

Name: Dr. Alaa Alnajashi
King Abdullaziz University

Investigating the Accuracy of Large Language Models 'ChatGPT-4' in Grading Students' Writing According to a Specific Rubric

Abstract:

The advancement of artificial intelligence (AI) is rapidly transforming various aspects of our lives. Notably, the development of chatbots, such as ChatGPT-4, has introduced new prospects for language teachers by serving as grading tools, potentially saving hundreds of hours of teacher time and ensuring a higher rate of accuracy. This study evaluates ChatGPT-4's accuracy in grading student assignments against a rubric and compares its results with those of human raters. Data were sourced from a high-stakes, end-of-year exam marked by human teachers at the university level. The assignments were inputted into ChatGPT-4 along with the rubric, which was then prompted to grade the 40 paragraphs according to the same rubric. Subsequently, a precision test was conducted to evaluate the accuracy of the machine's grading. The results of the precision test showed a high level of accuracy, showcasing the value of ChatGPT-4 in marking assignments according to a rubric. Educational institutions could benefit from this innovative and easy-to-use feature, which can perform the grading task swiftly and with a high degree of precision.

Investigating the Accuracy of Large Language Models 'ChatGPT-4' in Grading Students' Writing According to a Specific Rubric

Abstract:

The advancement of artificial intelligence (AI) is rapidly transforming various aspects of our lives. Notably, the development of chatbots, such as ChatGPT-4, has introduced new prospects for language teachers by serving as grading tools, potentially saving hundreds of hours of teacher time and ensuring a higher rate of accuracy. This study evaluates ChatGPT-4's accuracy in grading student assignments against a rubric and compares its results with those of human raters. Data were sourced from a high-stakes, end-of-year exam marked by human teachers at the university level. The assignments were inputted into ChatGPT-4 along with the rubric, which was then prompted to grade the 40 paragraphs according to the same rubric. Subsequently, a precision test was conducted to evaluate the accuracy of the machine's grading. The results of the precision test showed a high level of accuracy, showcasing the value of ChatGPT-4 in marking assignments according to a rubric. Educational institutions could benefit from this innovative and easy-to-use feature, which can perform the grading task swiftly and with a high degree of precision.

Keywords · ChatGPT -4 , Automated chatbots scoring, Human raters , AI rater.

Introduction:

The groundbreaking revolution in Artificial Intelligence (AI) has sent ripples across various sectors, and English Language Teaching (ELT) is no exception. ELT professionals worldwide are increasingly recognizing the potential of AI to transform the way English is taught and assessed. Furthermore, AI can also assist in automating mundane administrative tasks. Tasks such as grading assignments allowing teachers to spend more time focusing on actual teaching and one-on-one student interactions. But, the question is how accurate is large language models such as Chat GPT4 in grading students writing according to a specific question. This study aims to answer the following research questions:

1. How accurate is ChatGPT-4 in grading students' writing according to a specific rubric?
2. What is the precision rate of ChatGPT-4 in grading students' writing according to a rubric compared to human raters?

Literature Review:

Artificial intelligence has been recognized for its potential strength in personalized learning through the evaluation and feedback of essay writing (Chiu et al., 2023; Farrokhnia et al., 2023; Zhu et al., 2023). This capability has the potential to alleviate teacher workload and prevent

burnout (Farrokhnia et al., 2023). Looking at the history of artificial intelligence and machine learning you will notice that the process of automated error correction started in a simple way and involved simple string matching and substitution. Currently, many researchers employed data-driven methods to extract features from the data and use machine learning algorithms to build a model for detecting and correcting errors. These methods are based on machine learning (Huang et al, 2022). Although numerous studies have been conducted on automated error correction systems (AES), many experts have pointed out gaps and shortcomings in this field for various reasons (Huang et al, 2022; Almusharraf & AlOtabi, 2022; Wu, 2023). Dikli and Bleyle (2014) argued that the bulk of AES research concentrates on writing samples from native English speakers linked with major testing organizations. In this study, the data will be drawn from a real high-stakes exam, which can be classified as A2 level according to the Common European Framework of Reference (CEFR).

Numerous studies have examined the accuracy of human raters versus automated raters such as Grammarly and Criterion (Shermis, 2014; Ke, 2019; Dikli & Bleyle, 2014; Almusharraf & AlOtabi, 2022). A few recent studies have compared the differences between ChatGPT and Grammarly in providing corrections for grammar mistakes (e.g., Wu, 2023); however, their focus was on GPT-3.5, not GPT-4. Huang et al. (2022) developed a natural language processing system to correct grammatical errors. However, as far as I know, no study to date has compared the performance of human raters with that of ChatGPT4 when it acts as a rater according to a specific rubric. This study aims to contribute to the evolving field of large language models and language teaching.

Although the specific technical aspects of ChatGPT have not been comprehensively disclosed, it is understood to be based on Instruct GPT (Ouyang et al., 2022). This foundation utilizes instruction tuning (Wei et al., 2022) combined with reinforcement learning from human feedback (Christiano et al., 2017). In a recent paper by (Wu et al, 2023) they mentioned that Chat GPT3.5 performs well with grammar corrections as it doesn't solely focus on rectifying mistakes one after another. Rather, it often opts to alter the surface-level expression of certain phrases or modify the sentence construction. To date, the modified version of Chat GPT-4 has not been widely used in TESOL studies as it is not a free version and requires a subscription. This new model can perform extraordinary functions like reading large files and performing required tasks, including synthesizing, summarizing, or analyzing. Additionally, in the new version released in November 2023, the model can perform all these tasks through a simplified interface or front-end, eliminating the need for users to access the backend or utilize any Application Programming Interface (API). An API key is a code that identifies and authenticates an application or user.

The Language Models (LLMs) connected to ChatGPT are not trained using a corpus from a particular domain like essays; instead, they are trained with text extracted from the Internet. The broad, domain-general nature of these LLMs supporting ChatGPT necessitates thorough research into their effectiveness as Automated Writing Evaluation (AWE) tools before considering their application in such a context (Escalante et al, 2023).

Dai et al (2023) indicate that ChatGPT demonstrates the ability to generate detailed and coherent feedback that effectively summarizes students' performance, surpassing human instructors in this regard. Additionally, the model shows high agreement with instructors when assessing

assignment topics. Furthermore, ChatGPT proves valuable in offering feedback on the process of students completing tasks, contributing to the development of students' learning skills. However, Yoon (2023) asserts that ChatGPT, lacking specialized training for the feedback generation task, proves ineffective in providing constructive feedback on coherence and cohesion for English Language Learner students.

Dataset:

The dataset was extracted from a large high-stake exam for the foundation year at an anonymous university, 40 paragraphs from the final exam in an English language institute of foundation year students, at a proficiency level equivalent to A2 according to the Common European Framework of Reference for Languages (CEFR), were randomly selected for the initial phase of the research. It is pertinent to mention that the grading system in this institute relies on a specific rubric and peer-reviewed grading to ensure accuracy and consistency among teachers in adhering to the rubric.

Methodology:

This study employed a quantitative method, in the first phase 40 paragraphs were uploaded to a Python environment, "PyCharm," alongside the rubric used by human raters. Subsequently, ChatGPT4 was used, via an API key, to assess the paragraphs using this rubric. The evaluation focused on three areas of the rubric: 1) spelling and punctuation, 2) grammar, and 3) content (refer to Appendix A for the rubric details). The cumulative scores for each paragraph were computed. The program was configured to prompt ChatGPT to bypass any paragraphs it could not assess, resulting in 10 skipped paragraphs (refer to Appendix B for a screenshot of PyCharm environment for Python). While trying to solve this issue, a new update to ChatGPT-4 was introduced that enabled direct file uploads to ChatGPT-4 interface so the issue was resolved with this new update. As I uploaded the rubric file and the paragraph file into the newest update of ChatGPT4 and I asked it to mark it according to the rubric given in a separate file. The model's outputs from the latest update of ChatGPT4 were remarkably precise and swift and it managed to mark all the 40 paragraphs in 5 minutes without skipping any paragraphs (refer to Appendix C for a screenshot of the updated ChatGPT4 interface).

After generating the ChatGPT results, the dataset was downloaded from GPT-4. Then to test for the machine accuracy I used precision scoring which is a common metric to evaluate machine performance in artificial intelligence field (Escalante et al, 2023) as will be thoroughly discussed in the result section.

Results:

To analyze the accuracy of the machine marking we conducted a precision test which is a very established test in the area of machine learning (Escalante et al, 2023). The formula for calculating precision is as follows:

$$\text{Precision} = \frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Positives}}$$

If the AI's marking matches the human rater's marking (or is very close, within one grade of error), we can count it as a true positive value. If the AI's marking deviates significantly from the human rater's marking, it would be considered a false positive. With a margin of ± 5 , the calculated precision is approximately 0.675, or 67.5%. This means that with this margin, 67.5% of the AI's markings align with the human standard and are considered correct (True Positives), while 32.5% are incorrect (False Positives).

Discussion:

The aim of this study is to evaluate the accuracy of ChatGPT-4 in grading assignments that have already been graded by human graders for high-stakes examinations and to assess the machine's precision in comparison to human grading. Initially, some system errors related to the backend, Python, and the OpenAI key were encountered as the system was skipping many long paragraphs. While I was trying to resolve the issue, a new version of ChatGPT-4 was introduced in November 2023, known as the "plus edition". These issues were resolved with this update, as it has shown the ability to quickly analyze multiple files via ChatGPT-4's frontend (OpenAI, 2023). I uploaded the grading rubric and the assignments into the new edition of ChatGPT-4 and provided a prompt to grade the assignments according to the rubric. The results, as detailed in the results section, demonstrate that ChatGPT-4 achieved a high precision rate (more than 60%) with human tutors and completed the task in minutes, highlighting the efficiency of deep learning where a single model can process vast amounts of data swiftly.

The significance of this study lies in demonstrating the potential of large language models in grading assignments according to a rubric, marking a significant advancement for educators, particularly English teachers. Large language models may demonstrate lower bias and higher accuracy than humans. Dai et al. (2023) report that ChatGPT is capable of generating detailed feedback that effectively summarizes student performance, outperforming human instructors in this regard. The study confirms that ChatGPT-4 can grade assignments based on a specified rubric. However, Yoon (2023) cautions that ChatGPT4 may not provide accurate feedback on cohesion and coherence.

Despite limited research in this field, the findings of this study, along with the few others mentioned in the literature section above, are vital in this rapidly evolving field. They encourage stakeholders and teachers to adopt the innovative features of language learning models like GPT-4. Although GPT-4 is not free, it offers valuable functionalities that can streamline the grading process and enhance accuracy, as our analysis has demonstrated. Such advancements could significantly reduce teachers' workloads and ensure precision in grading.

Limitations and Future Research Directions :

This study has a few limitations; the dataset contains grades without any feedback. Future studies could focus on comparing both human grades and feedback with AI ratings and feedback. Nonetheless, this study is the first step towards exploring the capabilities of large language models in acting as language graders and we still need many studies in this new area.

Conclusion and pedagogical implications:

This study has investigated the applicability of large language models, such as GPT-4, in marking student writing exams according to a rubric. The results have shown a high precision rate compared to human raters. Pedagogically, this study is significant as it highlights the potential of large language models like GPT-4 in reducing teachers' workload and providing systematic and unbiased ratings according to the stated rubric. Therefore, it is crucial to utilize these new tools within the educational system, especially as they continue to evolve. According to a recent event on OpenAI's website announcing new updates (Openai, 2023), they will introduce custom GPT models for specific purposes. Institutions might use these advancements to design their own custom GPT models that can rate and provide feedback on student writing or work. More research is needed in this area, and this study has taken the first step.

References:

- Almusharraf, N., & AlOtaibi, H. (2022). An error-analysis study from an EFL writing context: Human and automated essay scoring approaches. *Technology, Knowledge and Learning*. <https://doi.org/10.1007/s10758-022-09592-z>.
- Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education Artificial Intelligence*. <https://doi.org/10.1016/j.caeai.2022.100118>
- Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. <https://doi.org/10.48550/arXiv.1706.03741>
- Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y., Gasevic, D., & Chen, G. (2023, April 13). Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. <https://doi.org/10.35542/osf.io/hcgzi>
- Dikli, S., & Bleyl, 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.sdl.idm.oclc.org/10.1016/j.asw.2014.03.006>
- 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, 57. <https://doi.org/10.1186/s41239-023-00425-2>
- Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. <https://doi.org/10.1080/14703297.2023.2195846>.
- Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, 26(2), 5–24.
- Huang, J., & Whipple, P.B. (2023). Rater variability and reliability of constructed response questions in New York state high-stakes tests of English language arts and mathematics: implications for educational assessment policy. *Humanities and Social Sciences Communications*, 10, 860. <https://doi.org/10.1057/s41599-023-02385-4>
- Juan Long. (2022). A Grammatical Error Correction Model for English Essay Words in Colleges Using Natural Language Processing. *Mobile Information Systems*, vol. 2022, Article ID 1881369, 9 pages. <https://doi.org/10.1155/2022/1881369>
- Ke, Z. (2019). Automated essay scoring: A survey of the state of the art [Paper presentation]. *International Joint Conference on Artificial Intelligence 28th Annual Meeting*. <https://doi.org/10.24963/ijcai.2019/879>

Long, O., Wu, J. , Almeida, d., Carroll L Wainwright, Mishkin P., Zhang C. , Sandhini A.I, Katarina S., Alex Ray, et al. (2022). Training language models to follow instructions with human feedback. arXiv.

OpenAI. (2023). GPT4. Retrieved November, 2023, from <https://www.openai.com>

Shermis, M., & Burstein, J. (2003). Automated essay scoring. Routledge.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53–76. <https://doi.org/10.1016/j.asw.2013.04.001>

Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; & Le, Q. V. (2022). Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Wu, H., Wang, W., et al. (2023). ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark. arXiv.org. <https://doi.org/10.48550/arXiv.2303.13648>

Yoon, S. Y., Miszoglud, E., & Pierce, L. R. (2023). Evaluation of ChatGPT Feedback on ELL Writers' Coherence and Cohesion. arXiv preprint arXiv:2310.06505.

Zhu, C., Sun, M., Luo, J., Li, T., & Wang, M. (2023). How to harness the potential of ChatGPT in education? *Knowledge Management & E-Learning*, 15(2), 133–152. <https://doi.org/10.34105/j.kmel.2023.15.008>

Appendix A

Rubric:

1)Content: (40The student wrote **in complete sentences, in the form of a paragraph.**

32 The student wrote **in complete sentences, in the form of a paragraph** and included **at least 7 sentences**

24 The student wrote **in complete sentences**, and included **at least 6 sentences**

16 The student wrote **four sentences**

8 The student include **three complete sentences**

0 The student mostly copied words and phrases from the table with very little attempt to organize them or put them into logical sentences .

2)Grammar:

Key Grammar Points:

- ☐ Correct use of personal and possessive pronouns
- ☐ Correct use of adjective phrases (very + adj, adj and adj, good at +noun, good with +noun, etc).
- ☐ Correct word order (Subject-Verb-Object, Subject-Verb-Adjective, or Subject-Verb-Adverb)
- ☐ Correct use of present simple (including 3rd person singular as needed/appropriate)

15 All simple sentences were written correctly. The student also **successfully attempted** combining some sentences with “and”. There were **no** fragments (incomplete sentences). The “Key Grammar Points” listed above are **used accurately with minimal errors.**

12 Most simple sentences were written correctly. The student **attempted** to combine some sentences with “and”. There **may be a few** fragments (incomplete sentences). The “Key Grammar Points” listed above are **used mostly correctly**.

9 There are **some errors in sentence structure**, even for simple sentences. There may be **some** fragments. The “Key Grammar Points” listed above have **some errors** in their use.

6 Errors in sentence structure **are common**. **Fragments** (incomplete sentences) **are also common**. The “Key Grammar Points” listed above are **used incorrectly more often than they are used correctly**.

3 Errors in sentence structure are **very frequent**. **Fragments** (incomplete sentences) are also **very frequent**. The “Key Grammar Points” listed above are **only occasionally used correctly**.

0 The student made almost no attempt to put the response into sentences and gave almost no evidence of ability to incorporate the “Key Grammar Points” listed above **OR** The writing is **mostly or completely off topic**.

3) Spelling & Punctuation:

Key Points:

Spelling of personal and possessive pronouns, the “be” verb, and high frequency words should be assessed.

Commas are used correctly when a prepositional phrase begins a sentence and in lists.

Apostrophes are used correctly **if** contractions are used (contractions are not required).

Capital letters should be used at the beginning of sentences and for names, countries, cities, and languages.

15 Full stops are used consistently and accurately. There are no run-on sentences and no fragments. The “Key Points” listed above are used accurately with minimal errors.

12 Full stops are usually used accurately. There are few run-on sentences and/or fragments. The “Key Points” listed above are used correctly most of the time.

9 Full stops are used about half of the time. There are some run-on sentences and/or fragments. The “Key Points” listed above have some errors in their use.

6 Full stops are frequently **not** used. There are many run-on sentences and/or fragments. The “Key Points” listed above are used incorrectly more often than they are used correctly.

3 Full stops are rarely used. There may be only 1 or 2 full stops in the entire response. The “Key Points” listed above are only occasionally used correctly.

Appendix B

Screen shot of the code in “Pycharm” environment for Python

```
main.py x
7
8 # Function to read text from a text file
9 usage
10 def read_text_from_file(file_path):
11     with open(file_path, "r", encoding="utf-8") as file:
12         return file.read()
13
14 # Load the rubric from the text file
15 rubric_text = read_text_from_file("/Users/alaanashi/Desktop/rubric.txt")
16
17 # Load the paragraphs from the Excel file
18 df = pd.read_excel("/Users/alaanashi/PycharmProjects/Human_rator/venv/lib/human.xlsx")
19 paragraphs = df["Answer"]
20
21 # Initialize variables to store grades for each criterion
22 spelling_punctuation_grades = []
23 grammar_grades = []
24 content_grades = []
25
26 # Iterate through each paragraph and grade it using GPT-4
27 for paragraph in paragraphs:
28     # Prompt GPT-4 to grade the assignment according to the rubric
29     prompt = f"Grade the assignment according to the rubric:\nRubric: {rubric_text}\nAssignment: {p
30
31     # Call GPT-4 for grading
32     response = openai.Completion.create(
33         engine="text-davinci-003", # GPT-4 engine
34         prompt=prompt,
35         max_tokens=1500, # Adjust this value as needed
```

```

# Extract the grade for each criterion from the response
output_text = response.choices[0].text
grades = output_text.split("\n")

try:
    # Ensure there are grades before attempting to convert
    if len(grades) == 3:
        spelling_punctuation_grade = float(grades[0].strip())
        grammar_grade = float(grades[1].strip())
        content_grade = float(grades[2].strip())

        # Append the grades to the respective lists
        spelling_punctuation_grades.append(spelling_punctuation_grade)
        grammar_grades.append(grammar_grade)
        content_grades.append(content_grade)
    else:
        print(f"Skipping paragraph: {paragraph}")
except ValueError:
    print(f"Error grading paragraph: {paragraph}")

# Append the grades to the respective lists
spelling_punctuation_grades.append(spelling_punctuation_grades)
grammar_grades.append(grammar_grades)
content_grades.append(content_grades)
else:
    print(f"Skipping paragraph: {paragraph}")

```

Appendix C

The recent interface of GPT4 (released in November 2023)

ChatGPT 4 ▾



How can I help you today?

Help me pick
an outfit that will look good on camera


Suggest some codenames
for a project introducing flexible work arrangements

Tell me a fun fact
about the Roman Empire

Write a text message
asking a friend to be my plus-one at a wedding

 rubric.txt
Document

 Paragraphs.xlsx
Spreadsheet

 Grade the paragraphs in the excel sheet -under the paragraphs column- according to the rubric in the rubric file then append the result in the excel sheet. |

