LLMs in Short Answer Scoring: Limitations and Promise of Zero-Shot and Few-Shot Approaches

Imran Chamieh[1], Torsten Zesch[2], Klaus Giebermann[1]
[1]Hochschule Ruhr West, Germany,
[2]CATALPA, FernUniversität in Hagen, Germany

Abstract

This study investigates the potential of Large Language Models (LLMs), in particular GPT and LLaMA, for automated scoring of short answer responses. We focus on zero-shot and few-shot settings, but also compare with fine-tuned models and a supervised upper-bound. Our results show that LLMs perform much worse in those settings on a performance level that is not feasible for practical purposes. Fine-tuning LLMs brings their results on roughly the same level as supervised results, but as they are less efficient there currently seems to be no basis for applying LLMs for short answer scoring.

1 Introduction

The constantly increasing demand placed on educators in today's educational landscape requires innovative solutions to replace traditional assessment methods. Manual assessment, especially for large-scale exams, presents challenges for scalability, consistency and timely feedback to students Ramesh and Sanampudi (2022). Automated scoring has emerged as a potential solution, promising faster, more objective and feedback-rich assessments Galhardi and Brancher (2018).

Extensive research has explored automated scoring, but many systems require large amounts of training data to achieve reliable performance Patil and Adhiya (2022). Our focus is on finding a system that demonstrates strong performance across different datasets while minimizing the need for huge number of training examples. Large Language Models (LLMs) seem promising in this regard Naveed et al. (2023). Thus, in this paper, we explore LLMs performance in scoring open-ended student answers across three datasets. We compare two prominent LLMs, Generative Pre-trained Transformer (GPT) and Large Language Model Meta AI (LLaMA), under different training settings, including zero- and few-shot learning, and fine-tuning specifically applied to the GPT model. Additionally, we benchmark their performance against established baselines, specifically Google's pre-trained language model BERT Devlin et al. (2018) and classical SVM , known for its robustness in classification tasks Cortes and Vapnik (1995). This evaluation aims to deepen our understanding of how LLMs handle various assessment tasks and shed light on their potential to enhance automated scoring in education, particularly with limited training data.

2 Related Work

Very few studies have explored the performance of LLMs in zero- and few-shot settings within the context of automated scoring. Wu et al. (2023) introduced the Matching exemplars as Next Sentence Prediction (MeNSP) method, by employing a zero-shot prompt learning method using pre-trained language models. Their results indicate that few-shot learning offered limited improvement in performance.

Latif and Zhai (2024) compare the performance of a fine-tuned GPT-3.5 model with BERT and demonstrated that GPT-3.5 achieved higher scoring accuracy. It showed a remarkable average increase of 9.1% compared to BERT when applied to a single dataset of six assessment tasks. This finding emphasized the need for domain-specific fine-tuning LLMs to enhance their performance.

On the other hand, many studies investigated the neural networks and machine learning models to build scoring tools. Steimel and Riordan (2020) demonstrate how pretrained transformer models could be adapted for content scoring using an instance-based approach. By pooling token representations across all model layers, this approach achieved state-of-the-art performance on short answer scoring tasks. Bexte et al. (2023) conduct a comparison between instance-based and similarity-based methods on multiple datasets. They investigated the influence of different training set sizes

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 309–315
June 20, 2024 ©2024 Association for Computational Linguistics

on the performance of these methods using learning curve experiments. It found that a fine-tuned SBERT model does often yield the best results.

Overall, existing research offers limited insight into how LLMs perform in zero-shot and few-shot settings.

3 Experimental Setup

We tested the GPT family of models introduced by OpenAI, specifically GPT-3.5 and GPT-4.[1] Additionally, we tested Meta AI's LLaMA-2 models LLaMA-7b, LLaMA-13b, and LLaMA-70b [2]. Finally, Google's BERT model and SVM were included as baselines for comparison. For testing, we randomly selected 20% of each task from the datasets. We observed that LLMs usually produce in addition to the score an explanation, or repeat the scores of the given shots, rather than providing only the score, so we applied a filtering function that retrieves only the last integer of the LLMs response, and if no integer was found, we assigned a randomly generated number between 0 and the maximum possible score of the current task.

3.1 Datasets & Evaluation

We performed experiments on three widely used answer scoring datasets that are freely available.

ASAP Automated Student Assessment Prize[3] contains 10 prompts covering a broad range of disciplines. All answers were scored by two humans on a 0-2 or 0-3 scale depending on the task.

MindReading contains responses from children (ages 7-13) on questions from the Strange Story and Silent Film tasks, where answers scored on a 0-2 scale Kovatchev et al. (2020).

Powergrading is a short-answer dataset focused on knowledge about the United States for the citizenship exam. Answers are scored on a 0-1 scale Basu et al. (2013).

In this study, we differentiate between the terms 'Task' and 'Prompt'. 'Task' refers to a specific question from the datasets used. While,'Prompt' is a set of instructions designed for the LLM, including scoring guidelines, relevant context, and the student answer to be scored. For few-shot model, the prompt also includes randomly selected answer samples for each score within the task of the studied dataset.

For each task, we calculated Quadratic Weighted Kappa (QWK) Cohen (1968) as a standard metric used to quantify the agreement between machine scoring and human expert scoring. Finally, we averaged QWK scores across all tasks, for each dataset, to obtain a single overall performance metric.

3.2 Prompting

For the Powergrading dataset and 5 tasks within the ASAP dataset, we explored zero-shot performance. Note that zero-shot was not suitable for other ASAP tasks due to their reliance on long-form text or image data, nor for the MindReading dataset, where the questions are unavailable. To investigate the effectiveness of few-shot model for score prediction, we employed a variety of prompt designs and evaluated them on different numbers of shots. Initial testing (1, 3, 5, and 10 shots) with three prompt designs – Newline, Semicolon, and Space delimiters – revealed minimal variance in results, unlike to what Sclar et al. (2023) found (see Appendix B). Based on these results, we proceeded with the Newline delimiter prompt design for subsequent experiments from 0 to 10 shots, as it showed consistent performance across initial tests.

3.3 Fine-tuning

We extended our study to unveil the potential of the LLMs by fine-tuning a GPT-3.5 model. Fine-tuning involves adjusting the pre-trained model's parameters to adapt to specific characteristics of the task under study. For training phase we used 80% of the data to fine-tune GPT-3.5-turbo-1106.

4 Results & Discussion

Table 1 gives an overview of our results. The supervised system is a reference point that we use to compare zero-shot and few-shot result with.

4.1 Fine-tuning

Contrary to the results in Latif and Zhai (2024), which was conducted on a limited dataset, our results in Table 1 show that BERT actually scores slightly higher QWK over all datasets. This suggests a potential overfitting issue in GPT-3.5 model. In particular, in three tasks of Powergrading dataset, we observed that the model consistently scored all answers as 1. In general, fine-tuned results are in the same ballpark as supervised results, but computationally much more expensive.

[1] https://openai.com/

[2] https://ai.meta.com/

[3] https://www.kaggle.com/c/asap-sas

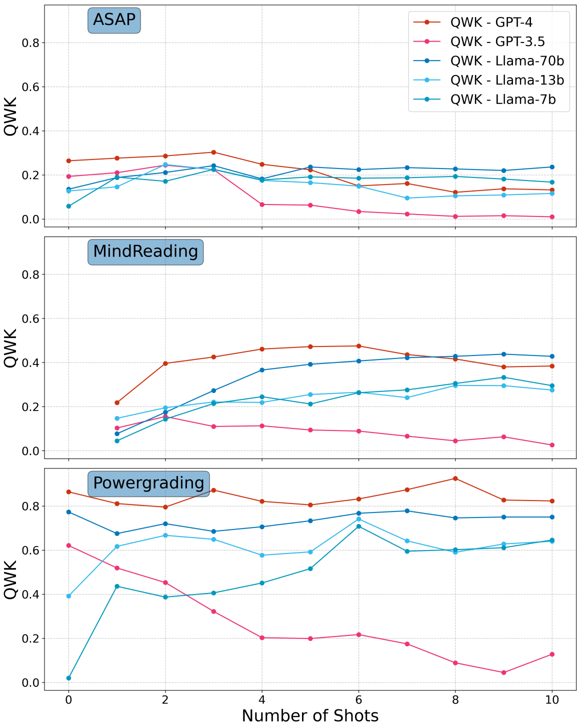

Figure 1: Impact of number of shots on scoring performance

4.2 Zero-shot

LLMs performance in zero-shot settings varied significantly across datasets. GPT-4 showed promising results on Powergrading dataset, while all models performed poorly on ASAP dataset. This suggests that LLMs are not yet mature enough for reliable zero-shot automated scoring. The good performance on Powergrading dataset can be attributed to the simplicity of the questions, which are related to USA citizenship test, and the scoring range (0,1). In contrast, even with detailed prompt and rubrics scoring instructions, LLMs struggled with ASAP dataset, indicating their limitations on tasks that require complicated reasoning or relay on domain-specific knowledge.

4.3 Few-shot

Our initial expectation was that incorporating few-shot into the prompt would enhance the model performance, as observed in the previous study Wu et al. (2023). However, our results indicate that only LLaMA models on Powergrading and MindReading datasets showed a slight improvement in performance with an increasing number of shots (up to 6 shots). In contrast, GPT-3.5 exhibited a weird behavior, with performance decreasing as the number of shots increasing, in particular on Powergrading dataset.

		QWK		
		ASAP	MR	PG
supervised	BERT	.74	.87	.94
	SVM	.46	.74	.80
fine-tuning	GPT-3.5	.61	.81	.83
0-shot	GPT-4	.26	.22	.86
	GPT-3.5	.19	.10	.62
	LLaMA-70b	.14	.08	.77
	LLaMA-13b	.13	.15	.39
	LLaMA-7b	.06	.05	.02
3-shot	GPT-4	.30	.43	.87
	GPT-3.5	.22	.11	.32
	LLaMA-70b	.24	.27	.69
	LLaMA-13b	.22	.22	.65
	LLaMA-7b	.23	.21	.41

Table 1: Comparison of model performance in terms of Quadratic Weighted Kappa (QWK)

The poor performance of LLMs in ASAP dataset is attributed to two key factors. First, answers in ASAP dataset tend to be longer compared to answers in other datasets, as shown in Figure 2 (see Appendix), where the average length of answers is approximately 50 words, so adding few-shot for each score increases the prompt size rapidly, which might badly affect the output. Additionally, the dataset's complexity, as questions heavily depend on domain-specific knowledge indicates challenges for general models in such domains. Similarly, in MindReading dataset, not only the questions are not available, but these questions are also derived from strange stories or silent films and they rely on specific knowledge that LLMs may not be trained on. On the other hand, the questions presented on Powergrading dataset related to general knowledge about USA, which made it easy for the LLMs to predict the scores which were limited to 0 and 1. Additionally, the short length of answers enables LLMs to effectively memorize it's task, making score prediction easier.

5 Conclusion

This study explores the potential of LLMs in automated scoring tasks, specifically zero- and few-shot, and fine-tuned settings across three diverse datasets.

Overall, our findings reveal strong performance from zero-shot and few-shot models on general knowledge. GPT-4 achieved performance very close to the upper bound BERT and outperformed SVM model. LLaMA models showed promising results; while not reaching GPT-4 levels, their performance remained consistent across different numbers of shots. In contrast, GPT-3.5 appeared overfitting as more shots introduced. This highlights the potential of few-shot LLMs for short answer scoring, especially on tasks involving general knowledge questions.

However, LLMs face challenges when confronted with tasks that require complicated reasoning or domain-specific knowledge, as noticed by their poor performance in ASAP dataset. The complicated nature of the questions in these subjects appears to cause difficulties for LLMs, highlighting the need for further improvements in dealing with nuanced and specialized content within educational datasets.

With regard to the fine-tuned model, our study revealed unexpected results as it failed to meet our performance expectations for automated scoring. It became clear that the model was overfitting at certain questions 'tasks', particularly noticeable when examining the performance of the Powergrading dataset.

Limitations

We only test commercial LLMs, but argue that open source LLMs would very likely yield even worse results. So the overall conclusion of the paper that LLMs are not yet ready to be used in zero-shot or few-shot settings for short answer scoring would stand unchanged. However, in future work we want to test a wider range of LLMs to gain further insights into their capabilities.

Ethical Considerations

LLMs are trained on large data sets that may contain unintentional biases, potentially leading to unfair scoring based. LLMs as black-box can lack transparency, making it difficult to provide an interpretation how they predict the scores. Another significant concern is student data privacy. If an LLM is hosted online, student answers are send to the provider and could end up in the model training data.

References

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2023. Similarity-based content scoring-a more classroom-suitable alternative to instance-based scoring? In *Findings of the association for computational linguistics: Acl 2023*, pages 1892–1903.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lucas Busatta Galhardi and Jacques Duílio Brancher. 2018. Machine learning approach for automatic short answer grading: A systematic review. In *Advances in Artificial Intelligence-IBERAMIA 2018: 16th Ibero-American Conference on AI, Trujillo, Peru, November 13-16, 2018, Proceedings 16*, pages 380–391. Springer.

Venelin Kovatchev, Phillip Smith, Mark Lee, Imogen Grumley Traynor, Irene Luque Aguilera, and Rory T Devine. 2020. " what is on your mind?" automated scoring of mindreading in childhood and early adolescence. *arXiv preprint arXiv:2011.08035*.

Ehsan Latif and Xiaoming Zhai. 2024. Fine-tuning chatgpt for automatic scoring. *Computers and Education: Artificial Intelligence*, page 100210.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Shweta Patil and Krishnakant P Adhiya. 2022. Automated evaluation of short answers: A systematic review. *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2021*, pages 953–963.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Kenneth Steimel and Brian Riordan. 2020. Towards instance-based content scoring with pre-trained transformer models. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34.

Xuansheng Wu, Xinyu He, Tianming Liu, Ninghao Liu, and Xiaoming Zhai. 2023. Matching exemplar as next sentence prediction (mensp): Zero-shot prompt learning for automatic scoring in science education. In *International Conference on Artificial Intelligence in Education*, pages 401–413. Springer.

A Models Hyperparameters

1. **SVM with TF-IDF vectorization:** We used a linear SVM model with a fixed Regularization parameter C=1.0 and utilized TF-IDF vectorization with a maximum vocabulary size of 1000 features.

2. **BERT:** We used the pre-trained BERT ("bert-base-uncased") model. Training data is processed with the BertTokenizerFast tokenizer and padded to a uniform length (512). we trained the model for 20 epochs with batches size = 8. After each epoch we run evaluation and kept the model with the lowest validation loss for evaluation on testing data.

3. **GPT:** For fine-tuning we utilized the OpenAI-recommended GPT-3.5-turbo-1106 model. Where as, GPT-4 is not yet available for fine-tuning. Training, validation, and test data were formatted in JSONL files as required. We employed the default values (auto) for learning rate, num_epochs, and batch_size.
   For few-shot experiments both GPT-3.5-turbo and GPT-4-turbo-preview were tested using the default parameters.

4. **LLaMA:** We utilized LLaMA-7b, LLaMA-13b, and LLaMA-70b for the few-shot model with the following parameter:
   temperature: 0.6 (Adjusts randomness of outputs. Higher values increase randomness, lower values promote determinism.). **top_p:** 0.9 (Controls text generation. Samples from the top 90% of most likely tokens, allowing for some variation.). **max_seq_len:** we choose different values between 512 and 2056, depending on the dataset and number of shots (It refers to the maximum length of input sequences the model can process.). **max_gen_len:** 5 as we want only the score. (It sets a limit on the maximum length of generated responses.). **max_batch_size:** int = 4

B Prompt designs

New line delimiter Evaluate student response to the United States Citizenship Exam. Return only the score, 1 if it is correct, and 0 if it is wrong. Question: What is one right or freedom from the First Amendment? (Return only the score):

Answer: the right to assemble -> Score: 1

Answer: freedom of speech -> Score: 1

Answer: freedom of religion.s -> Score: 1

Answer: right to pursue happiness. -> Score: 0

Answer: right to bear arms -> Score: 0

Answer: privacy -> Score: 0

Answer: free speech -> Score:

Semicolon delimiter Evaluate student response to the United States Citizenship Exam. Return only the score, 1 if it is correct, and 0 if it is wrong. Question: What is one right or freedom from the First Amendment? (Return only the score): Answer: the right to assemble -> Score: 1; Answer: freedom of speech -> Score: 1; Answer: freedom of religion.s -> Score: 1; Answer: right to pursue happiness. -> Score: 0; Answer: right to bear arms -> Score: 0; Answer: privacy -> Score: 0; Answer: free speech -> Score:

Space delimiter Evaluate student response to the United States Citizenship Exam. Return only the score, 1 if it is correct, and 0 if it is wrong. Question: What is one right or freedom from the First Amendment? (Return only the score): Answer: the right to assemble -> Score: 1 Answer: freedom of speech -> Score: 1 Answer: freedom of religion.s -> Score: 1 Answer: right to pursue happiness. -> Score: 0 Answer: right to bear arms -> Score: 0 Answer: privacy -> Score: 0 Answer: free speech -> Score:

	1 Shot		3 Shot		5 Shots		10 Shots	
Dataset	**qwk**	**Acc**	**qwk**	**Acc**	**qwk**	**Acc**	**qwk**	**Acc**
ASAP	**.152**	.417	.068	.319	.057	.287	.029	.234
	.077	.349	.051	.309	.034	.256	.010	.198
	.105	.355	**.083**	.343	**.058**	.288	**.037**	.223
Mindreading	.217	.516	.137	.419	**.133**	.400	.101	.403
	.237	.534	.128	.436	.122	.423	.104	.443
	.195	.500	**.139**	.423	**.133**	.420	**.108**	.428
Powergrading	**.656**	.909	**.077**	.500	.105	.767	**.156**	.793
	.403	.832	.058	.538	**.320**	.876	.098	.842
	.373	.843	.028	.507	.199	.838	.149	.813

Table 2: Impact of Delimiters (New Line, Semicolon, Space) of prompt on Accuracy and QWK using GPT-3.5

	1 Shot		3 Shots		5 Shots	
Dataset	**qwk**	**Acc**	**qwk**	**Acc**	**qwk**	**Acc**
ASAP	.244	.465	**.280**	.486	.220	.453
	.250	.460	.264	.482	.230	.487
	.244	.452	.250	.472	**.233**	.478
Mindreading	.133	.415	.408	.625	.448	.643
	.136	.416	**.447**	.655	**.507**	**.688**
	.129	.411	.409	.631	.476	.673
Powergrading	**.788**	.941	**.798**	.947	.777	.934
	.774	.940	.794	.941	.749	.922
	.781	.945	.794	.946	**.786**	.927

Table 3: Impact of Delimiters (New Line, Semicolon, Space) of prompt on Accuracy and QWK using GPT-4

	1 Shots		3 Shots		5 Shots		10 Shots	
Dataset	**qwk**	**Acc**	**qwk**	**Acc**	**qwk**	**Acc**	**qwk**	**Acc**
ASAP	**.190**	.465	**.225**	.503	**.190**	.483	**.168**	.455
	.154	.448	.195	.463	.137	.414	.112	.400
	.161	.462	.219	.477	.157	.448	.137	.429
Mindreading	.043	.190	.206	.523	.213	.517	**.296**	.565
	.094	.455	**.252**	.578	.193	.541	.212	.487
	.082	.436	.236	.569	**.274**	.588	.289	.540
Powergrading	.324	.698	.390	.709	**.500**	.829	**.646**	.901
	.278	.751	**.460**	.808	.461	.817	.583	.872
	.334	.732	.452	.806	**.500**	.801	.572	.866

Table 4: Impact of Delimiters (New Line, Semicolon, Space) of prompt on Accuracy and QWK using LLaMA2-7b-chat

	1 Shots		3 Shots		5 Shots		10 Shots	
Dataset	**qwk**	**Acc**	**qwk**	**Acc**	**qwk**	**Acc**	**qwk**	**Acc**
ASAP	.148	.428	**.237**	.518	.180	.470	**.132**	.430
	.200	.477	.186	.463	.123	.420	.090	.371
	.191	.470	.166	.461	**.184**	.469	.130	.416
Mindreading	.097	.367	**.221**	.505	**.255**	.528	**.285**	.562
	.110	.418	.217	.512	.219	.520	.219	.502
	.097	.395	.198	.476	.215	.491	.282	.557
Powergrading	.541	.848	.547	.818	.573	.862	.656	.884
	.620	.875	**.582**	.828	**.579**	.842	**.712**	.827
	.558	.835	.518	.826	.617	.872	.631	.875

Table 5: Impact of Delimiters (New Line, Semicolon, Space) of prompt on Accuracy and QWK using LLaMA2-13b

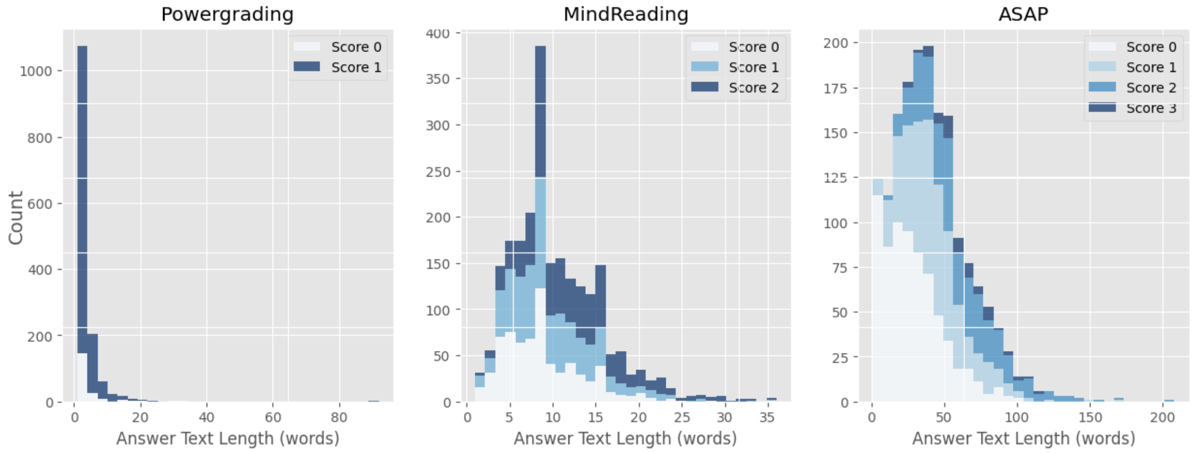

Figure 2: Length distribution of answers