Sakarya University Journal of Education

ISSN : 2146-7455
Publisher : Sakarya University

Vol. 14, No. 3, 515-538, 2024
DOI: https://doi.org/10.19126/suje.1475474

Research Article

Evaluating ChatGPT in Generating Feedback on Content and Organization Components of EFL Compare and Contrast Essays

Amine Hatun Ataş[1*]
Behice Ceyda Cengiz[2]
Berkan Çelik[3]

[1] Galatasaray University, Distance Education Application and Research Center, İstanbul, Türkiye, aminehatunatas@gmail.com

[2] Zonguldak Bulent Ecevit University, Faculty of Humanities and Social Sciences, Department of Translation and Interpreting, Zonguldak, Türkiye, behiceceydacengiz@gmail.com

[3] Van Yuzuncu Yil University, Başkale Vocational School, Department of Computer Technologies, Van, Türkiye, berkancx@gmail.com

*Corresponding author

Received: 30.04.2024
Accepted: 30.10.2024
Available Online: 07.11.2024

Abstract: ChatGPT, an innovative large language model that has impressed worldwide audiences with its exceptional generative capabilities, is now positioned to significantly transform the field of education. The purpose of this exploratory study is to investigate how accurately ChatGPT generates feedback on the content and organization components of EFL compare and contrast essays and the extent to which the feedback length provided by ChatGPT differs from that of the human teacher.To address these questions, a ChatGPT prompt incorporating evaluation criteria for content and organization components was developed, generating feedback on 10 compare and contrast student essays using the ChatGPT 3.5 version. The ChatGPT feedback and teacher feedback were assessed quantitatively and qualitatively according to the predetermined evaluation criteria. Furthermore, two types of feedback were compared descriptively and by conducting the Wilcoxon Sign Rank Test. The findings revealed that ChatGPT produced highly accurate feedback for both content and organization components, surpassing the teacher in the length of feedback provided. While the accuracy rate of the generated feedback was high, issues such as holistic assessment of the essay, false positives, failure to provide feedback where needed, and discrepancies in the depth of feedback compared to teacher feedback were identified. The results suggest that while ChatGPT shows promise in providing educational feedback, teacher-AI collaboration in giving feedback for EFL compare and contrast essays is important for delivering feedback that optimally benefits learners.

Keywords: Artificial Intelligence, Compare and Contrast Essay, ChatGPT, EFL, Feedback

1. Introduction

Exploring the etymology of the term feedback is advantageous for establishing a precise understanding of the concept. The initial definition of feedback entails the process of redirecting a portion of the output of a machine, system, or process back to its input. This terminology emerged from discussions in the field of electrical engineering and rocket science during the early 20th century (Burke & Pieterick, 2010). Steinmetz (1915) exemplified this concept by illustrating that when a cable is grounded, the current at its end undergoes reversal, flowing back into the cable, thus termed "feeding back" rather than existing from it (as cited in Merriam-Webster, n.d.). Later, in the field of education, Kulhavy (1977) defined feedback simply in his work *Feedback in Written Instruction* as "... any of the numerous procedures that are used to tell a learner if an instructional response is right or wrong" (p. 211). Feedback is a very important component in the field of education, more specifically in English as a Foreign Language (EFL) writing classes with many benefits for second language (L2) learners (Biber et al., 2011; Wilson & Czik, 2016). When L2 learners notice a gap between their current language use (interlanguage) and the target language form upon receiving feedback, they become more aware of their errors. According to the noticing hypothesis, form-focused corrective feedback that enables this noticing facilitates L2 learners' language acquisition processes (Schmidt & Frota, 1986). Other research also shows that L2 feedback that is centered on macrolevel aspects of L2 writing such as content, coherence, and cohesion helps L2 learners improve their writing performance (Bakla, 2020; Elola & Oskoz, 2016). Therefore, it becomes important to provide L2 learners with instructional writing feedback not only in the language component of L2 writings but also in such dimensions as content and organization since writing is a versatile skill and requires dealing with high level concepts such as ideation and style (Zhai & Ma, 2023).

Cite as (APA 7): Ataş, A. H., Cengiz, B. C., & Çelik, B. (2024). Evaluating the ChatGPT in Generating Feedback on Content and Organization Components of EFL Compare and Contrast Essays, Sakarya *University Journal of Education,* 14(3), 515-538. https://doi.org/10.19126/suje.1475474

Literature shows that L2 teachers experience some challenges related to providing effective writing feedback (e.g., Dikli & Bleyle, 2014; Fu et al., 2024). One of the biggest challenges is pertinent to the significant amount of time and effort required to offer feedback to students, particularly when dealing with multiple students across various classes (Steiss et al., 2024). Offering individualized feedback and assistance can be demanding and resource-intensive for teachers, particularly when they lack the time or resources to attend to each student's needs effectively (Baskara, 2023; Jackson et al., 2022). This challenge may even discourage some teachers from delivering effective feedback and result in superficial feedback (Noroozi et al., 2023). These challenges motivate L2 teachers and researchers to search for alternative ways of giving writing feedback (Huang, 2023). This need resulted in the development of computer programs designed to offer feedback on writing, known as Automated Writing Evaluation (AWE) systems, such as Grammarly, Pigai, and similar software (Zainurrahman & Rojab, 2024). However, recently, new versions of Generative Artificial Intelligence (GenAI), such as ChatGPT, started to replace and transform AWE systems due to their affordances. Unlike previous models, ChatGPT does not need to be trained on human datasets specific to a task or genre. Additionally, ChatGPT's 3.5 version is currently affordable and readily available for everyone (Steiss et al., 2024).

Recent advancements in AI technology suggest that ChatGPT holds significant potential in writing pedagogy, offering automated feedback on students' L2 writing and supporting teachers in the feedback process. Collaborating with ChatGPT while giving feedback on students' writing can reduce teachers' workload, help prevent the fatigue that comes from correcting numerous student assignments and produce more efficient feedback (Barrot, 2023a; Teng, 2024). However, for effective collaboration between teachers and ChatGPT, it is crucial to first examine key characteristics of ChatGPT's feedback, with accuracy being one of the most important qualities to assess (Steiss et al., 2024). Uncovering ChatGPT's capability to provide accurate feedback can inform teachers of optimal utilization of ChatGPT in their feedback practices.

Research on the use of ChatGPT as a provider of L2 writing feedback has been related to teacher and student perceptions about ChatGPT feedback (e.g., Bok & Cho, 2023; Xiao & Zhi, 2023), and comparison of ChatGPT and teacher feedback in terms of such features as the type of feedback, the level of supportive tone in the feedback, and the clarity of the directions given in the feedback (e.g., Banihashem et al., 2024; Guo & Wang, 2024; Steiss et al., 2024). However, there is a growing need for scrutinizing the accuracy of ChatGPT feedback in EFL writing context, as also highlighted in Guo and Wang's (2024) study. As GenAI chatbots continue to gain prominence, there arises a need for additional research to investigate the quality and accuracy of the generated response by these GenAI chatbots, recognizing the dynamic nature of these evolving technologies (Chaka, 2023). To fill in this research gap, the present study investigates how accurately ChatGPT generates feedback concerning the content and organization components of compare and contrast essays written in the context of a CEFR B1 level EFL Vocabulary and Composition class in an undergraduate Applied English and Translation program in Türkiye. The compare and contrast essay type was selected as the focus of this study due to its emphasis on critical thinking and organizational skills, which are essential components of academic writing at tertiary level. Additionally, this essay type is commonly taught in English for Academic Purposes (EAP) at this level. In order to assess whether ChatGPT, as an educational technology, provides accurate writing feedback evaluations in the classroom, teacher feedback was used as a benchmark for comparison which is also a common method adopted in previous studies (e.g., Mizumoto et al., 2024; Pfau et al., 2023). The rationale behind this comparison is to view ChatGPT as a feedback tool that can complement and enhance teacher feedback instead of serving as a substitute for teacher feedback (Guo & Wang, 2024). Content and organization were chosen as the main feedback focus in the current study - since the accuracy of ChatGPT feedback on linguistic features of L2 texts has already been evidenced in previous studies (e.g., Mizumoto & Eguchi, 2023; Mizumoto et al., 2024; Pfau et al., 2023). It was also noted in earlier studies that even less technologically advanced AWE systems can deal with surface-level errors related to language easily

whereas the capability of these technological tools related to giving feedback on content and organization are still questioned (Guo & Wang, 2024). As the second focus of the study, the length of teacher and ChatGPT feedback is compared. Earlier research on L2 writing demonstrates that feedback amount can be a factor that affects learners' perceptions and uptake of feedback (Thi & Nikolov, 2021; Zhang & Hyland, 2018). The reason for this comparison, therefore, is to provide the baseline for future studies which can analyze the effect of feedback length as a variable that can affect the utilization of ChatGPT feedback by L2 learners (Guo & Wang, 2024).

1.1. Chat generative pre-trained transformer (ChatGPT)

ChatGPT, abbreviated for Chat Generative Pre-Trained Transformer, is an AI-driven conversational agent created by the American startup OpenAI. ChatGPT offers multifunctional capabilities (OpenAI, 2024a). It taps into a vast knowledge base, drawing from diverse sources to generate human-like text, serving as a valuable resource for language input and practice. Unlike traditional databases, ChatGPT's corpus is structured as statistical patterns and associations (Barrot, 2023a). The latest iteration of ChatGPT, derived from the GPT 3.5 model, demonstrates improved proficiency in comprehending natural language, enhanced efficiency and accuracy in addressing inquiries, and increased adaptability (Rudolph et al., 2023; Su et al., 2023).

As an AI-driven technology designed to simulate human intelligence, ChatGPT exhibits exceptional proficiency in a wide range of writing tasks including choosing a topic, establishing the context, creating an outline, drafting the content, and making revisions, often comparable to the capabilities of humans (Barrot, 2023a). ChatGPT provides a lot of affordances in the context of EFL. To exemplify, it can be used for material and assessment generation (Pack & Maloney, 2023), improving writing skills and motivation (Song & Song, 2023), grammar check (Schmidt-Fajlik, 2023), question generation (U. Lee et al., 2023), and feedback (Su et al., 2023). Using GenAI can aid in certain situations, such as when working on early writing drafts or when lacking access to a well-trained educator, given the simplicity of automatic feedback generation with satisfactory quality using ChatGPT (Steiss et al., 2024). Studies approached feedback from different perspectives, including the comparison of the scoring of ChatGPT and human-generated feedback (Steiss et al., 2024), comparison of ChatGPT and teacher-generated feedback (Guo & Wang, 2024), or AI-enabled evaluation (Lee, 2023).

Despite the invaluable affordances, ChatGPT comes with significant limitations. To name a few, language models like ChatGPT have the capability to produce false or incorrect statements, often exhibiting low accuracy in various contexts. ChatGPT might also produce fabricated information instead of producing an "I don't know" response (Meyer et al., 2023). The complexity and depth of the responses can be restricted by the lack of sophisticated or iterative prompt engineering in output generation. In addition, submitting each prompt separately to ChatGPT to prevent the learning from previous prompts may also limit the breadth of its responses even further (Barrett & Pack, 2023). It is known that there is a tendency of AI models to hallucinate. When the essay topics primarily focus on argumentation and critical reflection rather than factual accuracy, these hallucinations are not a concern. Still, this does not change the fact that AI models may struggle with factual correctness in some cases (Herbold et al., 2023). Regarding the limitations with respect to feedback, ChatGPT might use different evaluation criteria than the teachers, which could lead to feedback that does not fit teachers' needs. When ChatGPT does not have adequate background information about the class and students, this could lead to inappropriate feedback (Guo & Wang, 2024). Furthermore, ChatGPT provides longer feedback on average, and this may increase the chances of the error rate related to inaccurate content or knowledge. Occasionally, ChatGPT offers constructive feedback alongside numerous instances of irrelevant or excessive information. ChatGPT might need iterative prompting to refine the outputs although it has great capacity to respond. While not as severe as other challenges previously reported, it can be a struggle for ChatGPT and human evaluators to give feedback for the

high-scoring essays with respect to crucial features prioritization, yet this can be solved via the improved prompting (Steiss et al., 2024).

Before the appearance of large language models (LLMs), educators and researchers have been using several approaches including automated feedback (Barrot, 2023b; Ranalli, 2018), AWE (Link et al., 2022), automated corrective feedback tools (Shadiev & Feng, 2023), or natural language processing (NLP) tools (Wang et al., 2020). LLMs have the capacity to analyze grammar, cohesion, and style all at once, while also offering feedback (Bonner et al., 2023). As an LLM type, ChatGPT is capable of offering versatile feedback automatically across various genres and contexts as it does not necessitate an independent training set like other AWE applications (Steiss et al., 2024). Teachers can leverage these affordances of ChatGPT to generate feedback on student writing (Guo & Wang, 2024). However, effective writing of prompts is needed to harness the full potential of ChatGPT.

1.2. Prompt engineering

Prompt engineering is becoming an essential aspect of understanding generative AI as it progresses to become deeply integrated in every aspect of our lives (Bozkurt, 2024). Guidelines and frameworks were developed to construct effective prompts. Giray (2023) put forward that prompts should encompass distinct tasks, contextual information crucial for task completion, a defined question to address, and specifications outlining the format for generating the response. Particularly, Giray (2023) structured the prompts with the elements "instruction, context, input data, and output indicator" (p. 2630). Spasić and Jankovic (2023) designed their prompts to incorporate the role of the AI model, which defines the persona it adopts while responding, the instruction that guides the model in producing the desired outputs, and seed-words that direct the AI's generated output through specific keywords or phrases. Guo and Wang (2024) suggested including more contextual information, such as language proficiency of students, into the prompts for more personalized feedback generation. OpenAI also provided some guidelines for best practices for prompt engineering. They suggested providing specific, detailed descriptions of the desired context, outcome, length, format, style, and other relevant aspects (OpenAI, 2024b). Lo (2023, p. 1) provided the CLEAR ("Concise, Logical, Explicit, Adaptive, and Reflective") framework as a standard method for creating prompts. Based on that framework, the prompt should be clear and precise as well as structured and coherent. It should include clear output specifications, allow flexibility and customization, and be refined and enhanced via continuous evaluation.

AI can be optimized as a valuable resource for enhancing productivity in delivering quality feedback when specific prompts are formed clearly and precisely (Carlson et al., 2023). Several studies have formed unique prompt structures for their needs in EFL contexts (Bonner et al., 2023; Carlson et al., 2023; Guo & Wang, 2024; Huang, 2023; Pack & Maloney, 2023; Schmidt-Fajlik, 2023; Steiss et al., 2024; Su et al., 2023). We examined these specific prompts that were constructed to provide feedback for students' writing. The prompts constructed to provide feedback on writing in these studies cover various aspects of feedback and assessment for EFL learners' writing. To be more specific, they include instructions for providing specific, actionable feedback on essays, evaluating paragraphs based on given criteria, correcting grammar and mechanics in sentences, and providing suggestions for improvement in writing quality, grammar, spelling, vocabulary, and organization. These prompts emphasize the importance of providing constructive feedback, using examples, and adhering to specific criteria or rubrics for evaluation. To sum up, the most common points of these prompts were the role, context, type of the writing, tone and simple language of the feedback, and evaluation criteria.

1.3. Feedback related studies in EFL essay writing

When scrutinizing studies conducted within the domain of GenAI and feedback in EFL writing, a multitude of findings regarding the benefits and limitations of AI tools emerge. Guo and Wang's (2024)

study focused on assessing how ChatGPT could aid EFL instructors in providing feedback on students' writing, initially by analyzing ChatGPT's ability to generate feedback for EFL students' argumentative essays. The findings demonstrated that ChatGPT generated a substantially greater volume of feedback compared to teachers. Furthermore, whereas teacher feedback primarily concentrated on content and language-related concerns, ChatGPT allocated its attention more evenly across the three feedback areas: content, organization, and language. In a similar vein, Wang et al. (2024) compared ChatGPT and teacher feedback on argumentative essays in terms of feedback accuracy and examined the factors affecting their evaluation. This study demonstrated that ChatGPT and teacher feedback had unique affordances and limitations. It was shown that ChatGPT had a considerable accuracy rate, demonstrating promising capability to give writing feedback. This capability, however, was influenced by the utilization of discourse markers and arguments' length. The limitations of ChatGPT were related to the fact that it limited its feedback to the linguistic form while giving affective feedback. Teacher feedback, on the other hand, was advantageous in terms of teachers' available contextual knowledge about the students' immediate needs and progress supported by their ability to have empathy with their students. Another similar study was conducted by Banihashem et al. (2024) who made a comparison between the quality of the feedback provided by ChatGPT and teachers on argumentative essays. Their study revealed that while ChatGPT feedback was more focused on giving informative feedback about how to write an essay, teacher feedback was centered on locating the problems in the essay. It was also shown that there was not any significant relationship between the essay quality and feedback quality. Another comparison study was carried out by Steiss et al. (2024) who evaluated the quality of ChatGPT and teacher feedback in terms of certain criteria such as being criteria-based, accurate, indicating ways for improvement. In that study, teachers who received comprehensive training in providing writing feedback were found to be more effective than ChatGPT in delivering feedback across all areas, except when it came to feedback based on specific established criteria.

Another body of research looked into the effect of AI-assisted language learning on the development of Chinese EFL learners' writing skills. These studies demonstrated the efficacy of AI-assisted language learning on the improvement of L2 learners' writing (Liu et al., 2021; Song & Song, 2023; Yan, 2023) and writing motivation (Song & Song, 2023). Likewise, Su et al.'s (2023) study on ChatGPT's role in guiding writing suggested that it could assist the learners with developing argumentative writing's structural, dialogical, and linguistic aspects. It was also shown that ChatGPT had competency to provide personalized feedback, evaluate content and organization, as well as analyzing language, and proofreading texts. However, its effectiveness was found to depend on the quality of questions and criteria provided by users.

Regarding the student perceptions about ChatGPT, Bok and Cho (2023) studied college students' views on using ChatGPT for revising paragraphs in an academic writing course. Students found ChatGPT helpful and reliable for feedback, appreciating its instant responses and flexibility. It effectively corrected errors in vocabulary, grammar, and paragraph structure. However, challenges included the lack of error descriptions, unclear feedback, inconsistency in responses, worries about reduced authorship, and doubts about its learning effectiveness. In another study related with student perspectives, Xiao and Zhi (2023) revealed that while Chinese college-level EFL learners viewed ChatGPT as a valuable tool for offering them instant feedback and individualized learning experiences, they were skeptical about the accuracy of the ChatGPT outputs.

The necessity for the current study originates from the increasing focus on how teachers can collaborate with ChatGPT to enhance the L2 writing feedback process (Guo & Wang, 2024). Working alongside ChatGPT has the potential to improve the quality and efficiency of feedback provided to students, supporting teachers in refining their feedback practices and decreasing their heavy workload (Barrot, 2023a; Teng, 2024). However, for collaboration to work well, it is important to evaluate key

aspects of ChatGPT feedback, particularly its accuracy (Steiss et al., 2024) and length of its responses (Thi & Nikolov, 2021; Zhang & Hyland, 2018). Gaining insight into its capabilities will help teachers integrate ChatGPT more effectively into their feedback practices. This study is also significant in the Turkish context where most studies related with the utilization of ChatGPT for providing L2 writing feedback mostly focused on the teacher and student perceptions (e.g., Punar Özçelik & Yangın Ekşi, 2024; Üstünbaş, 2024), paying little attention to the accuracy of ChatGPT feedback. As essay writing is a common component of EAP courses across Türkiye, this study provides results that may influence the practical use of ChatGPT-generated feedback in terms of its accuracy and length for the teachers delivering these courses.

In line with this, the purpose of this study is to evaluate the accuracy of ChatGPT in generating feedback on content and organization components of EFL compare and contrast essays, and to compare the length of feedback provided by ChatGPT with that of human teachers in the same points. Accordingly, the following two research questions (RQs) have been addressed:

RQ1: How accurately does ChatGPT generate feedback on the content and organization of EFL compare and contrast essays?

RQ2: How does the length of feedback given by ChatGPT and teacher on the content and organization components of compare and contrast essays differ and is there a significant difference between the two?

2. Method

2.1. Research design

This study adopted an exploratory research perspective. Effectively exploring a phenomenon requires adopting two key orientations: flexibility, which involves being adaptable in the search for data, and openness, which entails being receptive to various sources for obtaining that data. The emphasis in exploration always lies on inductively generating new concepts and empirical generalizations. During exploration, both quantitative and qualitative data may be collected. While qualitative data often dominate in exploratory studies, they are supplemented with descriptive statistics whenever possible and appropriate (Stebbins, 2001). Our study aims to explore new ground by investigating how accurately ChatGPT generates feedback on the content and organization components of EFL compare and contrast essays and the extent to which the length of the feedback provided by ChatGPT and that of the human teacher differs in EFL context. As this area is not well understood yet, we are using an exploratory research approach to uncover valuable insights. This study will establish a basis for future research examining the direct application of ChatGPT in the classroom, potentially incorporating the student perspective by analyzing the accuracy and length of the feedback generated by ChatGPT in the specified context. Also, this method allows us to be flexible and open in collecting data, helping us to develop new ideas and generalizations through exploration.

2.2. Context and sample

The compare and contrast essays used in the current study were collected from a Vocabulary and Composition course taught by the second author at the department of Foreign Languages and Cultures at a state university in Türkiye. The medium of instruction was English in that department. The writers of the essays enrolled in the afore-mentioned course were 10 Turkish EFL freshmen with CEFR B1 level in English. The essays were randomly selected from this class with homogenous writing proficiency, with all students experiencing difficulty in essay writing due to challenges with grammar and writing skills. According to the short demographic questionnaire that asked about their age, gender, and participation in preparatory school conducted at the beginning of the semester, seven were female and three were male while their ages ranged from 19 to 21. Before their studies at the department, all of these students attended one-year intensive preparatory program where they

practiced the four language skills, including writing, through 20 hours of weekly instruction and reached a B1 level, as evidenced by the proficiency exam conducted at the end of the program. At the preparatory school, they learnt to write different types of texts such as informal e-mails, blog posts and formal letters of complaint. They also received instruction on how to write an opinion essay. The three-hour Vocabulary and Composition course, which lasted 14 weeks, had two hours of instruction allocated for the writing component of the course and included the teaching of different types of essays, such as cause and effect essay, compare and contrast essay, and argumentative essay respectively, in its syllabus. The decision to focus on compare and contrast essays in the current study, rather than on cause and effect essays, which were taught as the first essay type, was made to give learners additional essay-writing practice. This decision was based on the second author's observation, who also teaches the course, that students struggled with essay writing. Before writing the compare and contrast essay, these students received four weeks of instruction and were expected to write their essays during the final week and submit them to the teacher, which accounted for 20% of their final grade. The instruction covered content-related topics, such as writing an effective hook and thesis statement in the introduction paragraph, using examples and explanations as supporting sentences in similarity and difference paragraphs and organization related topics such as unity, coherence, cohesion and the use of appropriate linking words for this essay type. In-class activities included teacher lectures, as well as group and pair work that engaged learners in analyzing sample essays and completing exercises on connectors, punctuation rules, and grammar topics closely related to compare and contrast essays. The essays they were required to write had to consist of four paragraphs, including introduction, similarity, difference and conclusion paragraph. Although this essay type can include four or five paragraphs, the decision to use four was based on the way it was taught in class, due to the difficulty students had with essay writing. To prevent struggles associated with writing essays on unfamiliar subjects and provide them with greater flexibility and comfort, 10 different compare and contrast essay topics were offered in order to allow them to choose one they felt comfortable with.

2.3. Data collection/generation tools

a) Rubric

The rubric used for providing feedback on compare and contrast essays was developed by integrating elements from two main sources. Some of the items were derived from the instructional content related to compare and contrast essays, as outlined by Buitrago and Díaz (2018). Additionally, for some of the items in the organization section of the rubric, an academic writing book titled “Writing to communicate: Paragraphs and Essays” which is written by Boardman and Frydenberg (2002) was consulted. The rubric was made up of sections related with content and organization. Under the content category, there were 3 content-related questions for the introduction paragraph, 4 questions for the similarity paragraph, 4 questions for the difference paragraph and 3 questions for the conclusion paragraph. As for the organization category, there were 5 questions in total which are pertinent to unity, coherence and cohesion aspects. To promote the validity and reliability of the rubric, expert opinion was gained from 2 academicians: one having a PhD degree in English Language Teaching and the other having 15 years of teaching experience in a university context. Both of these academicians also had more than 10 years of experience in teaching and grading college-level essay writing. The rubric was piloted with two essays by these academicians, who later provided some wording suggestions for some items to make them clearer. These suggestions were discussed and incorporated into the final version of the rubric.

The rubric was utilized to assess the accuracy of the feedback provided by ChatGPT (See Appendix A). Researchers added an evaluation range to the rubric criteria for this purpose. A three-point evaluation scale was defined as below:

1: Very poor: The feedback is entirely incorrect/irrelevant.

2: Average: The feedback is correct but not comprehensive enough.

3: Very good: The feedback is entirely accurate and comprehensive.

Additionally, a comment section was added for noting differences between feedback provided by ChatGPT and teacher feedback, facilitating the qualitative analysis.

b) Prompt

This part explains how we structured our prompts to get feedback from ChatGPT. As ChatGPT works based on the given prompts, how a person structures a prompt is strongly linked to the output ChatGPT creates. First, we tried two zero-shot prompts ("Can you provide feedback on the student's essay? /Can you provide feedback on the content and organization of the student's essay?") to observe what feedback ChatGPT generates. While these prompts generated a substantial amount of feedback, it lacked sufficient structure. Then we analyzed the general guidelines suggested for prompt engineering as well as the various prompts formed in EFL studies on feedback to improve the quality of feedback. In light of these studies, two distinct prompts were generated for feedback on content and organization.

Specifically, in order to create the necessary prompt structure, we specified the role (the persona the AI model adopts when responding) (Bonner et al., 2023; Steiss et al., 2024; Su et al., 2023), student level (the proficiency level of the student the AI model is addressing) (Huang, 2023), expectation (the desired outcome or standard the AI model's response should meet) (Huang, 2023; Steiss et al., 2024), tone (the emotional or stylistic approach the AI model uses in its responses) (Steiss et al., 2024), and criteria (the specific standards or metrics for the AI model's output) (Carlson et al., 2023; Pack & Maloney, 2023). We formed few-shot prompts and in the end, specified fine-tuned prompts. For the fine-tuned prompt, we iteratively prompted ChatGPT to obtain the best refined feedback results based on the evaluation criteria we provided. Ultimately, this process resulted in fine-tuned prompts tailored for optimal feedback generation. We refined our last effective prompt to the following:

> As an English language instructor, generate feedback based on the comparison-contrast essay provided in this session. Students' English level is [B1]. Use a friendly and encouraging tone with simple language. If needed, provide examples of how the student could improve the essay. Instead of rewriting the paragraph, give specific examples and guidelines on how to revise. Be clear and specific in your feedback, and try to include as many corrections as possible. While giving feedback, just focus on ***[****the criteria lists given to you for the introductory, similarity, difference and conclusion paragraphs****.] [****the organization criteria list given to you.****]***

For the complete working prompt, see Appendix B.

2.4. Procedure

To initiate the research process, the essays written previously by students in an essay writing class taught by the second author of the current study were obtained. Later evaluation criteria for these essays were developed according to the instructional content covered in that class. Thus, the essays, evaluation criteria, and prompts served as necessary input required for the feedback to be provided by ChatGPT. For accuracy check, the second author of the study initially provided her own feedback on the 10 essays using the identified evaluation criteria while the other two authors ran ChatGPT sessions for gaining feedback on these essays. The ChatGPT feedback was conducted using April 2024 version of ChatGPT 3.5. Content feedback and organization feedback prompts were run in separate sessions to prevent ChatGPT from learning from sessions and generating unwanted pieces of feedback. After the process of ChatGPT feedback was completed and the feedback was stored in a text document, the accuracy check of those feedback was done by the second author based on the feedback she gave on

the same essays before. That is, the researcher compared her own feedback on each item in the rubric with ChatGPT feedback and gave scores for their accuracy. For reliability purposes, this accuracy analysis was also done by another expert in English Language Teaching (ELT) who used the same rubric to give feedback on the student essays. In relation to the second research question, the length of the ChatGPT feedback was calculated using a word count function.

The time spent generating feedback was also noted to clarify the procedure. Accordingly, the total duration for teacher feedback was 11 hours and 40 minutes, with an average of 1 hour and 10 minutes spent on each essay. Producing feedback for 10 essays via ChatGPT took 2 hours and 13 minutes, averaging 13.3 minutes per essay.

2.5. Data analysis

To answer the first research question, quantitative data obtained from the evaluation rubric was analyzed descriptively. The frequencies of the "Very poor," "Average," and "Very good" categories were reported for each criterion of the content and organization components. Additionally, the mean (M) and standard deviation (SD) values were reported. Furthermore, the feedback produced by ChatGPT was compared to teacher feedback qualitatively. Accordingly, comments generated for each essay were categorized into themes using thematic analysis (Braun & Clarke, 2006) and were further elucidated with relevant quotations.

Regarding the second research question, the word counts of ChatGPT and teacher feedback given for content, organization, and total essays were analyzed descriptively. The minimum (min.), maximum (max.), M, and SD values were reported. To determine if there was a significant difference between the length of teacher feedback and ChatGPT feedback, the non-parametric Wilcoxon Signed Ranks Test was conducted using IBM SPSS Statistics Version 23. In this examination, the dependent variable, feedback length, is at the continuous level, with quantification reliant on word count. The independent variables comprise two matched pairs, which were total ChatGPT feedback length and total teacher feedback length; ChatGPT feedback length on content and teacher feedback on content; ChatGPT feedback length on organization and teacher feedback on organization. Three distinct comparisons were executed employing a single test. Consequently, a Bonferroni correction was implemented to mitigate the effects of multiple comparisons, thereby establishing the significance threshold at .016, which was adjusted from .05 divided by 3.

2.6. Reliability and validity

The criteria list used for generating ChatGPT feedback for the EFL compare and contrast essays, encompassing content and organization aspects, was developed by the second author who is an expert in the field of English language teaching with graduate degree in the ELT department and experience in teaching and researching writing for almost 10 years. The resulting criteria list underwent further scrutiny by two academicians, each with over 10 years of experience in teaching essay writing at college level. These academicians piloted the rubric with two compare and contrast essays to validate its effectiveness. The piloting resulted in revisions in the wording of some items, producing the final version of the rubric.

The ChatGPT prompt was tested and refined by two experts, both with over 10 years of experience in the field of Computer Education and Instructional Technology (CEIT), who are also the authors of this paper. The refinement process was guided by prompt criteria outlined in the literature. The two essays used during the rubric testing were also employed to check the functionality of the prompt. The ChatGPT feedback generated for the two essays was reviewed by the second author, further clarifying the final version of the developed prompt. The ChatGPT feedback was checked as a precaution against the potential risk of generating inaccurate feedback. For reliability of data analysis, accuracy

assessment was conducted by 2 researchers. One of them was the second author of the study while the other assessor was a researcher with a PhD in ELT and taught essay writing at university for over 10 years. To ensure a true understanding of the criteria, the second author provided explanations for each item and guided the other researcher in assessing the accuracy of a separate essay, which was not one of the 10 chosen essays. Then, ten pieces of ChatGPT feedback data were independently evaluated by two EFL experts based on the rubric. The agreement rate of these evaluation sets was computed individually for each essay. Two essays yielded a consensus of 91%, while two others reached 95.5%. The consensus of 100% was attained for the remaining six essays. Instances of discordance were subjected to deliberation by two domain experts in order to reach a consensus. Miles and Huberman (1994) set an 80% agreement level as an acceptable threshold, and the acquired values fulfill this criterion of reliability.

2.7. Ethical Procedure

The ethics committee report for this study was obtained from the Zonguldak Bülent Ecevit University Human Research Ethics Committee with the decision dated 29.05.2014 and numbered 2014/08-13.

3. Findings

3.1. Feedback accuracy

The accuracy of ChatGPT feedback has been evaluated using a three-point scale, taking the teacher feedback, which serves as a reliable benchmark of expert judgment, as a baseline. Regarding the findings related to content feedback, it is found that ChatGPT's feedback aligns nearly with three criteria of the teacher's feedback for the introduction paragraph of the essay. Relatively less alignment of ChatGPT feedback is found in the feedback provided for the conclusion paragraph. Overall, the mean values of the assessments are above 2.50, with a calculated total content mean value of 2.63 (*SD* = .17) (See Table 1).

Table 1

Feedback Accuracy of ChatGPT Feedback on the Content of the Compare and Contrast Essays

Feedback Accuracy	**Introduction Paragraph (IP)**			**Similarity Paragraph (SP)**				**Difference Paragraph (DP)**				**Conclusion Paragraph (CP)**		
	IP1	*IP2*	*IP3*	*SP1*	*SP2*	*SP3*	*SP4*	*DP1*	*DP2*	*DP3*	*DP4*	*CP1*	*CP2*	*CP3*
Very poor	0	0	0	2	1	1	2	0	1	1	3	2	1	1
Average	0	0	1	1	1	2	3	3	1	1	0	2	4	3
Very good	10	10	9	7	8	7	5	7	8	8	7	6	5	6
Mean	2.97			2.53				2.63				2.43		
SD	.10			.39				.37				.35		

Note1: A total of 10 essays are evaluated for each criterion.

Note 2: M and SD values are reported for each paragraph.

It has been determined that ChatGPT feedback is 100% accurate regarding the criteria of having an attention-grabbing hook sentence in the essay (IP1) and including background information in the introductory paragraph (IP2) that describes the context of the topics. However, it has been observed that in the criteria of restating the thesis (CP1), summarizing the similarities and differences written in previous paragraphs (CP2), and stating the students' own opinion about the topic in the conclusion

paragraph (CP3), ChatGPT's feedback is less accurate compared to the feedback given for the other paragraphs.

According to the findings related to organization feedback, it has been found that ChatGPT's feedback on the organization of the essay generates mostly accurate feedback in terms of unity, coherence, and cohesion. Particularly, its feedback on the cohesion of the similarity (O3) and difference paragraphs (O4) has been found to be 100% accurate. The lowest accuracy rate was identified in the feedback generated for the unity criterion (O1). In general, the average scores for the accuracy evaluations exceed 2.50, with a computed total mean score for organization of 2.84 (*SD* = .22) (See Table 2).

Table 2

Feedback Accuracy of ChatGPT Feedback on the Organization of the Compare and Contrast Essays

Feedback Accuracy	***O1 Unity***	***O2 Coherence***	***O3 Cohesion of SP***	***O4 Cohesion of DP***	***O5 Cohesion of CP***
Very poor	1	0	0	0	0
Average	3	1	0	0	2
Very good	6	9	10	10	8
Mean	2.50	2.90	3.00	3.00	2.80
SD	.70	.31	.00	.00	.42

Note1: A total of 10 essays were evaluated for each criterion.
Note 2: M and SD values are reported for each criterion.

The unity criterion, for which relatively less accurate feedback is generated, pertains to the relevance and consistent adherence to the central theme throughout both individual paragraphs and the entire essay. The coherence of paragraphs discussing similarities (O3) and differences (O4), evaluated with ChatGPT's feedback being entirely accurate, is judged based on the standard of employing appropriate connectors to ensure clear transitions between the sentences.

Overall, according to the qualitative analysis, issues have been identified concerning the feedback provided by ChatGPT on both content and organization, particularly in relation to unity criterion. During the feedback generation process, an inclination has been noted for ChatGPT to concentrate on individual paragraphs, consequently neglecting the central theme of the essay. In essence, it falls short of providing a comprehensive evaluation of the essay. Regarding the depth of feedback, while in some instances, ChatGPT offers more informative feedback compared to teacher feedback, in others, it provides less informative feedback. The issue of non-informativeness was predominantly observed in the feedback provided for similarity and difference paragraphs. Lastly, it is found that ChatGPT is incapable of detecting certain errors. Moreover, there have been observations indicating that ChatGPT erroneously identifies non-existent errors as mistakes, as demonstrated by instances falling under the false positive theme (See Table 3).

Table 3

Qualitative Examination of the Accuracy of Feedback Provided by ChatGPT

Theme	***Code***	***Explanation***	***Quotation***
Unity	NA	It concerns the matter of relevance and the continual preservation of the central theme throughout both individual paragraphs and the entirety of the essay.	**Teacher feedback:** "*The similarities are off-topic and do not express a similarity that job and education shares. Take these similarities out and write new ones, please."* **ChatGPT feedback**: "*You've mentioned two similarities, taking notes and listening to the teacher, but they could be presented more clearly. For example, "Both classical education and modern learning techniques emphasize the importance of active listening and note-taking in the classroom."* **Feedback rubric criterion:** O3
Depth of feedback	Informativeness	It denotes the extent of feedback depth, offering thorough explanations in instances where a student makes an error.	**Teacher Feedback:** "*The essay does not present ideas in a well-organized way*." **ChatGPT feedback**: "*Your essay needs better coherence to help your reader follow your ideas more easily. Within each paragraph, your sentences should flow logically from one to the next. For instance, in your "Difference" paragraph, you mention taking notes, then suddenly jump to discussing the importance of the study environment. It would be clearer if you introduced each difference one at a time and provided examples or explanations for each. This way, your reader can understand each point before moving on to the next."* **Feedback rubric criterion:** O2
	Non-informativeness	It pertains to feedback that is superficially given when a student commits an error	**Teacher Feedback**: "*Use connectors to start the difference paragraph (e.g., Despite similarities), to introduce each difference (The first difference is that, the second difference is that). While describing the differences, please use such connectors as "On the contrary, while, despite, though, whereas". To provide examples, please use such connectors as "For example, For instance "and, "thus, therefore" for explanations*." **ChatGPT feedback**: "*Your use of connectors to link sentences and paragraphs is limited, which affects the overall clarity of your essay."* **Feedback rubric criterion**: O4
Capability	No feedback	It denotes the lack of any feedback given when it is necessary.	**Teacher feedback***:* "*Your last sentence is not meaningful. Hollywood is not a film. Please, remove it and state your own opinion about the topic*." ChatGPT could not generate any meaningful feedback. **Feedback criterion**: CP3
	False positive	It denotes a circumstance where feedback mistakenly suggests the existence of an error or characteristic when it is actually not present.	**ChatGPT feedback**: "*There is a need to start the conclusion with a clear transition: "In conclusion, while high school and university share some similarities, they also exhibit significant differences…*" Feedback was not provided as it was deemed unnecessary by the teacher. **Feedback criterion**: O5

Note1: Feedback criteria lists are provided on Appendix A.
Note 2: NA: Not Applicable, O: Organization, CP: Conclusion Paragraph

As presented in Table 3, the qualitative analysis of comments written on the accuracy of ChatGPT feedback revealed three key themes: unity, feedback depth, and capability. Concerning unity theme, as quoted, although a student wrote about similarities in her/his essay, ChatGPT did not correctly identify the relevance of these similarities to the main topic and instead provided feedback suggesting that the presented similarities were just unclear. In terms of feedback depth, which also affects its

length, two main themes emerged: informativeness and lack of informativeness. As seen in the examples presented in the Table 3, while the teacher's feedback notes that the student's ideas were not well-organized, ChatGPT's feedback explains this in greater detail. On the other hand, in some instances where the teacher offered a more detailed explanation, ChatGPT provided only a brief statement in its feedback. Regarding capability theme, we observe that in some instances, although the teacher provided feedback, ChatGPT did not give any feedback on the same points. On the other hand, there are cases where the teacher deemed feedback unnecessary, but ChatGPT still provided feedback.

3.2. Feedback length

The length of feedback provided by both ChatGPT and the teacher was compared based on the word count. The word count of the 10 student essays ranged from a min. of 184 to a max. of 239, with a mean of 216.00 and a SD of 19.26. Table 4 presents the descriptive statistics of feedback length.

Table 4

Feedback Length

Feedback focus	***Min.***	***Max.***	***Mean***	***SD***
ChatGPT feedback length on content	358	718	499.10	93.06
ChatGPT feedback length on organization	304	617	475.60	88.56
Total ChatGPT feedback length	1223	1548	1398.50	115.87
Teacher feedback on content	199	462	343.00	84.31
Teacher feedback on organization	52	246	146.00	56.28
Total teacher feedback length	406	786	611.60	129.51

Note: The feedback length is determined by the number of words.

Wilcoxon Signed Ranks tests were conducted to compare total ChatGPT feedback length and total teacher feedback length, ChatGPT feedback length on content and teacher feedback on content, as well as ChatGPT feedback length on organization and teacher feedback on organization. The tests elicit that ChatGPT created significantly higher length of feedback in three comparisons (Total ChatGPT-Teacher: Z = -2.803, p = .005; Content feedback- ChatGPT-Teacher: Z = -2.599, p = .009; Organization feedback-ChatGPT-Teacher: Z = -2.803, p = .005) at .016 significance level, based on two-tailed tests. Positive ranks were used for the calculation of Z-values (See Table 5).

Table 5

Wilcoxon Signed Ranks Test Statistics

	ChatGPT-Teacher *(Total)*	**ChatGPT-Teacher** *(Content)*	**ChatGPT-Teacher** *(Organization)*
Z	-2.803[a]	-2.599[a]	-2.803[a]
Asymp. Sig. (2-tailed)	.005	.009	.005

Note: a. Based on positive ranks.

4. Discussion

The first research question addressed how accurately ChatGPT generated feedback on the content and organization of EFL compare and contrast essays. The findings of the study demonstrated that the accuracy of ChatGPT feedback on the content and organization components of EFL compare and contrast essays was considerably high for each paragraph. These findings are in line with those of earlier studies. In their study, Banihashem et al. (2024) reported not finding any significant inaccuracy

in the feedback given by ChatGPT on argumentative essays. Likewise, the studies by Wang et al. (2024) and Steiss et al. (2024) substantiated the accuracy of the feedback provided by ChatGPT for argumentative texts. In a similar vein, Su et al. (2023) noted that ChatGPT was considerably competent in giving feedback on the argumentative essays in terms of content and organization aspects.

The feedback provided by ChatGPT on the introduction paragraph was found to be completely accurate in terms of content. However, the content-wise accuracy of feedback in similarity and difference paragraphs was lower than that in the introduction paragraph although those paragraphs still had a high accuracy rate. Additionally, the accuracy of content feedback was relatively lower in the conclusion paragraph than in the other paragraphs. The findings show that when there is a need for the linking of some ideas in different paragraphs, ChatGPT can fail to give accurate feedback since it cannot merge the ideas from different paragraphs effectively and considers the essay as a whole consisting of related parts. This can be considered as a unity problem, which represents ChatGPT's inability to look at the essay holistically. This finding is also evident in the analysis of feedback accuracy in the organization component, which demonstrates that unity is the feedback aspect having the least accuracy rate in that component. This problem was also highlighted in Steiss et al.' s (2024) study which revealed that ChatGPT failed to identify the mistake when a student confused a proper name for another proper name, which signified that ChatGPT did not understand the text as a problem related to unity.

As qualitative data show, ChatGPT occasionally fails to detect if the relevance of the ideas is maintained throughout the essay or not. Other identified distinctions in ChatGPT feedback related to depth of feedback and capability. These findings suggest that while using ChatGPT feedback, teachers should check if unity is achieved through maintenance of the main theme within and across the paragraphs. They also need to examine whether the feedback adequately addresses identified problems in a student essay, and whether additional feedback from the teacher is necessary. Furthermore, whether ChatGPT ignores some mistakes or gives wrong feedback although there is no need for feedback also needs to be checked by teachers. Concerning the capability theme, Guo and Wang's (2024) study also showed that ChatGPT could sometimes give irrelevant feedback. Under the capability theme, the issue related to overlooking necessary feedback was also noted by Wang et al. (2024) who put forward that teacher feedback gave more focused feedback addressing critical problems in essays, which can be ignored by ChatGPT. Related to the informativeness theme, the situations where ChatGPT or teacher gave more in-depth feedback than one another were observed. The latter situation can be explained by ChatGPT's lack of contextual information about the students and their progress as also emphasized in other studies (Guo & Wang, 2024; Wang et al., 2024). The in-depthness of the ChatGPT feedback can be attributed to its tendency to provide more directive feedback, unlike teachers, who, as noted in Guo and Wang's (2024) study, tend to offer more indirect feedback.

As for the second research question which investigated the difference in the length of feedback provided by ChatGPT and teacher, it was shown that ChatGPT provided longer feedback than teachers, which was corroborated by Guo and Wang (2024) and Wang et al. (2024). Considering that ChatGPT has paramount capacity to provide more voluminous and detailed feedback in just a few seconds in comparison to teachers who spend a greater amount of time, it can be stated that ChatGPT proves to be an efficient tool in terms of time and effort required for feedback effort. Therefore, when all the affordances and limitations of ChatGPT are taken into consideration, it can be argued that collaboration between teacher and ChatGPT is required for an optimal integration of ChatGPT feedback in L2 writing classes.

Finally, despite the capacity of LLMs to give feedback by analyzing grammar, cohesion, and style all at once (Bonner et al., 2023), its success is very much dependent on the prompt structure. For taking

advantage of ChatGPT, prompts should be written thoroughly (Carlson et al., 2023). Previous studies have already provided guidance for writing prompts, yet each situation might come with unique requirements and characteristics, and therefore, prompts that are tailored to the specific situations at hand should be ensured. In this way, the above-mentioned issues about the feedback output of ChatGPT can be eased, and this might help educators maximize the effectiveness, accuracy, quality, and practicality of feedback generation by creating their own fine-tuned prompts that align closely with the desired context and criteria.

5. Conclusion

Conclusively, it can be affirmed that ChatGPT delivers exceedingly precise feedback concerning both content and organizational aspects within EFL compare and contrast essays. It is inferred that ChatGPT provides feedback closely resembling teacher feedback for the introduction paragraph in terms of content criteria; nevertheless, it exhibits decreased accuracy in generating feedback for the conclusion paragraph based on the relevant criteria. This could be because assessing the conclusion paragraph requires a comprehensive view of the entire essay, including factors like paraphrasing the thesis statement, summarizing similarities and differences, and ensuring coherence across the paragraphs. ChatGPT's deficiency in maintaining unity in this aspect may have contributed to the relatively lower mean score. When providing feedback based on organizational criteria, it is deduced that ChatGPT performs well in generating feedback on the coherence of similarity and difference paragraphs. However, it produces less accurate feedback when evaluating the essay for unity and providing corresponding feedback. Although the error rate is low, upon examining the errors, it is concluded that ChatGPT excels in paragraph and sentence-level evaluations but encounters difficulties in evaluating the essay holistically. The comparatively greater length of ChatGPT feedback compared to teacher feedback highlights its strength over teacher feedback and indicates a potential solution to the general problem teachers face in providing detailed feedback to all students. The conclusion reached is that there is a necessity for collaboration between teachers and ChatGPT, rather than delivering ChatGPT feedback directly to students, at least with this version of ChatGPT. This suggests that ChatGPT feedback should be reviewed by teachers before being shared with students. This study shows that the precision of ChatGPT feedback is notably elevated contingent upon the prompt criteria, thereby offering guidelines outlining the facets of the generated ChatGPT feedback necessitating scrutiny by educators and researchers.

6. Implications

The results of this study provide a guideline for points to consider before utilizing ChatGPT feedback in research and classroom applications. Prior to providing ChatGPT feedback to students, it should be evaluated whether the feedback comprehensively addresses the essay, provides sufficient explanations, and identifies any areas where feedback is lacking or incorrect. Additionally, the high accuracy rate of ChatGPT feedback obtained in this study indicates that ChatGPT shows promise in providing educational feedback. Diversifying feedback evaluation criteria and conducting in-depth content analysis of generated feedback represent significant areas for research and application. These efforts are crucial for understanding how LLMs interpret written text, identifying areas of difficulty, and determining where they might outperform humans.

This study also provides implications for general prompt engineering and prompt engineering for generating feedback with GenAI tools, especially with ChatGPT. When the text provided to ChatGPT contains any automatic numbering/bullets, ChatGPT fails to recognize these numbers/bullets. Hence, it is advisable to avoid using automatic numbering/bullets and to use manual numbering/bullets for the piece of text intended for ChatGPT. As ChatGPT can pause at times, it is important to compare the generated feedback with the number of criteria provided to ChatGPT. This ensures that ChatGPT's

pauses do not lead to incomplete or insufficient feedback. There is always a risk for ChatGPT to incorporate the exact criteria provided or utilize sample structures from the criteria in its feedback. For this reason, it is crucial to ensure that the generated feedback does not replicate the exact structures from the given criteria. When the prompts are entered separately in different sessions, it generates more detailed feedback. In organization prompts, it is necessary to present the essay as a whole without any paragraph distinctions so that ChatGPT can evaluate the overall organization of the essay effectively. Additionally, it is necessary to connect the sections of the essay that refer to each other. For instance, ChatGPT needs to check the introductory paragraph again to see if the thesis statement is restated in the conclusion section since thesis statements are initially presented in the introductory paragraph of the compare and contrast essay.

7. Limitations and Recommendations

In this study, the accuracy and length of ChatGPT feedback were experimentally analyzed by researchers. Evaluating the effectiveness of ChatGPT feedback from the perspective of students and examining its direct impact on students' essays are areas that remain underexplored and are suggested as future research topics. Additionally, technology-related challenges in utilizing ChatGPT for feedback generation by teachers were not investigated, given the emphasis on research over practical classroom implementation. Future research could involve comparing the teacher-generated feedback with that of ChatGPT feedback in terms of efficiency.

Within this research, the feedback length was calculated by considering the total word count of the content generated by ChatGPT and teachers. In future research, determining feedback idea units and conducting comparisons based on feedback types could be beneficial in understanding the potential of ChatGPT feedback. Concerning LLM model, feedback was generated using the GPT-3.5 version. Comparing the feedback produced by different LLMs could be valuable in understanding the potential of models in providing educational feedback.

This study produced feedback specifically tailored for compare and contrast type essays. It is suggested that further research examines the accuracy and effectiveness of AI-generated feedback for various other essay types within the EFL context. Further studies can also focus on the prompt structures and provide a comparative analysis of generated feedback by different prompt structures. Finally, this study sampled 10 student essays. Working with a larger sample could more clearly elucidate potential issues/strengths in ChatGPT's feedback generation.

References

Bakla, A. (2020). A mixed-methods study of feedback modes in EFL writing. *Language Learning & Technology, 24*(1), 107–128. https://doi.org/10125/44712

Banihashem, S, K., Kerman, N.T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peergenerated or AI-generated feedback? I*nternational Journal of Educational Technology in Higher Education, 21*(23), 1-15. https://doi.org/10.1186/s41239-024-00455-4

Barrett, A., & Pack, A. (2023). Not quite eye to AI: Student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education*, *20*(59). https://doi.org/10.1186/s41239-023-00427-0

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

Barrot, J. S. (2023b). Using automated written corrective feedback in the writing classrooms:

Effects on L2 writing accuracy. *Computer Assisted Language Learning*, *36*(4), 584-607. https://doi.org/10.1080/09588221.2021.1936071

Baskara, F. (2023). Integrating ChatGPT into EFL writing instruction: Benefits and challenges. *International Journal of Education and Learning, 5*(1), 44-55. https://doi.org/10.31763/ijele.v5i1.858

Biber, D., Nekrasova, T., & Horn, B. (2011). The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. *ETS Research Report Series, 2011*(1), i-99. https://doi.org/10.1002/j.2333-8504.2011.tb02241.x

Boardman, C.A., & Frydenberg, J. (2002). *Writing to communicate: Paragraphs and Essay* (2nd ed.). Pearson Education.

Bok, E., & Cho, Y. (2023). Examining Korean EFL College Students' Experiences and Perceptions of Using ChatGPT as a Writing Revision Tool. *Journal of English Teaching through Movies and Media, 24*(4), 15-27. https://doi.org/10.16875/stem.2023.24.4.15

Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology*, *23*(1), 23-41. https://doi.org/10.56297/BKAM1691/WIEO1749

Bozkurt, A. (2024). Tell me your prompts and I will make them true: The alchemy of prompt engineering and generative AI. *Open Praxis, 16*(2), 111-118. https://doi.org/10.55982/openpraxis.16.2.661

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

Buitrago, C. R., & Diaz, J. (2018). Flipping your writing lessons: Optimizing time in your EFL writing classroom. In J. Mehring & A. Leis (Eds.), *Innovations in flipping the language classroom.* (pp. 69–91) Springer.

Burke, D. M., & Pieterick, J. (2010). *Giving Students Effective Written Feedback*. Maidenhead: Open University Press.

Carlson, M., Pack, A., & Escalante, J. (2023). Utilizing OpenAI's GPT-4 for written feedback. *TESOL Journal*, *759*, e759. https://doi.org/10.1002/tesj.759

Chaka, C. (2023). Generative AI chatbots-ChatGPT versus YouChat versus Chatsonic: Use cases of

selected areas of applied English language studies. *International Journal of Learning, Teaching and Educational Research*, *22*(6), 1-19. https://doi.org/10.26803/ijlter.22.6.1

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

Elola, I., & Oskoz, A. (2016). Supporting second language writing using multimodal feedback. *Foreign Language Annals, 49*(1), 58-74. https://doi.org/10.1111/flan.12183

Fu, Q. K., Zou, D., Xie, H., & Cheng, G. (2024). A review of AWE feedback: types, learning outcomes, and implications. *Computer Assisted Language Learning, 37*(1–2), 179–221. https://doi.org/10.1080/09588221.2022.2033787

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

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,* 8435-8463*.* https://doi.org/10.1007/s10639-023-12146-0

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, *13*(1), 18617. https://doi.org/10.1038/s41598-023-45644-9

Huang, J. (2023). Engineering ChatGPT Prompts for EFL Writing Classes. *International Journal of TESOL Studies, 5*(4), 73-79. https://doi.org/10.58304/ijts.20230405

Jackson, D., Davidson, P.M., & Usher, K. (2022). *Feeding Back and Feeding Forward*. In: Successful Doctoral Training in Nursing and Health Sciences. Springer, Cham. https://doi.org/10.1007/978-3-030-87946-4_5

Köroğlu, & A. Çakır (Ed.), *Fostering foreign language teaching and learning environments with contemporary technologies*. (pp. 115-133). IGI Global. https://doi.org/10.4018/979-8-3693-0353-5.ch006

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*(2), 211–232. https://doi.org/10.3102/00346543047002211

Lee, A. V. Y. (2023). Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation. *Studies in Educational Evaluation*, *77*, 101250. https://doi.org/10.1016/j.stueduc.2023.101250

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies*, 1-33. https://doi.org/10.1007/s10639-023-12249-8

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

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, *49*(4), 102720. https://doi.org/10.1016/j.acalib.2023.102720

Liu, C., Hou, J., Tu, Y. F., Wang, Y., & Hwang, G. J. (2021). Incorporating a reflective thinking promoting mechanism into artificial intelligence-supported English writing environments. *Interactive Learning Environments, 31,* 3340–3359. https://doi.org/10.1080/10494820.2021.2012812

Merriam-Webster. (n.d.). Get Looped in on 'Feedback' In Merriam-Webster.com dictionary. Retrieved from https://www.merriam-webster.com/wordplay/the-history-of-feedback#:~:text=Feedback%2C%20which%20began%20as%20an,of%20coming%20out%20of%20it

Meyer, J. G., Urbanowicz, R. J., Martin, P. C., O'Connor, K., Li, R., Peng, P. C., ..., & Moore, J. H. (2023). ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, *16*(1), 20. https://doi.org/10.1186/s13040-023-00339-9

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

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

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics, 3*(2), 100116. https://doi.org/10.1016/j.rmal.2024.100116

Noroozi, O., Banihashem, S. K., Taghizadeh Kerman, N., Parvaneh Akhteh Khaneh, M., Babayi, M., Ashrafi, H., & Biemans, H. J. (2023). Gender differences in students' argumentative essay writing, peer review performance and uptake in online learning environments. *Interactive Learning Environments, 31* (10), 6302–6316. https://doi.org/10.1080/10494820.2022.2034887

OpenAI. (2024a). What is ChatGPT? Retrieved from https://help.openai.com/ en/articles/ 6783457-what- is- chatgpt

OpenAI. (2024b). Best practices for prompt engineering with the OpenAI API. Retrieved from https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the- openai-api

Pack, A., & Maloney, J. (2023). Potential affordances of generative AI in language education: Demonstrations and an evaluative framework. *Teaching English with Technology, 23*(2), 4-24. https://doi.org/10.56297/BUKA4060/VRRO1747

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research methods in Applied Linguistics, 2*(3), 100083. https://doi.org/10.1016/j.rmal.2023.100083

Punar Özçelik, N., & Yangın Ekşi, G. (2024). Cultivating writing skills: the role of ChatGPT as a learning assistant – a case study. *Smart Learning Environments, 11*(10), 1-18. https://doi.org/10.1186/s40561-024-00296-8

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

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

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

Schmidt, R., & Frota, S. N. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 237-326). Newbury House

Shadiev, R., & Feng, Y. (2023). Using automated corrective feedback tools in language learning: A review study. *Interactive Learning Environments*, 1-29. https://doi.org/10.1080/10494820.2022.2153145

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Spasić, A. J., & Janković, D. S. (2023, June 29-July 1). *Using ChatGPT standard prompt engineering techniques in lesson preparation: Role instructions and seed-word prompts* [Paper presentation]. 58th International Scientific Conference on Information Communication and Energy Systems and Technologies (ICEST), Nis, Serbia. https://doi.org/10.1109/ICEST58410.2023.10187269

Stebbins, R. A. (2001). *Exploratory research in the social sciences* (Vol. 48). Sage.

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

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

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

Thi, N. K., & Nikolov, M. (2021). Feedback Treatments, Writing Tasks, and Accuracy Measures: A Critical Review of Research on Written Corrective Feedback. *Tesl-Ej*, *25*(3), n3. Retrieved from https://files.eric.ed.gov/fulltext/EJ1332267.pdf

Üstünbaş, Ü. (2024). EFL Learners' Views About the Use of Artificial Intelligence in Giving Corrective Feedback on Writing: A Case Study. In *Fostering Foreign Language Teaching and Learning Environments With Contemporary Technologies* (pp. 115-133). IGI Global.

Wang, D., Su, J., & Yu, H. (2020). Feature extraction and analysis of natural language processing for deep learning English language. *IEEE Access*, *8*, 46335-46345. https://doi.org/10.1109/ACCESS.2020.2974101

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*, 101440. 46335-46345. https://doi.org/10.1016/j.tsc.2023.101440

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

Xiao, Y., & Zhi, Y. (2023). An Exploratory Study of EFL Learners' Use of ChatGPT for Language Learning Tasks: Experience and Perceptions. *Languages, 8*(3), 212. https://doi.org/10.3390/languages8030212

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

Zainurrahman, & Rojab, S. R. (2024). Examining Bing AI as a Solution to EFL Writing Feedback Challenges. *PROJECT (Professional Journal of English Education), 7*(2). Retrieved from https://journal.ikipsiliwangi.ac.id/index.php/project/article/view/21639

Zhai, N., & Ma, X. (2023). The Effectiveness of Automated Writing Evaluation on Writing Quality: A Meta-Analysis. *Journal of Educational Computing Research, 61*(4), 875-900. https://doi.org/10.1177/07356331221127300

Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing, 36*, 90-102. https://doi.org/10.1016/j.asw.2018.02.004

Article Information Form

Authors Notes: The authors would like to express their sincere thanks to the editor and the anonymous reviewers for their helpful comments and suggestions.

Authors Contributions: All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

Conflict of Interest Disclosure: No potential conflict of interest was declared by the authors.

Supporting/Supporting Organizations: No grants were received from any public, private or non-profit organizations for this research.

Ethical Approval and Participant Consent: It is declared that during the preparation process of this study, scientific and ethical principles were followed and all the studies benefited from are stated in the bibliography. The ethics committee report for this study was obtained from the Zonguldak Bülent Ecevit University Human Research Ethics Committee with the decision dated 29.05.2014 and numbered 2014/08-13.

Plagiarism Statement: This article has been scanned by Turnitin.

Appendix

Appendix A. Rubric for Assessing Content and Organizational Structure in Compare and Contrast Essays

Compare and Contrast Essay Content Feedback Criteria List				
	Assess the accuracy the feedback provided on a scale of 1-3.			
Feedback Criteria *(Introduction paragraph-IP)*	**1**	**2**	**3**	**Comment**
IP1-Is there an engaging hook in the introduction paragraph that grabs the reader's interest?				
IP2-Does the introduction paragraph include background information that contextualizes the topic being discussed?				
IP3-Does the introduction paragraph utilize an expression to present the thesis statement? *(e.g., This essay is written in order to ... The purpose of this essay is to... This essay aims at ____(ing)... This essay compares and contrasts... This essay discusses...)*				
Feedback Criteria *(Similarity paragraph-SP)*	**1**	**2**	**3**	**Comment**
SP1-Does the similarity paragraph begin with a topic sentence?				
SP2- Does the similarity paragraph present two similarities about the selected topics and compare them?				
SP3-Does the similarity paragraph incorporate examples and/or explanations to uphold the two similarities?				
SP4-Does the similarity paragraph conclude with a summarizing sentence including the two similarities?				
Feedback Criteria *(Difference paragraph-DP)*	**1**	**2**	**3**	**Comment**
DP1-Is there a topic sentence in the difference paragraph?				
DP2-Does the difference paragraph present two differences about the selected topics and contrast them?				
DP3-Does the difference paragraph incorporate examples and/or explanations to uphold the two differences?				
DP4-Does the difference paragraph conclude with a summarizing sentence including the two differences?				
Feedback Criteria *(Conclusion paragraph-CP)*	**1**	**2**	**3**	**Comment**

CP1-Does the conclusion paragraph provide a restatement of the thesis statement?				
CP2-Does the conclusion paragraph provide a summary of the similarities and differences discussed in the similarity and difference paragraphs?				
CP3-Does the conclusion paragraph express the student's personal viewpoint on the topic?				
Compare and Contrast Essay Organization Feedback Criteria List				
Feedback Criteria (O: Organization) (O1: Unity; O2: Coherence; O3, O4, O5: Cohesion)	**1**	**2**	**3**	**Comment**
O1-How well does the essay maintain unity? ***Unity:*** *This pertains to the issue of relevance and the consistent maintenance of the central theme within both individual paragraphs and the entirety of the essay. Unity within a paragraph is achieved when the supporting sentences enhance comprehension of the main point introduced at the paragraph's outset.*				
O2-How well does the essay maintain coherence? ***Coherence:*** *This pertains to the logical progression and linking of ideas within a sentence, the connection between sentences (the transitions between them) within a paragraph, and the continuity across paragraphs.*				
O3: Does the similarity paragraph incorporate connectors to begin the similarity paragraph, present each similarity, offer examples/explanations and provide a concluding statement? *(To begin with, the first similarity is..., for example, for instance, the second similarity is that..., also, as well as, as, both, most important, likewise/like, in the same manner /way, same/similar/similarly, the same as, too, in brief)*				
O4: Does the difference paragraph incorporate connectors to begin the difference paragraph, present each difference, contrast each topic, offer examples/explanations and provide a concluding statement? *(The first difference is that..., on the contrary, while, despite, though, whereas, for example, for instance, thus, therefore, in brief)*				
O5: Does the conclusion paragraph incorporate connectors to begin the paragraph, mention similarities, differences and state the student's personal opinion on the topic? *(In conclusion, as a result, to conclude, to sum up, both, in addition, on the contrary, as far as I am concerned, to my view, it is my impression that, from my point of view)*				

Appendix B. Prompts

Content prompt

As an English language instructor, generate feedback based on the comparison-contrast essay provided in this session. Students' English level is B1. Use a friendly and encouraging tone with simple language. If needed, provide examples of how the student could improve the essay. Instead of rewriting the paragraph, give specific examples and guidelines on how to revise. Be clear and specific in your feedback, and try to include as many corrections as possible. While giving feedback, just focus on the criteria lists given to you for the introductory, similarity, difference and conclusion paragraphs.

Introductory paragraph: "*paste here*"
Similarity paragraph: "*paste here*"
Difference paragraph: *"paste here"*
Conclusion paragraph: *"paste here"*
Introductory paragraph criteria list: *"paste here"*
Similarity paragraph criteria list: *"paste here"*
Difference paragraph criteria list: *"paste here"*
Conclusion paragraph criteria list: *"paste here"*

Organization prompt

As an English language instructor, generate feedback based on the comparison-contrast essay provided in this session. Students' English level is B1. Use a friendly and encouraging tone with simple language. If needed, provide examples of how the student could improve the essay. Instead of rewriting the paragraph, give specific examples and guidelines on how to revise. Be clear and specific in your feedback, and try to include as many corrections as possible. While giving feedback, just focus on the organization criteria list given to you.

Organization criteria list: *"paste here"*
Essay: *"paste here"*

Note: The criteria list can be found in Appendix A. The complete research data and prompts are available upon request from the corresponding author.