Criteria for Demonstration Examples for LLM Assessment of Science Essays

Mahsa Sheikhi Karizaki[1*], Zhaohui Li[2**], and Rebecca J. Passonneau[1]

[1] Pennsylvania State University, State College, PA 16801, USA
(mfs6614, rjp49)@psu.edu
[2] University at Buffalo, Buffalo, NY 14260, USA
zli253@buffalo.edu

Abstract. Science writing is a core practice in science education, yet teachers often find it challenging to provide constructive feedback in real time. Yet students understand and benefit more from feedback when it is timely. It is therefore natural to ask whether LLMs could ease the burden on teachers through automated or partially automated formative assessment. In this research, we investigate the effectiveness of prompting instruction-tuned LLMs to assess middle school science essays for the presence of main ideas specified in a rubric, varying the number and quality of demonstration examples. In a comparison of Llama-3-8b with three GPT models, we found that prompting GPT4o with three examples outperformed customized AI assessment tools. We also found consistent results across different selections of demonstration examples in the prompt. Our results add to a body of recent research demonstrating the potential benefits of LLMs in educational assessments, alongside the importance of prompt design.

Keywords: LLMs · Science Essay Assessment · Formative Assessment · LLM Prompt Design

1 Introduction

Science writing is a core science practice, and can help students learn to explain science ideas. Yet teachers find it challenging to provide constructive feedback in real-time [14]. The U.S. Department of Education supports AI innovations like large language models (LLMs) in education as partners with human teachers [8]. Recent research indicates that instruction-tuned LLMs have the potential to accurately assess students' responses to assessment questions in some domains. For instance, one study [26] shows that ChatGPT correctly evaluated 75% of students' answers in a mathematics learning game called Decimal Point, which involved self-explanation responses. However, another study [33], also in the domain of mathematics, showed that 82% of ChatGPT's suggestions were not novel or insightful. Instead of short answers in mathematics, we focus on LLMs

* Equal contribution.
** Equal contribution.

for formative assessment of ideas students express in their science essays, and how best to prompt LLMs for this task.

Performance of LLMs on diverse tasks can vary widely, depending on the exact structure of the prompts used in defining the task for the LLM. Further, much research has shown that providing LLM prompts with appropriate demonstrations—typically including both positive and negative examples—significantly improves performance over prompting without demonstrations (referred to as zero-shot versus few-shot) [18,19,21,35,4]. However, we are unaware of work that investigates how many and what types of examples to include in prompts for science essay assessment. In few-shot learning, it is generally assumed that higher-quality examples yield better results [4]. Contrary to this assumption, our results indicate that the quality of examples does not significantly influence the outcome. On the other hand, we find that the number of positive and negative examples does matter. Our results therefore provide guidance to teachers about whether LLMs can facilitate formative assessment of science essays, and what types of prompts are most effective.

Our study investigates the utility of large language models (LLMs) to assess science explanation essays, based on a rubric listing the main ideas in a curriculum that students should express in their essays. The use of LLMs is compared with two high-performing, customized assessment tools, one of which was employed throughout a 5-year NSF-funded project. Additionally, the study examines the impact of using different demonstrations. The goal is to determine whether LLMs can perform as well as customized AI assessment methods, and therefore make automated assessment more available to science teachers.

Three research questions were addressed: (1) How does LLM prompting for assessment of the main ideas in middle school science essays compare with customized AI tools? (2) What is the performance impact of including examples, and how does the balance between positive and negative examples affect performance? (3) How much does example quality matter? We found that with only three examples in the prompt, LLMs can outperform custom AI tools. Further, through exploration of multiple combinations of high versus moderate quality examples, we found that using two positive and one negative example matters more than the specific choice of examples.

We first discuss related work on the potential benefits and pitfalls of LLMs for assessment. Then we give an overview of the 5-year project that provided the data for our investigation, the role of science writing, and other AI-enabled assessment methods that were utilized on the project. The three subsequent sections present our methods, data, and experimental results on these three research questions. We then synthesize and discuss our findings, followed by a brief conclusion that summarizes our contributions.

2 Related Work

Automated assessment of essays has had a long tradition, going back to Project Essay Grade [27], and later the highly successful e-rater [7]. These and similar

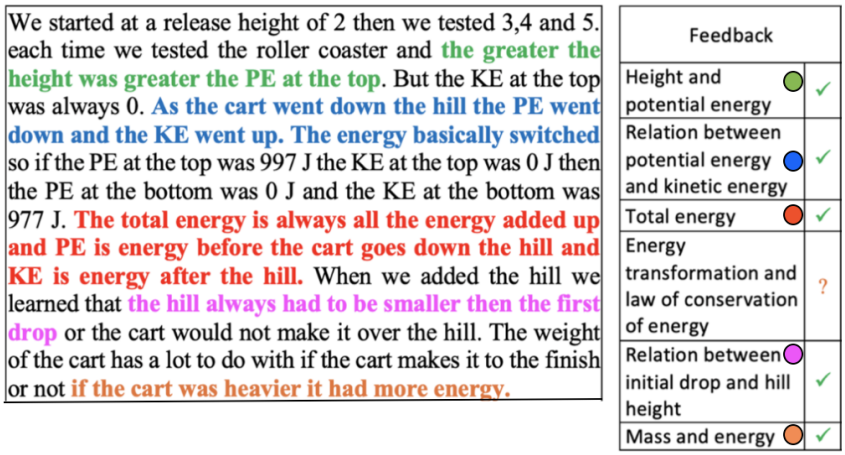

Fig. 1. Sample essay (left) and automated feedback table (right). The main ideas are color-coded. A green check indicates the relevant main idea was detected in the essay. A question mark indicates the main idea was not detected by the automated tool.

early systems relied on manual engineering of feature representations of input texts, and lacked the capacity to evaluate central scientific concepts or reasoning structures in students' essays. Transformer-based approaches to language-modeling, such as BERT [10], were shown to learn very effective feature representations automatically, leading to the use of pretrained language models (PLMs) for many tasks. Use of PLMs significantly expanded the scope of content analysis of text, improving performance in detecting topic relevance and coherence [32,3]. However, PLMs had other costs, such as the requirement for extensive labeled data for fine-tuning the representations to specialized domains and tasks, such as science writing [31].

Recent breakthroughs in generative LLMs, particularly those in the GPT family [5], have heightened interest in their zero-shot and few-shot capabilities for essay evaluation. A compelling case for LLMs over PLMs is that they avoid the high cost of collecting training data for fine-tuning. Educators and researchers have experimented with instruction-tuned models to score essays or identify missing ideas by carefully designing prompts, often leveraging in-context learning, chain-of-thought or self-consistency strategies [34,22]. For example, chain-of-thought prompting allows an LLM to provide a transparent rationale for its judgments, thus aiding in identifying key scientific ideas, hypotheses, or evidence usage [36]. Furthermore, LLMs such as BART and T5 can generate concise summaries of essays, enabling rubric-based comparisons to confirm whether students have correctly conveyed the core concepts of a scientific argument [17,28]. Empirical studies increasingly underscore the potential of these methods in real-world educational settings, including comparisons between ChatGPT and human graders [20], assessments of reliability and validity in essay grading [37], and the

role of LLMs as collaborative partners in student essay evaluation [13]. Recent studies have also shown that LLMs can provide consistent and explainable evaluation feedback, especially when guided by pre-specified rubrics [24,13].

While the advances discussed above are promising, effective use of LLMs shifts the up-front cost towards extensive prompt engineering from human experts [16,28]. Prior studies emphasize that prompts must align with instructional goals to yield precise and actionable feedback [23]. Yet, to the best of our knowledge, there is no existing research on optimally selecting demonstrations (both positive and negative examples) when prompting an LLM to identify the main ideas in science writing. To fill this gap, our work investigates how varying the number and composition of few-shot demonstrations impacts performance.

In addition, recent work by Min et al. [25] offers important insight into what drives performance in in-context learning. Their study shows that the effectiveness of few-shot prompting is less about whether the demonstrations contain correct labels, and more about how well the examples align with the label space, input distribution, and output format of the target task. This suggests that structural features and distributional alignment play a critical role in LLM performance.

3 Background

The essays for the work presented here are drawn from an NSF-funded project to develop an online curriculum for middle school science that investigated the role of science writing in science learning, and utilized a teacher co-design approach throughout the project. The project developed a web-delivered middle school curriculum about mass and energy where students conducted experiments in a simulated roller coaster environment, then entered short answers to different sets of explanation questions based on their experiments. Interleaved with these activities, they wrote and revised two explanation essays. Here we discuss the first essay that prompts students to explain six main ideas. An example essay is shown in Figure 1. Also shown is a checklist of the main ideas in the essay prompt that was provided as feedback to students, for which we used a content assessment tool called PyrEval [12]. Based on the feedback, students submitted revisions of each essay. This process aimed to enhance students' understanding of physics principles, as well as to improve their ability to express scientific concepts effectively. Improved ability to express science ideas in writing also benefits teachers' understanding of student learning. Previous work reports on evidence that students' revised essays show improvement [15].

In this paper, we investigate the performance of large language models (LLMs) to evaluate the elaboration of main ideas in science explanation essays, and compare LLM performance with PyrEval and another custom AI tool, VerAs. PyrEval [12] is a publicly available off-the-shelf tool that has been utilized in previous research on formative assessment of student essays, due in part to its minimal training data requirements and high modularity: only four to five essays are needed for PyrEval to automatically generate the content model it uses for

Prompt
"Your final task is to find one sentence in the student essay that is most relevant to the MAIN_IDEA. If it finds a matched relevant_sentence, make the final_result=1, make the relevant_sentence value equal to this sentence, otherwise make the relevant_sentence equal to NONE and final_result=0. Then output a JSON object that contains the following keys: relevant_sentence, final_result. "
Use the following format: "Output JSON: <json with relevant_sentence and final_result>"
"Positive Example [I]"
"Positive Example [II]"
"Negative Example [III]"
"Now, analyze the following student essay:"
"The student essay is as follows: " [Student Essay Text]
"The MAIN_IDEA is as follows: " [MAIN_IDEA Text]

Table 1. The LLM prompt template used in this research.

assessment. We selected five high-quality essays to generate an initial PyrEval content model, which we then manually curated through testing on initial essays. Thus one limitation is PyrEval's need for a manual curation step. In addition, PyrEval breaks an essay down into propositions, depending heavily on correct sentence-final punctuation, which middle school students do not always follow. VerAs [1] is a custom end-to-end neural architecture for science writing assessment. In contrast to PyrEval, VerAs requires thousands of training examples which makes it impractical for many educational settings. In contrast, LLMs require no task-specific training or fine-tuning.

Even though LLMs like ChatGPT have issues addressing ethical and pedagogical concerns such as academic integrity and critical thinking, they could still be very useful assessment tools for teachers [9,2]. These models can potentially help evaluate the main ideas in student essays, providing immediate and detailed feedback [11]. However, creating effective prompts to accurately detect these ideas can be challenging, especially for teachers unaware of prompt engineering, and who already have burdensome demands on their time. One possible solution to this problem would be to utilize a template based on a pre-designed prompt that merely requires teachers to fill in specific blanks, and to provide positive and negative examples related to an essay's main ideas. Streamlining the prompt method for teachers could allow them to leverage the power of LLMs, making the technology more accessible and practical for educational use.

4 Data

The data is from writing samples of students from three public middle schools who participated in a 3-week physics unit on the Law of Conservation of Energy

Dev		Test		Comb	
Avg	(Stdv)	Avg	(Stdv)	Avg	(Stdv)
308.68	(123.59)	381.92	(146.63)	345.30	(139.95)

Table 2. Average essay length in words in the Dev and Test sets, and Combined.

	Positive (1)	**Negative (0)**
Main_Idea 1	46	14
Main_Idea 2	43	17
Main_Idea 3	26	34
Main_Idea 4	37	23
Main_Idea 5	43	17
Main_Idea 6	42	18
Total	237	123

Table 3. Label distribution in the Dev Set.

and related energy concepts.[3] As mentioned above, students used roller coaster simulations and a digital notebook to conduct experiments and write essays. Students were prompted to think about six main ideas to include in their first essay (see Figure 1). For example, to adequately convey the relation between mass and energy, the student essay should contain a sentence that conveys the idea that "a cart with greater mass will have greater energy in motion or at rest." The LLMs' task is to indicate whether an input essay contains the main idea specified in the LLM prompt (see Table 1). The study as a whole included two sets of student essays from three years of classroom deployment.

A total of 120 essays were selected from both original and revised submissions to the first essay prompt, and randomly halved into a development set (Dev) for LLM prompt development, and a test set (Test) to replicate results on an unseen set. Table 2 shows average essay length to be about 308 words in the Dev set, 382 in the Test set, and 345 overall. Essays were manually annotated for the presence or absence of the six main ideas. Inter-rater agreement was assessed on 20% of the dataset using Cohen's Kappa ($\kappa = 0.768$), which indicated substantial agreement between the two annotators. After establishing this level of consistency, the remaining essays were labeled by a single annotator.

Table 3 shows the distribution of presence and absence labels for each of the six main ideas in the development set. On average, essays included twice as many main ideas as they omitted, with 237 main ideas expressed overall, and 123 times that main ideas were omitted: further, main ideas 3 and 4 were the most challenging for students. We later used the empirical distribution of 2:1 positive to negative cases as inspiration for prompt design in our experiments, and found that this ratio in the demonstration examples worked best.

[3] Student data collection had IRB approval to study, but not share, the collected data. Students and parents provided assent and consent to participate. No demographic data such as race or gender was collected.

5 Methods

The models under examination are among the most prominent instruction-tuned large language models available. Llama-3-8b (from Meta AI), is an advanced iteration in the Llama series, known for its efficiency and strong performance. GPT-3.5-turbo (from OpenAI), builds upon the success of GPT-3, offering enhancements in speed and understanding that make it a popular choice for many applications. GPT-4 (a later OpenAI GPT model) incorporates substantial improvements. GPT-4o is a later variant with further performance improvements. We compared their performance with and without demonstration examples, for up to three examples. Due to budget constraints, we could not test larger numbers of examples, as the number of combinations of positive and negative examples grows nearly exponentially with the total number of examples. For each LLM model, we used default parameter settings apart from temperature. For temperature, which controls diversity of output, we used zero to ensure consistent results across experiments. The demonstrations were incorporated into variations of the prompt shown in Figure 1, using different numbers of positive and negative examples, and varying the specific essays used as positive or negative examples by quality.

Through experiments documented below where we adapted state-of-the-art prompting [6] on the Dev set, we found that using two positive examples and one negative example performed better than using no examples, or one positive example, or one positive and one negative. The design of our final prompt was guided by established principles in prompt engineering: simplicity, structural consistency, and alignment with emerging standards [29]. By limiting variation in prompt wording and explicitly formatting the output in a JSON object, we minimized prompt-induced variance and supported content-focused assessment. Rather than asking whether performance could increase by using more than four examples, we focus here on how the specific example essays in the prompt affect performance. Thus we arrived at the high-performing prompt illustrated in Table 1, which we used for each of the main ideas from Figure 1 by replacing MAIN_IDEA_Text with a sentence that expressed that main idea, then inserting two positive example essays that expressed this main idea (I and II), and an essay that did not express it as a negative example (III).

The demonstration examples were drawn from the original five model essays, that had been manually curated and edited to create the PyrEval content model, indicated here as set 1 (S1): A, B, E, F, G. We used a random selection of three additional student essays as a contrasting set 2 (S2): C, D, H. S1 essays contained all six main ideas, whereas S2 essays did not. All positive examples for prompts were selected from S1. Thus there were $\binom{5}{2} = 10$ distinct positive pairs. Negative examples for a given main idea were either selected from S1 by dropping the sentence that contained that idea (symbolized as nA, nB, etc.), or from S2, yielding a total of eight negative examples. Conversion of S1 essays to negative examples omitted only one main idea, whereas S2 examples omitted multiple main ideas. Also, S2 examples were less well-written. For a given positive pair (e.g., A, B), we use only the six negative examples other than those derived

Model	Accuracy (%)	F1 (%)
Llama-3-8b (Zero-shot)	65.55	39.60
Llama-3-8b (Two-shot)	69.72	50.81
Llama-3-8b (Three-shot)	71.94	56.83
GPT-3.5-turbo (Zero-shot)	66.11	41.31
GPT-3.5-turbo (Two-shot)	72.77	61.63
GPT-3.5-turbo (Three-shot)	75.83	66.66
GPT-4 (Zero-shot)	68.88	49.23
GPT-4 (Two-shot)	75.56	64.95
GPT-4 (Three-shot)	78.89	71.20
GPT-4o (Zero-shot)	75.56	64.64
GPT-4o (Two-shot)	83.88	79.92
GPT-4o (Three-shot)	**86.11**	**83.80**

Table 4. The comparison of LLMs' performance with different types of demonstration.

from the positive pair (e.g., nE, nF, nG, C, D, H). This gives a total of 60 distinct conditions. The id for each condition is a string indicating the positive pair followed by the negative example; e.g., A&B_nE refers to essays A and B as positive examples, with E as a negative example.

Finally, we compared the final best-performing prompt on the best LLM with PyrEval and VerAs, using the reserved test set.

6 Experiments

First, we compared performance of zero-shot, two-shot, and three-shot demonstrations by sampling examples from S1 and S2. Table 4 shows the efficacy of instruction tuning in enhancing model performance: GPT-4o with three-shot (two positive and one negative example) demonstrated superior capabilities in both accuracy and F1 score across different settings. One-shot results are omitted because the experiment conducted—with either one positive or one negative example—produced consistently poor performance, indicating that a single example did not provide sufficient context. As a result, we discontinued further one-shot experiments. Two-shot and three-shot results shown here are the best combinations of positive and negative for each, which were one positive and one negative for two-shot, and two positive and one negative for three-shot.

Second, to test the impact of different demonstrations, we selected all combinations of high-quality essays as positive demonstrations paired with either S1 or S2 negative examples, giving 60 conditions as described above. Thus, we compared whether performance differed between S1 and S2 negative examples, and with different positive pairs. Figure 2 shows that of the 60 conditions, A&E_nD (on the right side with the deepest red color) has the highest F1 (84.73%) and accuracy scores (86.94%). There are some noticeable patterns: for instance, the positive pairs A&E and F&G consistently yielded higher F1 and accuracy. In contrast, pairs B&F and E&F had lower F1 and accuracy. Figure 2 shows that

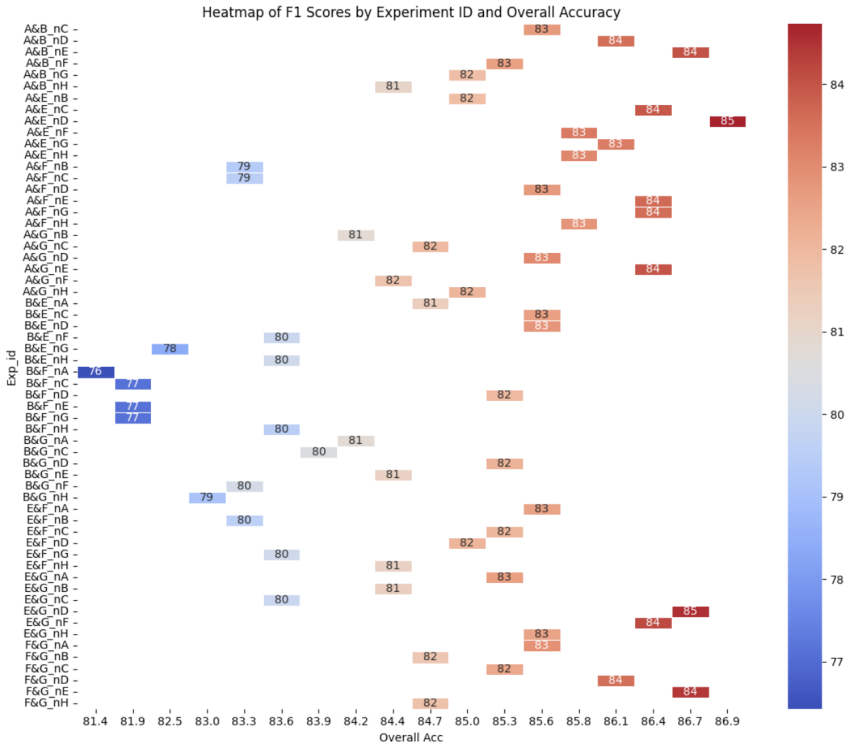

Fig. 2. GPT-4o performance for our 60 experimental conditions (y-axis) by overall accuracy (x-axis), with the F1 scores overlaid upon the accuracy rectangles. The heatmap colors range from cool (blue) to warm (red), with blue indicating lower F1 scores and red indicating higher F1 scores.

most high-performing pairs had essay A as one of the positive examples, while most low-performing pairs had essay B as one of the positive examples.

We summarize the results from Figure 2 by calculating separate averages for each positive pair and each negative example. The left side of Figure 3 shows the average accuracy and F1 for each positive pair, thus averaging over the use of each of the eight negative examples (see above). The right side of Figure 3 shows the average accuracy and F1 for each negative example, where there are 10 positive pairs for the original 3 negative examples (C, D, H), and 6 positive pairs for the negative examples created by modifying the original essays that expressed all 6 main ideas. That is, for nA, there are $\binom{4}{2} = 6$ positive pairs from B, E, F, and G. The best negative example appears to be D, showing highest average accuracy of 85.80% and highest average F1 of 83.10%. As shown in Figure 3, the best positive pair in the experiments appears to be A&E. This pair achieves the highest average accuracy of 86.02% and the highest average F1 of 83.37%.

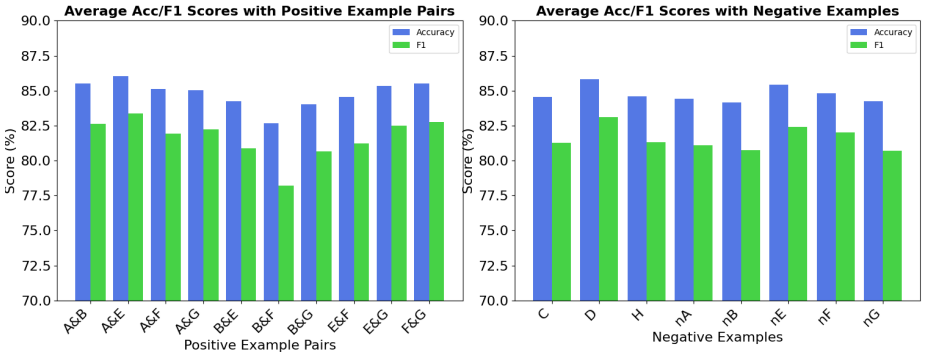

Fig. 3. The comparison of different demonstration examples for GPT-4o. The student quality negative examples are C, D, and H. The negative examples derived by modifying positive ones are nA, nB, nE, nF and nG.

Model	Acc (Dev set)	F1 (Dev set)	Acc (Test set)	F1 (Test set)
PyrEval	79.72	76.62	75.27	70.00
VerAs	78.88	76.14	75.27	71.09
GPT-4o*	**86.94**	**84.73**	**85.83**	**83.26**

Table 5. Comparison of our method (*with A&E and nD) against baselines.

Thus use of A&E as positive examples consistently yielded better performance compared to other positive pairs. The result also shows no significant difference between S1 versus S2 as negative examples.

To test the generalization of the prompt to our test set in comparison with PyrEval and Veras, we use A&E_nD. The results in Table 5 show that use of this prompt with GPT-4o outperforms the two custom AI models.

7 Discussion

For the first research question about the comparison of LLMs with custom AI tools, we found that with appropriate prompt design, GPT-4o outperformed PyrEval and VerAs by a large margin. The advantage of using LLMs is that they require only a few demonstration examples, whereas end-to-end neural tools like VerAs demand a large amount of training data, and the PyrEval pipeline architecture requires manual tuning of the content model.

For the second research question about whether to use examples and in what balance of positive to negative, Table 4 results showed that the 3-shot experiments performed significantly better than the zero-shot or two-shot experiments. The best three-shot performance used two positive and one negative example. We speculate that the 2:1 ratio of positive to negative examples in the demonstration performs well because it aligns with the data distribution (Positive: Negative = 2:1). Confirmation of whether demonstrations in prompts should align with the

empirical balance of positive and negative examples would require rigorous tests with multiple datasets.

For the third research question regarding whether quality of examples matters, we tested different combinations of positive and negative examples from S1 (High-quality) and S2 (Student-quality). Given the curriculum goal of improving students' science explanations, here a high quality essay means inclusion of the relevant main ideas. Set 1 essays were curated and revised versions of student essays, with minimal grammar issues and high content fidelity[30]. The experimental results, as shown in Figure 3 for the negative examples, indicate no significant difference between using S1 and S2 as negative examples, which suggests that the educator does not need to spend too much time picking up good negative examples. However, we did find that certain specific combinations, such as A&E_nD, performed best. Further exploration is needed to understand why these combinations work better and how to select the most effective demonstration examples.

Across the 60 conditions in our experiment with prompt quality, as shown Figure 3, prompts with nA generally led to better performance than those with nB as negative examples. By analyzing essays A and B, we found that essay A was structured in a list-like manner, highlighting the main ideas, whereas essay B had a more typical discursive essay structure. We speculate that similar to humans, it may be easier for LLMs to understand and process a list of main ideas rather than an essay written in a more naturalistic style.

This study has several limitations that should be considered when interpreting the results. First, the dataset consisted of 120 annotated essays, which may limit the generalizability of our findings across broader educational contexts and grade levels. Second, while the inter-rater agreement (Cohen's Kappa = 0.768) indicates substantial consistency, some variation in labeling may introduce minor noise into the development and test process, particularly on more complex main ideas. Third, although we examined overall LLM accuracy and F1 scores, future work should explore how reliably LLMs assess different types of main ideas across different STEM domains.

8 Conclusion

With the release of GPT4o, instruction-tuned LLMs have reached a high level of performance on diverse tasks, including assessment of students' science essays, given an appropriate prompt. Our results show that GPT4o prompted with two positive and one negative example outperforms custom AI tools developed for assessment of student STEM writing. We also observed certain trends that could merit further investigation, such as whether the proportion of positive to negative examples in a prompt should reflect the empirical data distribution.

References

1. Atil, B., Sheikhi Karizaki, M., J. Passonneau, R.: VerAs: Verify then assess stem lab reports. In: International Conference on Artificial Intelligence in Education. pp. 133–148. Springer (2024)
2. Barnes, T., Burriss, S., Danish, J., Finkelstein, S., Humburg, M., Limke, A., Molvig, O., Reichert, H.: Toward ethical and just AI in education research. Community for Advancing Discovery Research in Education (CADRE) (2024)
3. Beseiso, M., Alzahrani, S.: An empirical analysis of BERT embedding for automated essay scoring. International Journal of Advanced Computer Science and Applications **11**(10) (2020)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Bsharat, S.M., Myrzakhan, A., Shen, Z.: Principled instructions are all you need for questioning llama-1/2, GPT-3.5/4. arXiv preprint arXiv:2312.16171 (2023)
7. Burstein, J., Tetreault, J., Madnani, N.: The e-rater® automated essay scoring system. In: Handbook of automated essay evaluation, pp. 55–67. Routledge (2013)
8. Cardona, M., Rodríguez, R., Ishmael, K.: Artificial intelligence and the future of teaching and learning. office of educational technology. 2023
9. Čavojskỳ, M., Bugár, G., Kormaník, T., Hasin, M.: Exploring the capabilities and possible applications of large language models for education. In: 2023 21st International Conference on Emerging eLearning Technologies and Applications (ICETA). pp. 91–98. IEEE (2023)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423
11. Gan, W., Qi, Z., Wu, J., Lin, J.C.W.: Large language models in education: Vision and opportunities. In: 2023 IEEE international conference on big data (BigData). pp. 4776–4785. IEEE (2023)
12. Gao, Y., Warner, A., Passonneau, R.J.: PyrEval: An automated method for summary content analysis. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
13. Ishida, T., Liu, T., Wang, H., Cheung, W.K.: Large language models as partners in student essay evaluation. arXiv preprint arXiv:2405.18632 (2024)
14. Karizaki, M.S., Gnesdilow, D., Puntambekar, S., Passonneau, R.J.: How well can you articulate that idea? insights from automated formative assessment. In: International Conference on Artificial Intelligence in Education. pp. 225–233. Springer (2024)
15. Kim, C., Puntambekar, S., Lee, E., Gnesdilow, D., Karizaki, M.S., Passonneau, R.: NLP-enabled automated feedback about science writing. In: Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024, pp. 2431-2432. International Society of the Learning Sciences (2024)

16. Lewis, M.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)
17. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.703, https://aclanthology.org/2020.acl-main.703/
18. Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., Chen, W.: What makes good in-context examples for GPT-3? In: Agirre, E., Apidianaki, M., Vulić, I. (eds.) Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. pp. 100–114. Association for Computational Linguistics, Dublin, Ireland and Online (May 2022). https://doi.org/10.18653/v1/2022.deelio-1.10, https://aclanthology.org/2022.deelio-1.10/
19. Lou, R., Zhang, K., Yin, W.: Large language model instruction following: A survey of progresses and challenges. Computational Linguistics pp. 1–10 (2024)
20. Lundgren, M.: Large language models in student assessment: Comparing ChatGPT and human graders. arXiv preprint arXiv:2406.16510 (2024)
21. Maronikolakis, A., Köksal, A., Schütze, H.: Sociocultural knowledge is needed for selection of shots in hate speech detection tasks. arXiv preprint arXiv:2304.01890 (2023)
22. Marvin, G., Hellen, N., Jjingo, D., Nakatumba-Nabende, J.: Prompt engineering in large language models. In: International conference on data intelligence and cognitive informatics. pp. 387–402. Springer (2023)
23. Matelsky, J.K., Parodi, F., Liu, T., Lange, R.D., Kording, K.P.: A large language model-assisted education tool to provide feedback on open-ended responses. arXiv preprint arXiv:2308.02439 (2023)
24. Meyer, J., Jansen, T., Schiller, R., Liebenow, L.W., Steinbach, M., Horbach, A., Fleckenstein, J.: Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. Computers and Education: Artificial Intelligence **6**, 100199 (2024)
25. Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., Zettlemoyer, L.: Rethinking the role of demonstrations: What makes in-context learning work? In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 11048–11064. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022). https://doi.org/10.18653/v1/2022.emnlp-main.759, https://aclanthology.org/2022.emnlp-main.759/
26. Nguyen, H.A., Stec, H., Hou, X., Di, S., McLaren, B.M.: Evaluating ChatGPT's decimal skills and feedback generation in a digital learning game. In: European Conference on Technology Enhanced Learning. pp. 278–293. Springer (2023)
27. Page, E.B.: The imminence of... grading essays by computer. The Phi Delta Kappan **47**(5), 238–243 (1966)
28. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)

29. Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A.: A systematic survey of prompt engineering in large language models: Techniques and applications (2025), https://arxiv.org/abs/2402.07927
30. Singh, P., Passonneau, R.J., Wasih, M., Cang, X., Kim, C., Puntambekar, S.: Automated support to scaffold students' written explanations in science. In: Rodrigo, M.M.T., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education. AIED 2022. Lecture Notes in Computer Science, vol. 13355, pp. 679–691. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-11644-5_64, https://doi.org/10.1007/978-3-031-11644-5_64
31. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? In: Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18. pp. 194–206. Springer (2019)
32. Tay, Y., Phan, M., Tuan, L.A., Hui, S.C.: Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32.1 (2018)
33. Wang, R., Demszky, D.: Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In: Kochmar, E., Burstein, J., Horbach, A., Laarmann-Quante, R., Madnani, N., Tack, A., Yaneva, V., Yuan, Z., Zesch, T. (eds.) Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023). pp. 626–667. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.bea-1.53
34. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
35. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur) **53**(3), 1–34 (2020)
36. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
37. Yavuz, F., Çelik, Ö., Yavaş Çelik, G.: Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. British Journal of Educational Technology **56**(1), 150–166 (2025)