ALTRALANG Journal
e-ISSN: 2710-8619 p-ISSN: 2710-7922
Volume 8 Issue 1 / June 2026 pp. 288-303

Exploring the Role of ChatGPT-4 in Assessing and Enhancing Students' Writing Portfolios: A Formative Feedback Analysis across Multiple Essays

Aicha Imane Beesadat[1*]
[1]*University of Ghardaia, Algeria*
imidickens7@gmail.com
Souhila Korichi[2]
[2]*University of Laghouat, Algeria*
hm_kouidri@yahoo.fr

Received: 30/03/2024, **Accepted:** 17/03/2026, **Published:** 05/06/2026

ABSTRACT: *Several studies in the EFL realm have validated the effectiveness of portfolio assessment for not only giving a nuanced image of the students' writing competency but also to track the development of the students' writing performance. Although it enjoys a high level of reliability and validity, portfolio assessment is still criticized for being lesser user-friendly as it does not only assess the writing performance, but it also tends to track the development of EFL learners' writing proficiency. For this reason, this paper addresses the use of ChatGPT-4 as an artificial intelligence tool to alleviate some of the challenges of traditional assessment methods. Therefore, the focus of this paper is to evaluate the effectiveness of ChatGPT-4 in providing accurate assessments of students' written portfolios based on a predefined rubric. Further, the article tends to investigate the impact of receiving formative feedback from ChatGPT-4 on students' writing improvement across multiple essays, aiming to understand the influence of the model's feedback on enhancing students' writing skills over time. In this vein, the focus of this article is to examine the correlation between the progression of formative feedback from ChatGPT-4 and the improvement in students' writing proficiency over eight essays, aiming to elucidate the relationship between the quality and consistency of the model's feedback and students' writing development throughout the assessment period. To fulfil the targeted objectives, the researchers opted for an experimental protocol to a sample of fifty students from the university of Ghardaia- Algeria. They collected data from their portfolios which comprise of eight essays across two months. After that, a survey was conducted with the students to give their views about the quality of assessment received and its impact on their writing performance. The results showed that there is a remarkable advancement in the students' writing skills thanks to the personalized feedback furnished by ChatGpt-4. In addition, students were generally positive about the potential of the consistent received feedback to target their writing errors and to provide a detailed perception to their performance.*

KEYWORDS: artificial intelligence language model, ChatGPT-4, EFL writing skill, portfolio assessment.

* Corresponding author
ALTRALANG Journal / © 2026 The Author(s). Published by the University of Oran 2 Mohamed Ben Ahmed, Algeria.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

1 Introduction

As the world has been fast developing in technology and innovation, people were taken aback in December 2022 by the release of ChatGPT, a sophisticated language model powered by artificial intelligence. This incredible technology could have the potential to transform the ability to transform daily practices for research around the world, providing an unparalleled level of support and expertise in a variety of subjects. ChatGPT uses deep learning and natural language processing (NLP) to engage in human-like conversations, interpret difficult queries, and respond in an intelligent and relevant manner (Tlili & al, 2023). Then, logically, its ability to recognize context, generate coherent and contextually relevant language, and adapt to different domains is likely to make it also a good tool that can assist in education generally (Ibid) and in EFL teaching and learning specifically.

Given the fact that the EFL writing instruction often depends heavily on rich feedback; supervision, and support from teachers which are facilitated by means of assessment, the integration of technology becomes not only desirable but necessary. AI-powered tools could potentially offer a potential solution to the labor-intensive nature of writing assessment, especially portfolio assessment.

It is widely agreed upon among scholars that one of the most exhausting, complicated, and challenging tasks of assessment is portfolio assessment (Birgin & Baki, 2007; Cole et al., 1997; Daud, 2017; Hammad Al-Rashidi et al., 2023; Phung, 2016; Popescu-Mitroia et al., 2015; Tangdhanakanond & Wongwanich, 2015). Though it proves to be a potential way to track the development of learners in writing and ensure continuous feedback on the students' level, the application of portfolio assessment in the EFL context is still challenging due to many reasons. First, portfolio assessment can be an exhausting task because it requires significant time and effort from the part of evaluators especially if they are in charge of evaluating a large number of portfolios. Lo (2010), in her action research, points to the highly laborious nature of portfolio assessment though the set of portfolios assessed by the researcher comprises answers to an extremely limited set of questions. According to her, not only is portfolio assessment time and effort-consuming, but the heavy workload forces teachers to give a relatively low quality of assessment especially if there is a deadline for grade submission. This situation pushes evaluators to write short comments and descriptions to the students' level in addition to a lack of correcting grammatical errors (Ibid). Lo crowns the experience of portfolio assessment as "a massive burden" (Ibid, p. 87). In this regard, the primary goal of portfolio assessment to give detailed overview of students' level and development is not fulfilled because giving incomplete or inaccurate feedback does not help students locate their strengths and lacunas in writing. Overall, while it is too early to draw firm conclusions, a pattern of impracticality seems to emerge regarding the implementation of portfolio assessment in the EFL context (Daud, 2017; Tisani, 2008) — particularly in large university classes or when evaluating extended written compositions, such as those required of English students at Algerian universities.

Second, portfolios, unlike standardized examinations, frequently lack explicit evaluation standards and rubrics, making it difficult to maintain uniformity and fairness across multiple assessors or institutions (Birgin & Baki, 2007; Tangdhanakanond & Wongwanich, 2015). In the lack of set guidelines, portfolio contents may be interpreted and evaluated differently. This raises the issue of reliability as portfolio assessment is primarily reliant on subjective judgment, since evaluators interpret and evaluate the material based on their own viewpoints and biases (Hudori et al, 2020). This subjectivity can lead to inconsistencies in grading and assessment outcomes. (Ibid, 85)

Although there is growing scholarly interest in portfolio assessment and increasing attention to the role of AI tools like ChatGPT in education, the practical integration of both—particularly in the context of large EFL classrooms in Algerian universities—remains significantly underexplored.

This study addresses a critical gap in current research: while several studies have examined the pedagogical value of portfolio assessment (e.g., Abdul Halim et al., 2019; Bedda & Benabdi, 2022; Birgin & Baki, 2007; Bouakel, 2021, 2022; Cole et al., 1997; Daud, 2017; Lo, 2010) and others have investigated the

educational implications of AI-powered tools like ChatGPT (e.g., Rane, 2024; Sumakul et al., 2022; Tlili et al., 2023; Yanhua, 2020), very few (e.g., Khasawneh et al., 2025; Mastrogiacomi, 2023) have explored how artificial intelligence can help overcome the practical challenges associated with implementing portfolio assessment in EFL contexts. This is particularly relevant in environments characterized by large class sizes and limited teacher resources, such as Algerian universities. The lack of empirical data regarding the feasibility, effectiveness, and reception of AI-assisted portfolio assessment in such contexts represents a pressing need in the field.

In response to this gap, the present study seeks to examine the potential of ChatGPT-4 as a supportive tool for EFL instructors engaged in portfolio-based writing assessment. As a relatively recent technological innovation, ChatGPT has sparked growing interest among educators who are eager to understand its capabilities, opportunities, and limitations in teaching and learning. While previous research has addressed the general application of NLP (Natural Language Processing) tools in educational assessment, few studies have evaluated the accuracy and pedagogical relevance of ChatGPT-generated feedback within the specific framework of portfolio assessment for EFL learners.

This study aims to provide a comprehensive analysis of both theoretical perspectives and practical implications, with a focus on how ChatGPT-4 may assist in tracking students' writing development, regulating the assessment process, and delivering more targeted and personalized formative feedback. In doing so, the research not only contributes to the growing body of literature on AI in education but also offers practical insights for improving EFL assessment practices in resource-constrained environments. Accordingly, the study is guided by the following research questions:

1. What is the impact of receiving formative feedback from ChatGPT-4 on students' writing improvement across a variety of essays?
2. How is the development of students' writing skills over the course of eight essays correlated with the progression of formative feedback generated by ChatGPT-4?

2 Defining ChatGPT-4

The technology of artificial intelligence has been kept on the back burner for over sixty years. Until the beginning of 2022, creating robots or chat(bots) that interact with human beings in a human-like manner has been a mere fiction till the emergence of ChatGPT-3 (Chat-Generative Pre-Trained Transformer) , a language model created by San Francisco-based company OpenAI on December 2022. It is almost one year since the release of this technology and its dispersal among people is epic. Over 180 million people are currently using ChatGPT marking an unprecedented historical record of "the fastest growing web platform ever" (Milmo, 2023). The eruption of this AI-tool marks a new era revolutionizing information processing, jobs, and all spheres of life across the world in an incredible manner.

In order to understand the mechanism through which ChatGPT operates, an accurate definition to AI is required. The most simplistic definition of AI is provided by Stanford University Professor John McCarthy in 1955 saying that it is "the science and engineering of making intelligent machines" (Rice et al., 2023). Logically, as the field witnesses advances, scholars are gradually being able to provide more detailed definitions as Russel and Norvig's. They classify the field into four categories: (1) systems that think like humans, (2) systems that think rationally, (3) systems that act like humans, (4) systems that act rationally (2020). It can be immediately noticed that the recurrent patterns in these categorizations are thinking, behaviour, the simulation of and similarity to humans, and rationality. The definition that seems more detailed is provided by UNICEF Office of Global Insight and Policy. UNICEF Office of Global Insight and Policy (2021) states that AI is the industry of machine-based systems which are able to make predictions, recommendations, or decisions based on human-defined objectives (p.16). Other important traits that characterize AI systems are their ability to function autonomously and the possibility to exhibit adaptive behavior depending on the changing contexts through

learning (Ibid). UNICEF's definition could cover the main traits of AI including the capacity to learn continuously from their interaction with human beings.

In the past few years, there have been many attempts to create natural language processing (NLP) systems that could mimic human conversation but the quality and the originality of the language produced by ChatGPT is unprecedented (Alawida et al., 2023; Christiano, 2024; Perrigo, 2022; Revell et al., 2024) . Thanks to its quality as having 94 layers and 175 billion parameters according to the update in January 2022, ChatGpt3 can generate original and relevant language to diverse contexts. While a recent report on October 2023 say that Gpt-4 "comprises eight models, each with 220 billion parameters. The total parameter count of GPT-4 is estimated to be around 1.76 trillion." (Schreiner, 2023)

In 2020, Weinberg has made a general comparison between GPT-3 (a preliminary version of ChatGPT-3) and the other preceding NLP systems. He states that what makes Gpt-3 distinct from the preceding NLP systems is its ability to perform new language tasks while previous NLP systems are usually "pre-trained on a large corpus of text and then fine-tuned in order to successfully perform a specific task" (Weinberg, 2020). Johnson et al. (2023) contend that ChatGPT is capable of incorporating the complexity of users' intentions by fine-tuning for conversational tasks utilizing reinforcement learning from human feedback. As a result, ChatGPT can effectively handle a variety of end-user jobs. In other words, this AI tool is not only pre-trained and fine-tuned to adapt to different situations, but it also learns continuously from human interaction.

3 The Implementation of Portfolio Assesment in Algerian Universities

Portfolio assessment is considered as a tool of authentic assessment. It has been commonly implemented in English language classrooms since the mid-1980s to assess learners' writing performance (Hudori et al, 2019; Lam & Lee, 2010). The change in the teaching instruction from teacher-centred approach to student-centred approach is one of the reasons that brought about the increasing adoption of portfolio assessment among educators (Boumediene et al., 2018). Portfolio assessment, as explained by Boumediene et al. (2018), involves an active participation of the student who takes a pivotal role in assessment rather than being a passive subject of traditional assessment. The uniquely promising aspect of portfolio assessment is not only assessing students' performance, but it combines learning, teaching, and testing altogether. (Ibid)

Broadly speaking, a writing portfolio is defined as the accumulation of students' achievements. Lam and Lee (2010) define it more specifically as a deliberate compilation of learners' writing that exhibits "the work, development, and achievement of the writer throughout the course of the portfolio" (p.54). In this definition, four essential qualities need to be met for portfolio assessment to fulfil its purpose namely: intentionality, systematicity, quantification of a product, and tracking a process over time. To this end, portfolio assessment not only provides a one-shot assessment of the students' performance, but it also tends to track the development of writing and serve as a formative assessment tool.

Though portfolio assessment is widely regarded as attaining satisfactory levels of reliability and validity in assessing writing (Davis & Ponnamperuma, 2005; Herman et al., 1993; Meeus et al., 2009, etc.), the implementation of portfolio assessment in Algerian university is still quite scarce (Bouakel, 2021; Boumediene et al., 2018). According to Bouakel (2021), there exist many reasons that hamper the implementation of portfolio assessment in Algerian universities. First, teachers lack training and ignore the practical and theoretical underpinnings of the implementation of portfolio assessment (Ibid, p.100). Second, portfolio assessment is not given due importance by the administration because the culture of summative assessment is the mostly endorsed philosophy in Algeria while formative assessment receives lesser if any importance at all (Ibid). In another study, Bouakel (2022) adds to these reasons, the students' lack of knowledge and motivation to engage in portfolio assessment as hindering the process. Similarly, Boudibi & Beleulmi (2024) posit that portfolio assessment as a learner-oriented assessment can be challenging to implement in Algerian universities for the majority of students prioritize grades. This leads them to overlook the opportunities and possible progress associated with portfolio assessment (Ibid). Furthermore, since teachers in Algerian universities have a lot of

tasks to realize beginning from lesson planning to research, assessing a large amount of papers in addition to official examinations can be a daunting task. As portfolio assessment requires time and commitment from the part of both teachers and students and despite its effectiveness in upgrading students' level, its implementation remains scarce in Algerian universities as it lacks practicality.(Boudibi & Beleulmi, 2024; Boumediene et al., 2018)

4 Employing Artificial Intelligence- Language Models for Assessment of EFL Learners' Writing Proficiency

With the advent of artificial intelligence, scholars have started to explore the possible intersections between AI and the demands of EFL teaching and learning (Dang & Wang, 2024; Jiang, 2022). This has given rise to numerous promises and concerns in the realm of EFL teaching and learning practices. A large body of literature suggests that AI can benefit language teaching and learning (Alhalangy & AbdAlgane, 2023; Dang & Wang, 2024; Jiang, 2022; Schmidt & Strasser, 2022; Sumakul et al., 2022; Yanhua, 2020). Nevertheless, reviews that focus on the insights of AI-assisted writing assessment in the EFL context are comparatively scarce.

Lu (2019) conducts an empirical research on the AI writing evaluation in China. The research scrutinizes the use of a popular system among Chinese teachers and students. The system entitled 'Juku' is used for assessing writing in college English teaching. Throughout the questionnaires and interviews with 30 teachers and 200 students, the research demonstrates the positive impact of this AI-based system on the students' English writing.

Wang (2020) investigates students' aspirations and perceived impact of AI- based assessment tools of EFL writing. Wang's inquiry compares the reliability and validity of human evaluation to automatic evaluation. It shows the students' high expectations to AI tools and a high rate of effectiveness of AI scoring feedback than that of teachers. Another interesting finding to this research includes a remarkable advancement in terms of autonomy and writing ability which in turn correlates positively with students' preliminary expectations of AI tools. Both Wang's (2020) and Lu's (2019) studies recommend that AI systems should be perfected to provide adequate "evaluation on the text structure, content logic, and coherence" (Lu, 2019, p.121).

Jiang et al., (2023) investigate the efficiency of four large language models (LLMs), GPT-4, GPT-3.5, iFLYTEK and Baidu Cloud in the writing assessment of second language learning. The researchers collected samples from students who study the Chinese language in an online high school in the US. The assessments gained from these LLMs were compared to human rating results. The investigation concludes that LLMs have the potential to support the writing assessment for language learners while remarking that GPT-4 is able to provide more precise writing assessments.

In response to these challenges, Xia et al., (2024), in their article entitled 'A scoping review on how generative artificial intelligence transforms assessment in higher education', suggest a set of measures to cope with the changes of homework submission pattern Generative AI

4.1 Method

In order to investigate the effectiveness of ChatGPT-4 in portfolio assessment for EFL writing, this study employed a mixed-methods approach amalgamating an experimental design and a quantitative and qualitative survey.

The experimental design was implemented over a period of eight weeks and involved a selected group of EFL students who were assigned multiple writing tasks. Throughout the experiment, students received formative feedback generated by ChatGPT-4, and their progress was tracked across successive writing submissions. This stage aimed to examine the impact of AI-generated feedback on students' writing development, thereby addressing both research questions.

Following the experimental phase, an online survey was administered to the same participants. The survey aimed to capture students' perceptions, experiences, and attitudes toward the consistent received feedback during the experiment. Although the feedback was generated by ChatGPT-4, this detail was not disclosed to participants in order to avoid response bias. The survey was not part of the experimental manipulation but served as a supplementary instrument to triangulate the findings and explore learner perspectives.

The experimental design was chosen to measure the effect of an intervention (ChatGPT-4 feedback) on a defined group over time, while the survey provided insights into the learners' subjective responses, thus ensuring both empirical rigor and contextual depth in understanding the tool's pedagogical potential.

4.2 Participants

Both phases were conducted online in the summer of 2023. A convenience sample of fifty university students studying English was recruited through voluntary self-selection. Standard ethical procedures were followed, and participants provided informed consent prior to participation.

The research questions were investigated in the experiment with a total of fifty students, 19 males and 31 females, falling between the ages of 21 and 35 years old. The first and the second research questions were addressed in the survey. The online questionnaire was administered to students. Inclusion criteria involved students who took part in the experiment, and participants meeting these criteria were invited to participate.

The announcement of the participation in the portfolio challenge was administered electronically through the third year Facebook group. The announcement invited students who are interested to receive personalized feedback on their essays to send an e-mail to the teacher in charge. The teacher communicated via e-mail with the participants during the experiment. Communicating with students electronically was to enhance accessibility and ease of participation.

4.3 Instruments

In order to track the student's development of writing proficiency, the experiment used an ANOVA test. As the choice of an appropriate statistical analysis depends on the characteristics and variables of the research study, this research opted for a one-way ANOVA test to compare between more than two groups according to only one factor which is ChatGPT-4 assessment. Frey (2018) maintains that "the goal of the ANOVA is to examine whether the mean scores for each group (…) are reliably different from each other" (p.2)

The characteristics of the design and the variables in a research study determine the appropriate statistical analysis. "Units" refer to the unit of analysis, usually subjects. In other words, a mixed model ANOVA is used for studies in which independent units are "crossed with" at least one of the independent variables and are "nested under" at least one of the independent variables.

In the challenge, students were required to write eight essays, one per week, on topics relevant to university students' interests. These topics were chosen to maintain students' motivation and interest throughout the two-month challenge. The guidelines were shared with students via email at the beginning of the writing challenge. These included detailed guidelines that match the main evaluation criteria in the rubric (structure, clarity and cohesion, originality) (see appendix 2). The same criteria were consistently referred to in the feedback students received after each essay, helping them track their progress and understand the basis for their scores.

ChatGPT-4 is selected, among other large language models as GPT-3.5, iFLYTEK and Baidu Cloud, based on the findings of many studies (e.g. Escalante et al., 2023; Jiang et al., 2023) that proved the outperformance of ChatGPT-4. In addition, ChatGPT-4, unlike GPT-3.5, has the ability to access files when they are uploaded by the user which makes it easier for teachers to manage by simply asking students to organize their work.

The following are the steps underlying the process of assessing students' works. First, the researchers uploaded the students' written works in a supported format such as Word document (.docx) or PDF. Second, the file uploaded was accompanied with a prompt asking the model to act like an EFL teacher who has to assess students' writing essays on the given rubric while also correcting the mistakes, and providing instructions that serve as formative recommendations. In addition, the prompt instructed the model to give alternatives to the mistakes and errors in the essay then ask it to re-write the corrected essay based on the feedback given. Furthermore, the prompt included a command to track the progress in performance in every area in the rubric after inserting the previous essay. After that, the researchers made sure the model gave an optimized assessment where students not only realize their mistakes and errors but also get a corrected version which familiarizes them with correct structures.

The rubric included the following areas: structure and organization; grammar and mechanics, vocabulary and diction, content and development. Each criterion is evaluated on a 4-point scale, with higher points indicating stronger performance in each area. For example, in 'Structure and Organization', a clear, logical structure, and seamless transitions with strong introduction, body, and conclusion would earn the highest score of 4 points, while a piece lacking clear organization or logical progression would receive a lower score. Similarly, in 'Content and Development', an essay with fully developed ideas; insightful, relevant, and well-supported arguments would merit the highest score, whereas content that deviates or lacks relevance to the intended audience would score lower. This pattern applies across all criteria. The total score is the sum of points across all criteria, providing an overall assessment of the written content.

The model applied the rubric consistently, providing both scores and formative feedback. The same evaluation format was used for all eight essays per student, allowing for tracking of progress. To ensure reliability, the researchers cross-checked the samples of the AI-generated feedback against human judgment to confirm alignment with the rubric standards.

After receiving the feedback, the researchers sent the feedback to the students. In this stage, researchers ensure that students receive summative as well as formative assessment where learners's attention is drawn to the areas where they progressed while also highlighting the areas that need more focus. At every time the students submit their essays, the process was repeated.

The survey that was delivered to students aimed at understanding their perceptions on the process while exploring the extent to which this formative assessment helped students' progress with their writing proficiency. It is worth mentioning that the survey was not piloted. However, the researchers ensured that it was clear and logically structured, although this may have limited the opportunity to refine some questions.

4.4 Ethical Consideration

In the original design, students were not informed that their writing would be assessed by an AI tool to preserve the authenticity of their writing and the survey was not formally approved by an ethics committee due to the researchers' limited awareness at the time. Upon review, this omission raises ethical concerns regarding transparency and informed consent. In future studies, full disclosure regarding the use of AI in assessment should be ensured, and approval from an appropriate ethics committee should be obtained.

4.5 Data Collection

In the experiment, students were required to submit their essays each week all along the eight weeks to come. Personalized feedback on students' essays escorted with an account of their progress is sent to students via e-mail.

The researchers used an analytic rubric where each aspect of performance is rated separately. Analytic scoring provides a clearer image about student's performance on different aspects of the language. One lacuna associated with analytic rubrics is that they may run the risk of over-assessment that is both tiring and time-

consuming especially when assessing portfolios for large classes. The approach to developing this analytic rubric adheres to two principles. First, the results generated will get both the numerical and verbal aspects, in order not to make the process of scoring an end by itself. Instead, the nature of portfolio fosters learning by providing additional remarks when needed. Second, analytic rubric presumably produces high level of reliability as agreed by a number of scholars. (E.g. East, 2009; Jonsson & Svingby, 2007; Uzun et al., 2019; Yaqub & Tabassum, 2016)

The rubric has five criteria (the systematic coherence, relevance and appropriateness, grammar accuracy and spelling, range of vocabulary, and formats and Punctuation). The Systematic Coherence criterion evaluates the organization and flow of the written content, considering the clarity of structure, progression of ideas, and logical sequencing. A higher score indicates a well-organized and coherent piece with smooth transitions between sections. Relevance and Appropriateness assesses the content's alignment with the topic, purpose, and intended audience, ensuring it meets research objectives or communication goals effectively. Grammar Accuracy and Spelling focuses on the correctness of grammar and spelling, rewarding precision and adherence to language conventions. Range of Vocabulary evaluates the diversity and appropriateness of words used, favouring rich, varied language over repetition or simplicity. Formats and Punctuation assesses adherence to formatting guidelines and accurate use of punctuation marks, with higher scores for precise formatting and punctuation.

In the second part of the study, the data were collected through an online survey administered to participants via SurveyMonkey. The survey was designed to collect both qualitative and quantitative data and gather feedback from participants regarding their experience with the intervention and its impact on their writing proficiency. Both open-ended and closed-ended questions were included in the study.

Closed-ended questions provided structured response options, allowing for quantitative analysis and statistical comparisons. Strategic integration of open-ended questions enabled the qualitative viewpoints of the participants to be captured, enabling a deeper understanding of their experiences. In order to address several facets of the study questions, the survey was divided into multiple sections. These areas included participants' perceptions of the effectiveness of the intervention, their satisfaction with the provided feedback, and their overall writing improvement.

The questionnaire began with demographic questions to collect information on participants' age and gender. Next, participants were presented with Likert scale items to assess their agreement with statements related to the intervention's impact on their writing skills. The purpose of these measures was to investigate the extent at which the participants felt the intervention was effective, how beneficial they thought the feedback was, and how satisfied they were in general.

In order to get detailed feedback and comments on their experience with the intervention, participants were also asked to reply to open-ended questions. Through these questions, participants were able to provide their opinions, recommendations, and any challenges they had during the intervention.

The online survey was administered using SurveyMonkey, which facilitated data collection and management. Participants were invited to complete the survey via email invitations and were provided with instructions on how to access and submit their responses securely. Participation in the survey was voluntary, and participants were assured of anonymity and confidentiality in their responses.

Upon completion of the data collection period, the responses were exported from the survey platform into a spreadsheet for analysis. Participants' identities were anonymous to ensure confidentiality, and the data were stored securely in accordance with ethical guidelines.

4.6 Data Analysis

To address the first and the second research questions concerning the impact of receiving formative feedback from ChatGPT-4 on students' writing improvement and the analysis of the progression of writing proficiency across a variety of essays, this study conducted a one-way repeated-measures ANOVA. The ANOVA was performed separately for each student to assess the variability in writing proficiency scores within

individuals over time. Each student's writing proficiency scores for the eight essays were treated as repeated measures, with the essay number serving as the independent variable:

$$(\sigma^2) = \Sigma (x_i - \mu)^2 / N$$

Where:

- σ^2 is the variance,
- x_i represents each individual score,
- μ is the mean of all the scores,
- Σ denotes the summation of values, and
- N is the total number of scores.

The formula used to calculate the variance (σ^2) of the writing proficiency scores is derived from the sum of squared deviations from the mean (μ) of the scores. Each deviation ($x_i - \mu$) is squared and summed across all scores (Σ), then divided by the total number of scores (N). This formula quantifies the average squared distance of each individual score from the mean, providing a measure of the spread or dispersion of the data points around the mean value.

While initial descriptive statistics indicated a mean variance of approximately 1.3061, suggesting moderate variability in student performance, the primary statistical test focused on evaluating whether changes in proficiency over time were statistically significant. The ANOVA revealed a significant effect of essay number on writing proficiency, with results showing:

- $F(7, 343) = 5.82$,
- $p < 0.001$,
- $\eta^2 = 0.106$ (indicating a small-to-moderate effect size).

This result allows us to reject the null hypothesis, which posited no significant difference in writing proficiency across the eight essays. In contrast, the alternative hypothesis—that students' scores would change significantly over time was supported.

Post-hoc pairwise comparisons using the Bonferroni correction further identified that the most substantial improvement occurred between Essay 1 and Essay 5 ($p < 0.01$), and between Essay 5 and Essay 8 ($p < 0.05$). These findings suggest a steady and statistically meaningful progression in writing proficiency as students continued to receive feedback and revise their work.

While variance calculations offered a preliminary understanding of score dispersion, the ANOVA results provided a more conclusive account of trends in writing development. The observed statistical significance supports the interpretation that students' writing improved meaningfully over the duration of the study.

These quantitative findings were reinforced by the survey data. The survey results revealed valuable insights into students' perceptions of the feedback provided on their essays. A total of fifty students participated in the survey, with responses varying across different aspects of the feedback process. A majority of students (30 out of 50) reported being satisfied with the feedback, while 15 students expressed being very satisfied. However, a smaller proportion of students (5 out of 50) remained neutral.

When asked about the helpfulness of feedback in identifying errors, students generally perceived it as helpful, with the majority of students (35 out of 50) finding it very helpful, and an additional 11 students considering it extremely helpful. Only a few students (4 out of 50) found it slightly or moderately helpful. Additionally, the feedback was reported to have a significant impact on writing skills improvement, as indicated

by 36 students. However, a smaller number of students (9 out of 50) rated the feedback as extremely helpful in this regard.

Concerning the clarity of feedback, most students (40 out of 50) found the feedback to be clear, while 8 students rated it as very clear. Only 2 students expressed neutrality towards the clarity of the feedback. Moreover, a majority of students (25 out of 50) agreed that the feedback accurately assessed their essays, while 20 students strongly agreed. However, 5 students remained neutral in their response.

The majority of students (40 out of 50) indicated a likelihood of continuing to use the feedback process for future writing assignments. Among them, 21 students expressed being very likely to continue, while 19 students rated themselves as likely to continue. The feedback progression over multiple essays was perceived as helpful by a significant number of students (43 out of 50), with 19 students finding it extremely helpful.

Some students identified specific areas for improvement in the feedback, including grammar, lexis, and creativity. However, the majority of students (38 out of 50) did not specify any areas for improvement. The feedback process received positive ratings generally, with 23 students rating it as good and 21 students rating it as excellent. Only a small number of students (6 out of 50) remained neutral in their overall experience. Finally, Seven students suggested that the feedback process could be improved by providing longer feedback time, while the majority of students (43 out of 50) did not offer any additional comments or suggestions.

Generally, the findings suggest a positive reception of the feedback, with a majority of students expressing satisfaction and finding it helpful in improving their writing skills. The clarity and accuracy of the feedback were also positively rated by most participants, indicating the effectiveness of the feedback in providing actionable insights into their writing performance.

Furthermore, the high likelihood of students to continue using the feedback process for future assignments underscores its perceived value and importance in their writing development. The feedback progression across multiple essays was also perceived as beneficial by a significant number of students, highlighting the importance of continuity and consistency in feedback delivery. However, some students identified specific areas for improvement in the feedback, particularly in grammar, lexis, and creativity. Addressing these areas could further enhance the effectiveness of the feedback process and contribute to students' writing proficiency.

4.7 Discussion

Using one-way repeated-measures ANOVA, the analysis revealed statistically significant differences in writing proficiency scores across the essays, demonstrating clear evidence of student improvement over time. The ANOVA findings showed that variance within individual students' scores was not random; rather, patterns emerged that suggested measurable growth in writing skills across the essay sequence. This progression indicates that the feedback provided by ChatGPT-4 played a role in facilitating reflective writing practices, error correction, and linguistic improvement.

Importantly, these quantitative results are supported by students' survey responses. Most students reported being satisfied or very satisfied with the feedback received, particularly appreciating its clarity and usefulness in identifying errors. These perceptions align with the observed upward trend in scores, suggesting a strong correlation between student engagement with the feedback process and measurable writing improvement. Notably, the progression across multiple essays was rated as helpful or extremely helpful by the majority of students, reinforcing the value of consistent and timely feedback.

These findings are consistent with existing literature on the role of Large Language Models like ChatGPT-4 in improving writing through formative assessment which emphasizes the importance of personalized, timely, targeted, and actionable feedback provided by AI tools in fostering students' writing development (E.g., Escalante et al., 2023; Jiang et al., 2023; Khasawneh et al., 2025; Lu, 2019; Mizumoto & Eguchi, 2023; Wang, 2020; Xia et al., 2024).

Moreover, while the variance analysis highlighted the statistical significance of differences in scores, its pedagogical value lies in identifying nuanced trends, such as plateaus or sudden improvements. These micro-patterns may reflect the effectiveness of feedback at certain stages or reveal areas where students struggled to

implement suggestions. For instance, though many students improved overall, some stagnation in creative expression was noted. This points to a potential limitation of AI-generated feedback: its relatively weaker support for fostering originality. Though students exhibited noticeable improvement in surface-level structures as grammar and structure, formative feedback on the aspect of creativity is still regarded as an important lacuna of AI tools. This aligns with Chakrabarty et al.'s (2024) study which proved that Large Language Models including ChatGPT-4 fell short of "assessing objectively the creativity of a piece of writing" based on the Torrance Test of Creative Thinking. Further, the Large Language Models- generated assessments differed significantly from the ones provided by human experts.

The study also found that students were highly likely to continue using such feedback tools in the future, highlighting the growing trust and perceived utility of AI in academic settings. This can signal a new shift from traditional assessment into AI assessment. Similarly, studies (E.g., Perkins, 2023; Xia et al. 2024) recommend that the assessment methods in higher education must be reviewed to account for the implications of Generative AI as the traditional assessment approach seems to become increasingly inadequate for giving accurate assessments. However, this growing reliance raises questions about the balance between automated assistance and human instruction. While AI can augment the feedback process, its role should be viewed as complementary to teacher guidance, especially when it comes to assessing deeper rhetorical strategies, creativity, or content development. Hence, educators and practitioners should benefit from the powerful capabilities of AI in assessing students' writing especially in areas such as grammar, mechanics, and structure while avoiding AI in assessing creativity. Then, human assessors should take in charge the aspect of creativity.

Despite its promising results, this study is not without limitations. The sample was limited to 50 students, all within the same academic context, which may limit the generalizability of the findings and reduce statistical power. In addition, concealing the fact that assessment was done via an AI tool may not be ethically permissible which further limits the scope of the study. Another notable limitation of this study is the lack of verifiable intra-rater reliability as the assessment output may lack consistency where ChatGPT-4's assessment cannot inherently be fully guaranteed because the output depends on several factors: the phrasing of the prompt, the context of the session, timing, etc. The survey results, while informative, are also self-reported and may be influenced by social desirability or recall bias.

Future research could expand this work by comparing AI-assisted feedback with traditional teacher feedback or peer feedback across different student populations. Further studies could address the AI's output consistency through repeated assessments to better understand its reliability as a formative assessment tool. Moreover, a qualitative analysis of essay content—beyond scores—could reveal how specific writing features evolve in response to feedback. Research could also explore the integration of AI feedback with human instruction in hybrid models to support not just grammatical accuracy but also argument content development, coherence, and creativity.

Conclusions

The emergence of ChatGPT, an advanced language model powered by artificial intelligence, can be considered as a giant leap towards education development. Thanks to its extraordinary capacity for comprehending and producing writing that is human-like, ChatGPT-4 has the potential to revolutionize how portfolios are evaluated and assessed. ChatGPT-4 can provide personalized feedback, insightful suggestions, and critical analysis of portfolio contents. Its ability to comprehend context, adapt to various domains, and generate coherent responses makes it an adequate tool for evaluating the depth and quality of portfolios. Moreover, ChatGPT's deep learning capabilities enable it to identify patterns, highlight strengths and weaknesses, and offer tailored recommendations for portfolio improvement. Through benefiting from the potentials of ChatGPT-4, portfolio assessment can become a more user-friendly tool that enables students to receive comprehensive feedback and enhance their understanding of their performance and progress. Artificial intelligence technology and portfolio assessment together have the potential to revolutionize the teaching and

learning processes through providing more accurate and comprehensive methods of assessing students' capacities while softening the workload associated with portfolio assessment.

References

- Abdul Halim, I., Hartati, A., & Wiji Lestari, Z. (2019). The Use of Portfolio Assessment of Writing Skill in Descriptive Text. *Journal of Applied Linguistics and Literacy*, *3(2)*(75). https://doi.org/10.25157/jall.v3i2.2420
- Alawida, M., Mejri, S., Mehmood, A., Chikhaoui, B., & Isaac Abiodun, O. (2023). A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity. *Information*, *14*(8), Article 8. https://doi.org/10.3390/info14080462
- Alhalangy, Abdalilah. G. I., & AbdAlgane, M. (2023). Exploring the Impact of AI on The EFL Context: A Case Study of Saudi Universities. *Journal of Intercultural Communication*, *23*(2), 41–49. https://doi.org/10.36923/jicc.v23i2.125
- Bedda, N., & Benabdi, F. (2022). Hinderers To The Use Of Portfolio In The Assessment Of Students' Writing. Case Study: Tertiary Teacher At Mustapha Stambouli University. *Revue Algérienne Des Manuscrits*, *18*(1). https://www.theses-algerie.com
- Birgin, O., & Baki, A. (2007). The Use of Portfolio to Assess Student's Performance. *Journal of TURKISH SCIENCE EDUCATION*, *4*(2), 75–90.
- Bose, J., & Rengel, Z. (2009). A model formative assessment strategy to promote student-centered self-regulated learning in higher education. *US-China Education Review*.
- Bouakel, T. (2021). E-portfolio Assessment between the Practical Difficulties and the FutureProspects. *Literature and Philosophy*, *13*.
- Bouakel, T. (2022). *Electronic Portfolio as an Alternative Form to Assess the Cultural Understanding within EFL Context: Case of 1 st Year EFL Students at the University of Chlef* [Djillali Liabes University, Sidi Bel-Abbès]. http://rdoc.univ-sba.dz/bitstream/123456789/3610/1/DS_Ang_Bouakel_Toufik.pdf
- Boudibi, N.-H., & Beleulmi, S. (2024). Exploring Algerian EFL University Instructors' Understanding and Practices of Learning-Oriented Assessment. *ATRAS Journal*, *5*(2), 205–219. https://doi.org/10.70091/Atras/vol05no2.16
- Boumediene, H., Berrahal, F. K., & Hardji, M. B. (2018). Using portfolio assessment to enhance reflective writing in EFL classrooms. *Literacy Information and Computer Education Journal(LICEJ)*, *9*(3), 2994–3001.
- Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., & Wu, C.-S. (2024). *Art or Artifice? Large Language Models and the False Promise of Creativity* (arXiv:2309.14556). arXiv. https://doi.org/10.48550/arXiv.2309.14556
- Christiano, P. (2024, August 10). *ChatGPT Unleashed: Pioneering A New Era Of Conversational AI - ExpertBeacon*. Expert Beacon. https://expertbeacon.com/chatgpt-unleashed-pioneering-a-new-era-of-conversational-ai/
- Cole, K. B., Struyk, L. R., Kinder, D., Sheehan, J. K., & Kish, C. K. (1997). Portfolio Assessment: Challenges in Secondary Education. *The High School Journal*, *80*(4), 261–272.
- Dang, A., & Wang, H. (2024). Ethical use of generative AI for writing practices: Addressing linguistically diverse students in U.S. Universities' AI statements. *Journal of Second Language Writing*, *66*, 101157. https://doi.org/10.1016/j.jslw.2024.101157
- Daud, A. (2017). Challenges of Using Portfolio Assessment as an Alternative Assessment Method for Teaching English in Indonesian Schools. *International Journal of Educational Best Practices*, *1*, 10–114. https://doi.org/10.31258/ijebp.v1n2.p106-114

- Davis, M. H., & Ponnamperuma, G. G. (2005). Portfolio Assessment. *Journal of Veterinary Medical Education*, *32*(3), 279–284. https://doi.org/10.3138/jvme.32.3.279
- East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, *14*(2), 88–115. https://doi.org/10.1016/j.asw.2009.04.001
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), 57. https://doi.org/10.1186/s41239-023-00425-2
- Frey, B. B. (2018). *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781506326139
- Hammad Al-Rashidi, A., Vadivel, B., Ramadan Khalil, N., & Basim, N. (2023). The comparative impacts of portfolio-based assessment, self-assessment, and scaffolded peer assessment on reading comprehension, vocabulary learning, and grammatical accuracy: Insights from working memory capacity. *Language Testing in Asia*, *13*(1), 24. https://doi.org/10.1186/s40468-023-00237-1
- Herman, J. L., Gearhart, M., & Baker, E. L. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment*, *1*(3), 201–224. https://doi.org/10.1207/s15326977ea0103_2
- Holmes, W., Persson, J., Chounta, I.-A., Wasson, B., & Dimitrova, V. (with Europarat). (2022). *Artificial intelligence and education: A critical view through the lens of human rights, democracy and the rule of law*. Council of Europe.
- Hudori, R. F. A., Tasnim, Z., Fardhani, A. E., & Sari, D. P. (2020). The Use of Portfolio Assessment in English Writing Classes. *IOP Conference Series: Earth and Environmental Science*, *485*(1), 012093. https://doi.org/10.1088/1755-1315/485/1/012093
- Jiang, R. (2022). How does artificial intelligence empower EFL teaching and learning nowadays? A review on artificial intelligence in the EFL context. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.1049401
- Jiang, Z., Xu, Z., Pan, Z., He, J., & Xie, K. (2023). Exploring the Role of Artificial Intelligence in Facilitating Assessment of Writing Performance in Second Language Learning. *Languages*, *8*(4), Article 4. https://doi.org/10.3390/languages8040247
- Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E., Scoville, E., Reese, T., Friedman, D., Bastarache, J., van der Heijden, Y., Wright, J., Carter, N., Alexander, M., Choe, J., … Wheless, L. (2023). Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Research Square*, rs.3.rs-2566942. https://doi.org/10.21203/rs.3.rs-2566942/v1
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002
- Khasawneh, M. A. S., Aladini, A., Assi, S. A., & Ajanil, B. (2025). Portfolio assessment in AI-enhanced learning environments: A pathway to emotion regulation, mindfulness, and language learning attitudes. *Language Testing in Asia*, *15*(1), 5. https://doi.org/10.1186/s40468-025-00345-0
- Kleppen, E. (2022, June 2). *How to Find Variance Using Python*. Built In. https://builtin.com/data-science/how-to-find-the-variance
- Lam, R., & Lee, I. (2010). Balancing the dual functions of portfolio assessment. *ELT Journal*, *64*(1), 54–64. https://doi.org/10.1093/elt/ccp024
- Lo, Y.-F. (2010). Implementing Reflective Portfolios for Promoting Autonomous Learning among EFL College Students in Taiwan. *Language Teaching Research*, *14*(1), 77–95. https://doi.org/10.1177/1362168809346509
- Lu, X. (2019). An Empirical Study on the Artificial Intelligence Writing Evaluation System in China CET. *Big Data*, *7*(2), 121–129. https://doi.org/10.1089/big.2018.0151

- Lubbad, M. (2023, October 6). The Ultimate Guide to GPT-4 Parameters: Everything You Need to Know about NLP's Game-Changer. *Medium*. https://medium.com/@mlubbad/the-ultimate-guide-to-gpt-4-parameters-everything-you-need-to-know-about-nlps-game-changer-109b8767855a
- Mahmud, F. (2023). Investigating EFL Students' Writing Skills Through Artificial Intelligence: Wordtune Application as a Tool. *Journal of Language Teaching and Research*, *14*, 1395–1404. https://doi.org/10.17507/jltr.1405.28
- Mastrogiacomi, F. (2023, July 10). *Future Proofing a ChatGPT-Proof Portfolio Evidence-Based Formative Assessment*. Conference: Conference: EdMedia + Innovate Learning Summit, 2023, Wien.
- Meeus, W., Van Petegem, P., & Engels, N. (2009). Validity and reliability of portfolio assessment in pre-service teacher education. *Assessment & Evaluation in Higher Education*, *34*(4), 401–413. https://doi.org/10.1080/02602930802062659
- Milmo, D. (2023, February 2). ChatGPT reaches 100 million users two months after launch. *The Guardian*. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app
- 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
- Nezakatgoo, B. (2011). The Effects of Portfolio Assessment on Writing of EFL Students. *English Language Teaching*, *4*(2), 231. https://doi.org/10.5539/elt.v4n2p231
- Office of Global Insight and Policy. (2021). *Policy Guidance on AI for Children*. United Nations Children's Fund (UNICEF). https://www.unicef.org/innocenti/media/1341/file/UNICEF-Global-Insight-policy-guidance-AI-children-2.0-2021.pdf
- Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*, *20*(2). https://doi.org/10.53761/1.20.02.07
- Perrigo, B. (2022, December 5). *AI Chatbots Are Getting Better. But an Interview With ChatGPT Reveals Their Limits*. TIME. https://time.com/6238781/chatbot-chatgpt-ai-interview/
- Perronoski Fuller, L., & Bixby, C. (2024). The Theoretical and Practical Implications of OpenAI System Rubric Assessment and Feedback on Higher Education Written Assignments. *American Journal of Educational Research*, *12*(4), 147–158. https://doi.org/10.12691/education-12-4-4
- Phung, H. V. (2016). Portfolio Assessment in Second/Foreign Language Pedagogy. *Hawaii Pacific University TESOL Working Paper Series*, *14*, 90–107.
- Popescu-Mitroia, M.-M., Todorescu, L.-L., & Greculescu, A. (2015). The Usefulness of Portfolios as Assessment Tools in Higher Education. *Procedia - Social and Behavioral Sciences*, *191*, 2645–2649. https://doi.org/10.1016/j.sbspro.2015.04.722
- Rane, N. (2024). Enhancing the quality of teaching and learning through ChatGPT and similar large language models: Challenges, future prospects, and ethical considerations in education. *TESOL and Technology Studies*, *5*(1), 1–6. https://doi.org/10.48185/tts.v5i1.1000
- Rettberg, J. W. (2022, December 6). ChatGPT is multilingual but monocultural, and it's learning your values. *Jill/Txt*. https://jilltxt.net/right-now-chatgpt-is-multilingual-but-monocultural-but-its-learning-your-values/
- Revell, T., Yeadon, W., Cahilly-Bretzin, G., Clarke, I., Manning, G., Jones, J., Mulley, C., Pascual, R. J., Bradley, N., Thomas, D., & Leneghan, F. (2024). ChatGPT versus human essayists: An exploration of the impact of artificial intelligence for authorship and academic integrity in the humanities. *International Journal for Educational Integrity*, *20*(1), Article 1. https://doi.org/10.1007/s40979-024-00161-8

- Rice, C., Taylor, J. B., Widom, J., Zegart, A., & Lin, H. S. (2023). *The Stanford Emerging Technology Review 2023: A Report on Ten Key Technologies and Their Policy Implications*. Stanford University. https://setr.stanford.edu/sites/default/files/2023-11/SETR_web_01_AI.pdf
- Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th edition). Pearson.
- Schmidt, T., & Strasser, T. (2022). Artificial Intelligence in Foreign Language Learning and Teaching. *Anglistik*, *33*(1), 165–184. https://doi.org/10.33675/ANGL/2022/1/14
- Schreiner, M. (2023, July 11). *GPT-4 architecture, datasets, costs and more leaked*. THE DECODER. https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/
- Sumakul, D. T. Y. G., Hamied, F. A., & Sukyadi, D. (2022). Artificial Intelligence in EFL Classrooms: Friend or Foe? *LEARN Journal: Language Education and Acquisition Research Network*, *15*(1), Article 1.
- Tangdhanakanond, K., & Wongwanich, S. (2015). State, Problems and Guidelines for Solving Problems in Implementing Student Portfolio Assessment in Elementary Schools in Thailand. *Procedia - Social and Behavioral Sciences*, *171*, 1381–1387. https://doi.org/10.1016/j.sbspro.2015.01.257
- Tisani, N. (2008). Challenges in Producing a Portfolio for Assessment: In Search of Underpinning Educational Theories. *Teaching in Higher Education*, *13*(5), 549–557. https://doi.org/10.1080/13562510802334830
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, *10*(1), 15. https://doi.org/10.1186/s40561-023-00237-x
- Uzun, N. B., Alici, D., & Aktas, M. (2019). Reliability of the Analytic Rubric and Checklist for the Assessment of Story Writing Skills: G and Decision Study in Generalizability Theory. *European Journal of Educational Research*, *8*(1), 169–180.
- Wang, Z. (2020). Computer-assisted EFL writing and evaluations based on artificial intelligence: A case from a college reading and writing course. *Library Hi Tech*, *40*(1), 80–97. https://doi.org/10.1108/LHT-05-2020-0113
- Weinberg, J. (2020, July 30). *Philosophers On GPT-3 (updated with replies by GPT-3)—Daily Nous*. https://dailynous.com/2020/07/30/philosophers-gpt-3/, https://dailynous.com/2020/07/30/philosophers-gpt-3/
- Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education*, *21*(1), 40. https://doi.org/10.1186/s41239-024-00468-z
- Yanhua, Z. (2020). The Application of Artificial Intelligence in Foreign Language Teaching. *2020 International Conference on Artificial Intelligence and Education (ICAIE)*, 40–42. https://doi.org/10.1109/ICAIE50891.2020.00017
- Yaqub, H., & Tabassum, R. (2016). *INTRA-RATER RELIABILITY OF HOLISTIC AND RUBRIC-BASED ASSESSMENT OF ESSAY WRITING IN PAKISTAN*.

Appendix 1

Criteria	Excellent (4 pts)	Good (3 pts)	Fair (2 pts)	Weak (1 pt)	Poor (0 pt)
Structure and Organization	Clear, logical structure with strong introduction, body, and conclusion; seamless transitions.	. Well-structured with minor lapses in flow or transitions.	Basic structure with some confusion in organization or progression.	Weak structure; ideas are poorly organized.	No structure or logical flow
Grammar and Mechanics	Free of grammatical, spelling, and punctuation errors.	Minor errors that do not affect understanding.	Noticeable errors that occasionally disrupt clarity.	Frequent errors that interfere with meaning.	Pervasive errors that prevent understanding.
Vocabulary and Diction	Rich, precise, and varied vocabulary; accurate word choice throughout.	Appropriate vocabulary with some variety.	Basic or repetitive vocabulary; some misuse of words.	Limited or inaccurate vocabulary; affects clarity.	Inappropriate or severely limited vocabulary.
Content and Development	Fully developed ideas; insightful, relevant, and well-supported arguments.	Clear and relevant ideas with some depth.	Ideas are somewhat relevant but lack depth or clarity.	Underdeveloped or unclear ideas; minimal relevance to topic.	Content is irrelevant, unclear, or missing.

Appendix 2

Guidelines of Writing the Essay:

1- Structure: Organize your essay with an introduction, body paragraphs, and a conclusion. Include a clear thesis statement that outlines your main argument or point of view.

2- Clarity and Cohesion: Ensure your ideas flow logically, using appropriate transitions between paragraphs. Use clear and concise language to convey your thoughts effectively.

3- Originality: Your essay should reflect your own thoughts and perspectives. Avoid plagiarism by properly citing any external sources used.

4- Proofreading: Before submission, carefully review and edit your essay for grammar, spelling, and punctuation errors. Ensure your writing is polished and coherent.