https://doi.org/10.36315/2025v2end013
Education and New Developments 2025

# ARTIFICIAL INTELLIGENCE ASSESSING CONTENT-FOCUSED SHORT ANSWERS

**Juuso Rytilahti, Erkki Kaila, & Erno Lokkila**
*Department of Computing, University of Turku (Finland)*

## Abstract

The capabilities of Artificial Intelligence (AI), and specifically large language models (LLMs) have changed the way teachers work. Using LLM or other AI-assisted tools to help review student submissions has quickly become common practice. These AI-driven automatic assessment tools still have a lot of open questions regarding their effectiveness, performance, and reliability. In this study, we observe the LLMs' capabilities to assess textual answers. The data set used consisted of 31 different computer science-related questions and 2981 answers written in English with detailed feedback and the correct answers. The LLM we used was GPT-4o from OpenAI. At first, the performance of the LLM was tested against a single question present in the data set producing scores for all of its answers (N=82) using multiple different variations of settings. The best-performing approach was then used to process the full data set. With the full set, the model got the exactly correct evaluation in 41,3% of the cases. With an accepted error margin of ±20%, the correctness was 74.7%. When observing the fully correct answers in the set (N=1802), the model was able to correctly identify 51.4% of them. The results can be used to guide future research endeavors in AI-driven automatic assessment research and to guide teachers on how to improve the performance of educational use of LLMs in different ways.

***Keywords:*** *Large language models, automatic assessment, content-focused short answers.*

## 1. Introduction

Large language models (LLM) have gained popularity as teaching-assisting tools. Their performance across various tasks is quite impressive. This study utilizes OpenAI's GPT-4o, one of the most popular state-of-the-art models available to assess short-answer data. One of the problems in LLM-based evaluation of student answers on courses is the need to specifically fine-tune a model with a specific set of questions and answers. As there are numerous courses where the amount of student submissions is not enough for fully creating custom AI models, the eyes of the educators turn toward LLMs without any specific fine-tuning, utilizing instead their performance on tasks based only upon user input. This creates a need to understand the best practices for achieving the best possible results.

This article aims to answer the following research questions:

**RQ1.** How reliably can an LLM assess short answers related to computer science?

**RQ2.** Is there a difference in results if we only consider answers that had received full points?

## 2. Background and related work

The usage of LLMs is not without limitations. LLMs output can also contain non-accurate information. This phenomenon is known as hallucination. LLM is hallucinating when the output contains plausible but factually incorrect or nonsensical information (Xu, Jain, & Kankanhalli, 2024). Additionally, in multiple-choice questions, the placement of choices, as well as the used symbols have been shown to greatly affect the LLM's performance on different multiple-choice-question benchmarks (Alzahrani et al., 2024). Furthermore, the current models are not able to formally reason but instead seem to follow the patterns that they have seen in their training data, as suggested by research conducted by Mirzadeh et al. (2024) in evaluating current state-of-the-art performance in mathematical reasoning tasks. The different model parameters affect it as well. One of these parameters is temperature. Temperature affects the likelihood of which the next token (word or part of word) will be in the model's output. Lower temperature leads to more consistent outputs, while high temperature creates more diverse output. Often temperature is claimed to control the "creativity" of the LLMs output, but the effects of it are not that straightforward, as noted by Peeperkorn et al. (2024).

Gao et al. (2024) noted in their systematic review that short-answer questions are a common focus of automatic assessment systems. They also noted that most current systems for automatically assessing text-based responses focus on assigning numerical values or automatic classification based on, for example, labels to provide feedback.

The performance of LLMs in auto-assessment of written works has been previously studied. Bui et al. (2024) utilized GPT-3.5 through web-UI for evaluating 200 argumentative essays comprised of 200-300 words and found that GPT-3.5 seemed to lack consistency across two independent rounds of scoring and that the gradings did not align closely with an experienced human reviewer. It should be noted that they utilized UI-version (higher temperature), which was likely at least partly the culprit. Flodén (2024) did a similar study utilizing GPT-3.5, automatically assessing 463 student answers. They produced a review three times using a specific prompt, and then averaged the gained grades, noting that "70% of ChatGPT's gradings were within 10% of the teachers' gradings and 31% within 5%". Kortemeyer, G. (2024) examined GPT-4 performance on short answer grading (ASAG) utilizing datasets Beetle and SciEntsBank. They found that the GPT-4 performs similarly to hand-engineered models but is worse than the pre-trained LLMs that received specialized training.

## 3. Research setup

The data set utilized in this study was Short Answer Feedback (SAF) (Filighera et al., 2022). The data set consists of college-level answers from computer science, more specifically communication networks topic. The data set is anonymized, licensed with cc-by-4.0, and retrieved from Hugging Face (Short Answer Feedback Interest Group 5, n.d.). For a more detailed explanation of the annotation process of the used data set, please refer to the original paper.

While the original data set contains answers from both German and English, this study only utilizes the English version. In this study, the utilized dataset consisted of 31 college-level English questions with a total of 2981 answers. Maximum scores of the questions were parsed from the data set. To unify the scoring, and make worded scoring a feasible approach, the scores were normalized to the scale of 0…1, rounding to one decimal (e.g., 0.25 was transformed to 0.3) before they were run through the pipeline. We then defined our acceptable threshold value to be 20%, which corresponds to two categories. This means that the grade given by the AI tool is considered acceptable if it is within the defined threshold.

The tool used for automatic assessment is open-source and publicly available at https://gitlab.utu.fi/tech/soft/tools/edu-ai-tools/TekstiArv. The examples used were randomly picked. Alzahrani et al. (2024) presented in their study that in multiple-choice questions (MCQs) the performance of the models is affected by the order of the questions (also known as position bias). We tried to mitigate this by shuffling and randomly selecting the essays used in the examples. The model used for evaluation was GPT-4o (gpt-4o-2024-08-06) from OpenAI.

The input to the model consisted of the following things: a system-role-based message where the model was described to be a teacher and the task was to assess the given answer based on the criteria. The grading scale was also provided. It should be noted that the rubric used for evaluating the answers was not given in the input of the model, as it was not available in the used data set. Then the LLM was given a question, reference answer, and a list of examples, one for each grade (if any were available) in randomized order. In practice, this meant each prompt included on average 5 different examples which included the answer, grade, and feedback of the said answer. The example selection was made randomly. The input given to the LLM in the role of the user included only the student's answer to be evaluated.

## 4. Methodology

We selected one of the available questions to test with different parameter settings. The decision to exclude initial testing to the submissions of only one question was made due to time constraints. First, we inspected the distribution of different grades on the questions and selected one with a wide grade distribution. Its grading was between 0-1, original grades on one rounding decimal, and the distribution of the answers was positively skewed (right-skewed). The selected question was "WHAT is the purpose of Reverse Path Forwarding and Reverse Path Broadcast? HOW do they work?" with a total of 82 different answers.

Then we ran the pipeline first for numeric grades (0-1, with a step of .1). The experiment was then run again by replacing numerical values with worded evaluations ("*Exceptional*", "*Excellent*", "*Commendable*", "*Proficient*", "*Adequate*", "*Developing*", "*Limited*", "*Deficient*", "*Inadequate*", "*Poor*", "*Failed/Unsubmitted*"). We did this twice, once with the same used examples in the same order, and a second time utilizing randomly selected examples to display the performance difference caused by the selection of the examples. We also ran the pipeline without any additional examples provided to the model

in input (only the grading scale, question, and reference answer provided in the input) with both numeric and worded evaluations to highlight the difference adding examples makes to the performance. All of these were run with temperatures 0.2, 0.4, 0.6, 0.8, and 1. Additionally, we would like to point out that there is a possibility that the performance could differ with a different amount of examples.

We also did a small-scale test of hybrid temperatures of the best-performing temperature from each approach listed in Table 1 by removing the grade from the output. Then we provided the full previous chat history and asked the model to output the grade "again", with the parameter temperature set to 0. For the final testing with the full data set, we selected the best-performing approaches of the initial run. To further validate the results, we decided to run both numeric and worded evaluations with the full data set with the same examples in the same order (Table 2).

## 5. Results

As noted in Chapter 4, we tested all of the different parameters on the initial testing with temperatures varying between 0.2-1. To display the difference that different settings make, we show the performance of the used model in the 0.2 temperature setting in Table 1 but decided to exclude other temperatures from the table for brevity. The only difference in the same seed was that the grading values were replaced to be either numerical or worded, the number of examples and order of them being the same on the runs with the same seed. On the initial testing, the amount of exact grade matches was low regardless of the used temperature. Additionally, the vast majority of the grades the model had given and fell inside the defined threshold on the full run were less than the ones the human assessor had given (Figure 1). We hypothesize that providing an analytical (or any) rubric in the model's input would likely improve results.

Even with the temperature set as 0.2 (low), there is still quite a significant difference between the different approaches. Additionally, the numeric grading outperformed the worded grading performance when utilizing the same examples in the same order. Although the number of automatically assessed submissions was low on the initial run, this trend was visible across different approaches. We also point out that in some approaches, the most high-performing was of higher temperature than 0.2, but this might be because of the rather limited amount of assessed submissions. Still, the highest performing was the numeric with examples and 0.2 temperature. As noted in the methodology, we tested the best-performing approach by removing the grading and prompted the model to only include in its output the (now) missing grade. Unfortunately, this approach produced inconclusive and mixed results and would need to be done on a larger scale to get anything conclusive.

*Figure 1. The normalized distribution of the grades of the full run, worded and numeric with examples in the same order, and the original score is visible to highlight the grade distribution difference.*

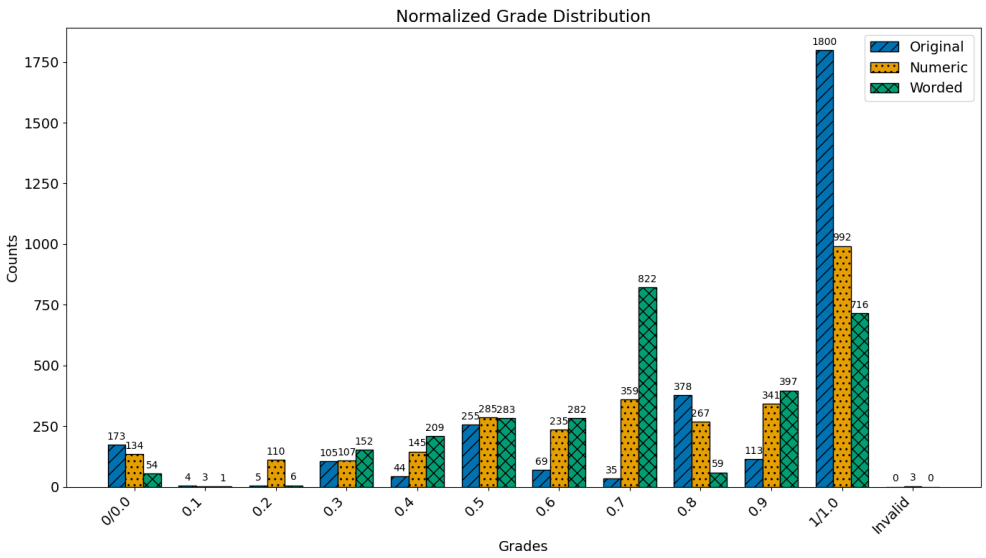

*Table1. The values with different configurations in 0.2 temperature (temp.). The "Same order" column means that the examples used in the input were the same in the same order. A total of 82 submissions were used. The accuracy is reported as a percentage.*

| Type | Had examples | Same order | >=-20% | Equal | <=+20% | Within the threshold (%) |
|---|---|---|---|---|---|---|
| Num. | Yes | Yes | 32.9 | 9.8 | 15.9 | 58.6 |
| **Num.** | **Yes** | **No** | **32.9** | **12.2** | **14.6** | **59.7** |
| Num. | No | N/A | 18.3 | 8.5 | 11 | 53.7 |
| Worded | Yes | Yes | 31.7 | 11 | 13.4 | 56.1 |
| Worded | Yes | No | 26.8 | 11 | 11 | 48.8 |
| Worded | No | N/A | 23.2 | 6.1 | 11 | 40.3 |

*Table 2. The final run consisted of 2981 submissions of the 31 questions, with randomly selected examples (fixed seed). The fully correct (%) is the correctly evaluated full-point answers / actual full points gained submissions in the data.*

| Type | >=-20% | Equal | <=+20% | Within the threshold (%) | Invalid (%) | Fully correct (%) | Full points amount |
|---|---|---|---|---|---|---|---|
| Num. | 24,1 | 41,3 | 9,3 | 74,7 | 0 | 51,4 | 1802 |
| Worded | 24.6 | 30 | 9.1 | 63.7 | 0,1 | 36.5% | 1802 |

## 6. Limitations

The wording, order, selection of examples, and other factors present in the input may affect the results. Furthermore, the validation of the selected (best) approach was done with answers to only one question with 82 answers, which probably is not a sufficient amount to be generalizable. The grade distribution may also skew the results. Furthermore, more diverse prompting strategies could potentially improve the results.

It should also be noted that we looked only at the final produced scoring, ignoring the produced qualitative feedback from our analysis. Inspection of feedback could create a more holistic overview of the LLM capabilities across different reviewing tasks.

As the data set was publicly available on the internet, the data set may be leaked into the used model's (GPT-4o) training data. This may skew the results. As a future work, for validating the results it is strongly recommended to be tested with unforeseen data.

## 7. Discussion

The results indicate that the LLM-based assessment is not quite yet ready to be utilized as a sole method for grading. The model got a little over 40% of the answers fully correct, but with a threshold of 20 %, the correctness raised to 74.7%. As such, the answer to the first research question is that the LLM cannot do the assessment as reliably as needed in the real-life context. It seems that the fully correct answers are assessed a little bit better (51.4%), so the answer to the second research question is positive.

The absence of a well-designed rubric from the input should be noted, as the LLM could only rely on the input given (grading scale, question, reference answer, and examples). The presence of a rubric could likely improve the results further. A decision was made to not include a rubric, as it was absent from the data set, and producing one with LLM might have a caveat: Wu et al. (2024) noted that a gap exists between LLM and human-generated rubrics. The grading results using numeric evaluation outperformed worded grading. It should be noted that the input did not include the definition of the worded grades.

We also want to highlight that LLMs are already used for directly providing students with interactive feedback. Hong et al. (2024) introduced CAELF, an interactive agent to produces interactive feedback with grading automatically to students. These and previous results from other studies raise a question of how these systems should be used. As the only method for assessment, the models clearly are not sufficient. However, in the case of providing a student additional feedback, we feel that, at least with the current state-of-the-art, LLMs could be useful tools. Moreover, such tools can be valuable assets in direct student use as well, as long as the students are also encouraged to challenge the model's output and be reminded about the importance of critical thinking.

## 8. Conclusion and future work

The results of this study seem to suggest that GPT-4o seems to give lower grades compared to the human assessors, and regardless of set-up, the numeric grading seems to outperform worded grading. Although the initial run was done on a low number of 82 submissions, the results indicate that a lower temperature seems to provide better results. All in all, the results still indicate that as a sole assessment tool, the quality of LLM is not yet at a good enough level.

The presented study can be used to guide future research endeavors. The approaches used provide valuable insights for researching LLM's capabilities in the automated assessment of written works. Testing with different grading scales, as well as removing the final grade and injecting it again with a lower temperature setting should be examined at a larger scale at a later date. Moreover, testing different subjects and different kinds of questions can provide better results.

### *Acknowledgements*

This work has been supported by FAST, the Finnish Software Engineering Doctoral Research Network, funded by the Ministry of Education and Culture, Finland.

### *References*

Alzahrani, N., Alyahya, H. A., Alnumay, Y., Alrashed, S., Alsubaie, S., Almushaykeh, Y., ... & Khan, H. (2024). *When benchmarks are targets: Revealing the sensitivity of large language model leaderboards*. arXiv preprint arXiv:2402.01781

Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies*, 1-18.

Filighera, A., Parihar, S., Steuer, T., Meuser, T., & Ochs, S. (2022). Your answer is incorrect… would you like to know why? introducing a bilingual short answer feedback dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1, pp. 8577-8591).

Flodén, J. (2024). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, *51*(1), 201-224.

Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C., & Srinivasa, A. R. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, *6*, 100206.

Hong, S., Cai, C., Du, S., Feng, H., Liu, S., & Fan, X. (2024). *"My Grade is Wrong!": A Contestable AI Framework for Interactive Feedback in Evaluating Student Essays*. arXiv preprint arXiv:2409.07453

Kortemeyer, G. (2024). Performance of the pre-trained large language model GPT-4 on automated short answer grading. *Discover Artificial Intelligence*, *4*(1), 47.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). *Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models*. arXiv preprint arXiv:2410.05229

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). *Is temperature the creativity parameter of large language models?*. arXiv preprint arXiv:2405.00492

Short Answer Feedback Interest Group 5. (n.d.). *saf_communication_networks_english* [Dataset]. Hugging Face. Retrieved January 30, 2025, from https://huggingface.co/datasets/Short-Answer-Feedback/saf_communication_networks_english

Wu, X., Saraf, P. P., Lee, G. G., Latif, E., Liu, N., & Zhai, X. (2024). *Unveiling scoring processes: Dissecting the differences between llms and human graders in automatic scoring*. arXiv preprint arXiv:2407.18328

Xu, Z., Jain, S., & Kankanhalli, M. (2024). *Hallucination is inevitable: An innate limitation of large language models*. arXiv preprint arXiv:2401.11817