyer, 2023; Tarakçı & Tarakçı, 2024), indicating that ChatGPT is still far from fully replicating the cognitive capabilities of the human brain, particularly in terms of language comprehension. This limitation is especially pronounced in understanding figurative meanings.

ChatGPT can be used effectively for scoring tasks of relatively low importance (such as homework) but should not be used in situations involving critical academic outcomes without expert supervision. Such systems are highly effective when used as a helpful tool for teachers (Dikli & Bleyle, 2014). It is also best to avoid using it for advanced writing tasks that currently require metacognition and metaphorical components.

To summarize, we reiterate the view (Mondal and Mondal, 2023) that AI can never replace humans, and even if it reaches the capacity to replace humans in the future, it should not do so but instead serve as an assistant to humans. This is because a good teacher has qualities that an AI cannot replicate, such as inspiring students, fostering connections and belonging, and seeing the world through students' eyes (Bryant *et al.*, 2020). According to Bryant *et al*. (2020, p. 2), some educational tasks "cannot and should not be automated". While there are situations where teachers can use AI for feedback, it is crucial to emphasize the value of human-centered education.

This study yields several recommendations for researchers and educators. We advise researchers to examine the grading consistency among experts and other AI programs across diverse kinds of writing. Moreover, further investigation is required on the formulation of prompts to enhance the reliability of AI in assessing writing proficiency. AI tools like ChatGPT can deliver feedback and evaluations that match the level of the prompts provided. Furthermore, we advise investigating subjects such as the distinctions between trained and untrained ChatGPT, as well as the impact of employing sample texts and scores during training. From a practical perspective, educators are encouraged to integrate AI technologies such as ChatGPT into the writing assessment process as supplementary tools rather than replacements for human evaluation. These tools can be effectively used to provide preliminary feedback, identify mechanical and surface-level errors, and support students' revision practices. However, teachers should maintain their pedagogical role by interpreting AI-generated feedback, addressing contextual nuances, and guiding students toward deeper writing improvement. For researchers, future studies could focus on developing hybrid feedback models that combine the efficiency of AI with the interpretive depth of human expertise, examining their impact on writing development and feedback literacy.

LIMITATIONS

This research possesses some limitations. Therefore, the interpretation must consider these limits. First, the study assessed solely the written compositions of 14 students. Additionally, a single writing was obtained from each student.

Another limitation of the study concerns the observation that the experts' feedback tended to become shorter as the evaluation process progressed. This observation was based on the researcher's review of the written feedback rather than on quantitative measures (e.g., average word count) or debriefing with the coders. Therefore, while the trend was evident in the data, it should be interpreted cautiously, as no additional triangulation was conducted to confirm this finding)

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Bayburt University, 4.11.2024-238376.

Data Availability Statement

Research data are available from the first author upon reasonable request.

Informed Consent

The study used students' previously submitted coursework as data. No additional tasks or writings were collected specifically for the purposes of this research. All data were anonymized prior to analysis, and

no personally identifiable information was used. Since the data consisted of routine coursework produced as part of normal classroom activities, and no direct intervention or additional participation was required, written informed consent was not applicable.

Contribution of Authors

Bahadır Gülden: Designing the study, and Collecting and analyzing the data. **Huzeyfe Bilge:** Analyzing the data, and Writing the manuscript. **Pınar Kanık Uysal:** Analyzing the data, and Reliability procedures.

Orcid

Bahadır Gülden https://orcid.org/0000-0003-1917-8813
Huzeyfe Bilge https://orcid.org/0000-0001-7664-488X
Pınar Kanık Uysal https://orcid.org/0000-0003-1208-9535

REFERENCES

Akaya, A.O., & Kurtuluş, A. (2011). 6. sınıf matematik dersi öğretim programının uygulanabilirliğine ilişkin öğretmen görüşleri [Teachers' opinions about the applicability of 6th grade mathematics curriculum]. *Education Sciences*, *6*(3), 2229-2245.

Altundal, B. (2024). A review of artificial intelligence and its use in education. *Artificial Intelligence in Educational Research*, *1*(1), 28-38. https://doi.org/10.5281/zenodo.11241932

Applebee, A.N., & Langer, J.A. (2011). EJ extra: A snapshot of writing instruction in middle schools and high schools. *English Journal*, *100*(6), 14-27. https://doi.org/10.58680/ej201116413

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7(*1), 52-62. https://doi.org/10.61969/jai.1337500

Bilge, H. (2024). Comparison of the texts selected from a Turkish textbook and the texts produced by artificial intelligence chatbots in terms of vocabulary. *Artificial Intelligence in Educational Research*, *1*(1), 1-16. https://doi.org/10.5281/zenodo.11246999

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

Bryant, J., Heitz, C., Sanghvi, S., & Wagle, D. (2020). *How artificial intelligence will impact K-12 teachers*. McKinsey & Company.

Bui, N.M., & Barrot, J.S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, *30*(2), 2041 2058. https://doi.org/10.1007/s10639-024-12891-w

Busch, P.A., & Hausvik, G.I. (2023). Too good to be true? An empirical study of ChatGPT capabilities for academic writing and implications for academic misconduct. In *Proceedings of the 29th Annual Americas Conference on Information Systems* (Paper 1829). Association for Information Systems. https://aisel.aisnet.org/amcis2023/sig_odis/sig_odis/21

Cai, Z., Duan, X., Haslett, D., Wang, S., & Pickering, M. (2024). Do large language models resemble humans in language use?. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 37-56). Association for Computational Linguistics. https://aclanthology.org/2024.cmcl-1.4.pdf

Creswell, J.W. (2022). *A concise introduction to mixed methods research*. SAGE publications.

de Winter, J.C.F. (2023). Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education*, *34*, 915-930. https://doi.org/10.1007/s40593-023-00372-z

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, *22*, 1-17. https://doi.org/10.1016/J.ASW.2014.03.006

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

Flower, L., & Hayes, J.R. (1981). A cognitive process theory of writing. *College Composition & Communication*, *32*(4), 365-387.

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*, Article 1166682. https://doi.org/10.3389/feduc.2023.1166682

Graham, S., Hebert, M., & Harris, K.R. (2015). Formative assessment and writing. *The Elementary School Journal*, *115*(4), 523-547. https://doi.org/10.1086/681947

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, *29*(7), 8435-8463. https://doi.org/10.1007/s10639-023-12146-0

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection*. arXiv preprint arXiv:2301.07597. https://doi.org/10.48550/arXiv.2301.07597

Han, Z., Battaglia, F., Udaiyar, A., Fooks, A., & Terlecky, S.R. (2024). An explorative assessment of ChatGPT as an aid in medical education: Use it with caution. *Medical Teacher*, *46*(5), 657-664. https://doi.org/10.1080/0142159X.2023.2271159

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). *AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays*. arXiv preprint arXiv:2304.14276. https://doi.org/10.48550/arXiv.2304.14276

Huang, S.J. (2014). Automated versus human scoring: A case study in an EFL context. *Electronic Journal of Foreign Language Teaching*, *11*, 149-164.

Jackaria, P.M., Hajan, B.H., & Mastul, A.R.H. (2024). A comparative analysis of the rating of college students' essays by ChatGPT versus human raters. *International Journal of Learning, Teaching and Educational Research*, *23*(2), 478-492. https://doi.org/10.26803/ijlter.23.2.23

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38. https://doi.org/10.1145/3571730

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield, D., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S. *et al.*, (2022). Language models (mostly) know what they know. Cornell University. arXiv preprint arXiv:2207.05221. https://doi.org/10.48550/arXiv.2207.05221

Kanık Uysal, P., Akın Arıkan, Ç., Acar Erdol, T., Bayrak Özmutlu, E., & Akyol, H. (2022). Examining Turkish course exam questions in terms of originality, page layout, item type, item writing criteria and cognitive level. *Education and Science*, *47*(210), 259-280. https://doi.org/10.15390/EB.2022.10896

Kiryakova, G., & Angelova, N. (2023). ChatGPT-A challenging tool for the university professors in their teaching practice. *Education Sciences*, *13*(10), Article 1056. https://doi.org/10.3390/educsci13101056

Klyshbekova, M., & Abbott, P. (2024). ChatGPT and assessment in higher education: A magic wand or a disruptor? *Electronic Journal of E-Learning*, *22*(2), 30-45. https://doi.org/10.34190/ejel.21.5.3114

Koo, T.K., & Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course. *Physical Review Physics Education Research*, *19*(1). https://doi.org/10.1103/PhysRevPhysEducRes.19.010132

Lo, C.K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*, Article 410. https://doi.org/10.3390/educsci13040410

Mahowald, K. (2023). *A discerning several thousand judgments: GPT-3 rates the article + adjective + numeral + noun construction*. arXiv preprint arXiv:2301.12564. https://doi.org/10.48550/arXiv.2301.12564

McMillan, J.H. (2018). *Classroom assessment principles and practice that enhance student learning and motivation improving results*. Pearson.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.

Mondal, H., & Mondal, S. (2023). ChatGPT in academic writing: Maximizing its benefits and minimizing the risks. *Indian Journal of Ophthalmology*, *71*(12), 3600-3606. https://doi.org/10.4103/IJO.IJO_718_23

Oğuz, B. (2024). Readability levels of narrative texts produced by artificial intelligence (ChatGPT) for 6th grade students. *Artificial Intelligence in Educational Research*, *1*(1), 17-27. https://doi.org/10.5281/zenodo.11247479

OpenAI. (2023). *GPT-4 Technical Report.* arXiv preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Özen, O. (2020). *Examining the open-ended question preparation skills of Turkish teachers* [Unpublished master's thesis] Atatürk University.

Parker, J.L., Becker, K., & Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education*, *62*(12), 721-727. https://doi.org/10.3928/01484834-20231006-02

Picken, J. D. (1988). A reassessment of error-count. *JALT Journal*, *10*(1&2), 79-90.

Qadir, J. (2023). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. *In Proceedings of the IEEE Global Engineering Education Conference* (pp. 1-9). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/EDUCON54358.2023.10125121

Rahman, Md. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, *13*, Article 5783. https://doi.org/10.3390/app13095783

Safdar, M., Siddique, N., Gulzar, A., Yasin, H., & Khan, A. (2024). Does ChatGPT generate fake results? Challenges in retrieving content through ChatGPT. *Digital Library Perspectives*, *40*(4), 668-680. https://doi.org/10.1108/DLP-01-2024-0006

Shi, H., Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, *36*(2), 187-209. https://doi.org/10.1017/S0958344023000265

Sok, S., & Heng, K. (2023). Opportunities, challenges, and strategies for using ChatGPT in higher education: A literature review. *Journal of Digital Educational Technology*, *4*(1), Article ep2401. https://doi.org/10.30935/jdet/14027

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*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Steyvers, M., Smyth, P., & Griffiths, T.L. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, *7*, 221-231. https://doi.org/10.1038/s42256-024-00976-7

Tarakçı, R., & Tarakçı, Ç. (2024). The comparison of artificial intelligence generated texts with textbook in teaching idioms and proverb. *Artificial Intelligence in Educational Science*, *1*(1), 39-53. https://doi.org/10.5281/zenodo.11242790

Uyar, A.C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. *International Journal of Assessment Tools in Education*, *12*(1), 20-32. https://doi.org/10.21449/ijate.1517994

Ülper, H. (2019). Yazılı metinleri ölçme ve değerlendirme [Measuring and evaluating written texts]. In N. Bayat (Eds.), *Yazma ve eğitimi* [Writing and teaching writing] (pp. 159-174). Anı Publishing.

Vargas-Murillo, A.R., Pari-Bedoya, I.N.M. de la A., & Guevara-Soto, F. de J. (2023). Challenges and opportunities of AI-assisted learning: A systematic literature review on the impact of ChatGPT usage in higher education. *International Journal of Learning, Teaching and Educational Research*, *22*(7), 122-135. https://doi.org/10.26803/ijlter.22.7.7

Waltzer, T., Cox, R.L., & Heyman, G.D. (2023). Testing the ability of teachers and students to differentiate between essays generated by ChatGPT and high school students. *Human Behavior and Emerging Technologies*, *2023*, Article 1923981. https://doi.org/10.1155/2023/1923981

Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity*, *51*, Article 101440. https://doi.org/10.1016/j.tsc.2023.101440

Weigle, S.C. (2002). *Assessing writing*. Cambridge University Press.

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, *28*(11), 13943-13967. https://doi.org/10.1007/s10639-023-11742-4

Yavuz, İ., & Bilgeç, İ. (2016). Açık Uçlu Sorularla Yapılan Matematik Sınavlarının Ölçme ve Değerlendirilmesinin İncelenmesi [Examining the measurement and evaluation of mathematics exams with open-ended questions]. *Journal of Research in Education and Teaching*, *5*(3), 183-193.

Zhao, H., Chen, H., Yang, F., Wang, Y., & Wu, X. (2024). Explainability for large language models: A survey. *ACM Computing Surveys*, *15*(2), Article 20. https://doi.org/10.1145/3639372

APPENDIX

Appendix 1. *Most frequently repeated sentence patterns in ChatGPT's feedback*.

Sentence Pattern	Frequency
There is almost no expression disorder in the text	11
There may be errors in word type and usage…	7
The text emphasizes that theater is a wide-ranging	6
The text states that…	6
There is no ambiguity in the text.	5