Comparison of Different LLMs on Grading Automation

Anton Chulakov
Innopolis University

Abstract

This project investigates the effectiveness of several Large Language Models (LLMs) in automating the grading of student papers. Using professor-evaluated student assignments as a benchmark, multiple models were compared based on their ability to reproduce the reference grading. A custom similarity metric was used to evaluate and rank the performance of each model. The findings suggest that model complexity does not always correlate with accuracy in numerical grading tasks.

1 Introduction

Automating the grading process with LLM has the potential to save time and ensure consistency. This project evaluates the performance of various LLMs against a professor's rubric-based evaluation of 57 student papers. The grading of each model was analyzed for similarity to human scores using a simple quantitative formula.

2 Related Work

Recent studies have explored the role of LLMs in educational assessment. One study compared GPT-4 to human graders in political science essays and found strong alignment in complex academic evaluation, albeit with limitations [1]. Another case study focused on ChatGPT-4's performance in essay evaluation, highlighting challenges in grading nuanced texts [2].

3 Dataset and Preprocessing

The dataset consisted of 57 anonymized student papers. Personally identifiable information was removed using a custom Python script `anonymize_data.py`. The anonymization ensured fairness and privacy in the evaluation.

4 Evaluation Method

Each model was used to assess the same student papers using `evaluate_ai.py`, which supports arguments such as model selection and file count (e.g., `--models aitunnel:gpt-4o --max-files 57`).

4.1 Professor's Evaluation

The professor used a rubric-based table. Each criterion had a maximum score. Final grades were calculated as:

$$\text{Final Score} = \frac{\sum \text{Points Awarded}}{\sum \text{Max Points}} \times 100$$

The "Feedback" column was excluded from the analysis due to its subjectivity.

4.2 Similarity Formula

Model scores were compared to professor scores using the formula:

$$\text{similarity} = 100 - \left(\frac{|\text{model\_val} - \text{ref\_val}|}{\text{max\_score}} \right) \times 100$$

5 Models Evaluated

The following models were tested:

- `gpt-4.1-nano`, `gpt-4.1`, `o3-mini`, `o3`
- `gpt-4o-mini`, `gpt-4o`, `gpt-4-turbo`, `gpt-3.5-turbo`
- `deepseek-chat-v3`, `deepseek-r1`, `deepseek-chat`
- `gemini-2.5-flash`, `gemini-2.5-pro-preview`
- `llama-4-maverick`, `llama-3.3-70b-instruct`

6 Results

Model outputs were saved as CSVs and compared to the reference evaluation. Visualization of similarity scores was done using both `matplotlib` and `seaborn`. While `matplotlib` handled figure creation and saving, `seaborn` was used to generate the heatmap and bar plot that compare model accuracy.

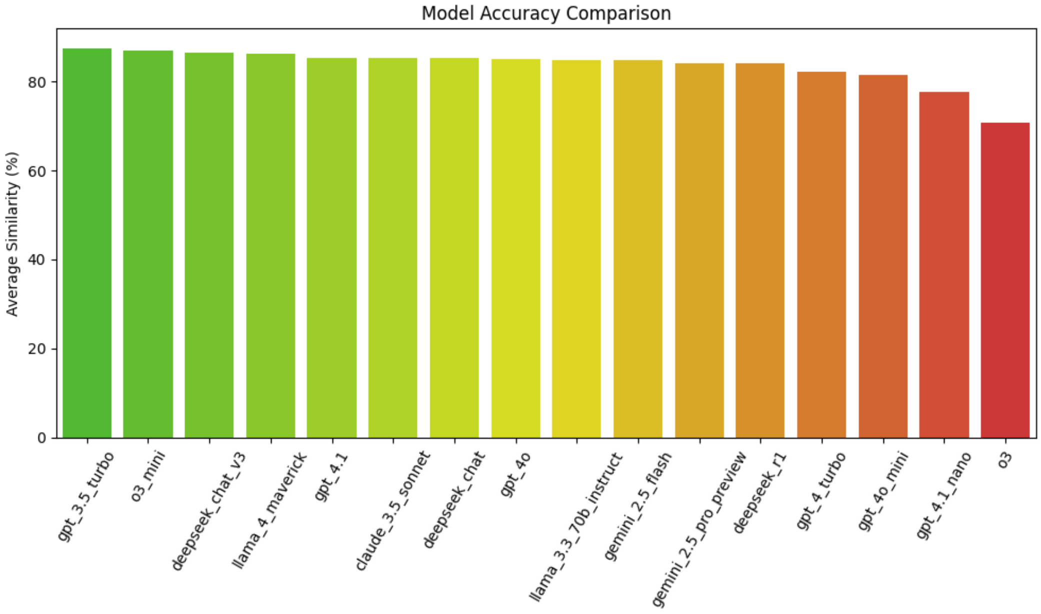

Figure 1: Model Accuracy Scores in Descending Order

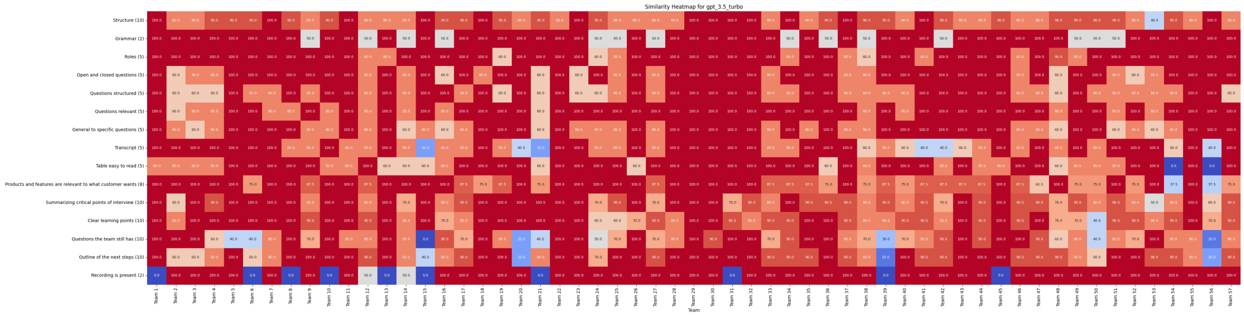

Figure 2: Heatmap of grading similarity for the best-performing model: `gpt_3.5_turbo`

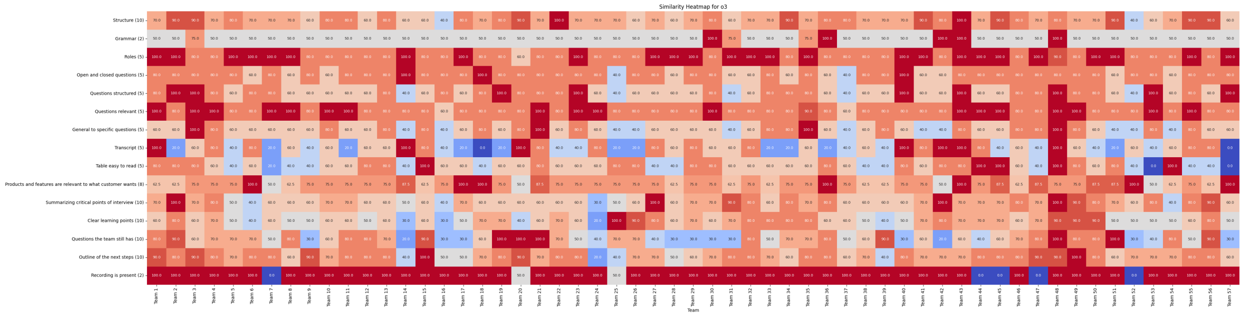

Figure 3: Heatmap of grading similarity for the worst-performing model: `o3`

Model Accuracy Summary and Ranking

Model	Accuracy (%)
gpt_3.5_turbo	87.51
o3_mini	86.88
deepseek_chat_v3	86.59
llama_4_maverick	86.18
gpt_4.1	85.39
claude_3.5_sonnet	85.31
deepseek_chat	85.25
gpt_4o	85.15
llama_3.3_70b_instruct	84.89
gemini_2.5_flash	84.88
gemini_2.5_pro_preview	84.04
deepseek_r1	84.00
gpt_4_turbo	82.23
gpt_4o_mini	81.57
gpt_4.1_nano	77.68
o3	70.76

Table 1: LLM Accuracy Rankings (Most to Least Accurate)

7 Discussion

Some of the simpler or older models outperformed newer ones. This suggests that larger models may overcomplicate answers or hallucinate when given tasks that require structured, numeric responses.

Interestingly, the top three performing models—gpt_3.5_turbo, o3_mini, and deepseek_chat_v3 — are comparatively affordable, with deepseek_chat_v3 even being free to use. This indicates that for rubric-based grading tasks, which are largely numerical and well-structured, using simpler or cheaper models can actually yield better results than more advanced and expensive alternatives.

8 Conclusion and Future Work

LLMs show promise for automated grading. However, model size or cost does not always equate to accuracy. Future work may explore combining numeric evaluation with textual feedback analysis or prompt-tuning models specifically for grading tasks.

References

[1] Tigran Melkonian, Arman Atabekyan, and Gevorg Petrosyan. Large language models in student assessment: Comparing chatgpt and human graders. 2024. Available at: https://www.researchgate.net/publication/381655103.

[2] Rui Zhou, Xia Wang, and Carol Lee. Llms in automated essay evaluation: A case study. 2024. Available at: https://www.researchgate.net/publication/380766203.