Comparing Few-Shot Prompting of GPT-4 LLMs with BERT Classifiers for Open-Response Assessment in Tutor Equity Training

Sanjit Kakarla   SANJIT.KAKARLA@GMAIL.COM
Conrad Borchers   CBORCHER@CS.CMU.EDU
Danielle R. Thomas   DRTHOMAS@CMU.EDU
Shambhavi Bhushan   SHAMBHAB@ANDREW.CMU.EDU
Kenneth R. Koedinger   KOEDINGER@CMU.EDU
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213, USA

Abstract

Assessing learners in ill-defined domains, such as scenario-based human tutoring training, is an area of limited research. Equity training requires a nuanced understanding of context, but do contemporary large language models (LLMs) have a knowledge base that can navigate these nuances? Legacy transformer models like BERT, in contrast, have less real-world knowledge but can be more easily fine-tuned than commercial LLMs. Here, we study whether fine-tuning BERT on human annotations outperforms state-of-the-art LLMs (GPT-4o and GPT-4-Turbo) with few-shot prompting and instruction. We evaluate performance on four prediction tasks involving generating and explaining open-ended responses in advocacy-focused training lessons in a higher education student population learning to become middle school tutors. Leveraging a dataset of 243 human-annotated open responses from tutor training lessons, we find that BERT demonstrates superior performance using an offline fine-tuning approach, which is more resource-efficient than commercial GPT models. We conclude that contemporary GPT models may not adequately capture nuanced response patterns, especially in complex tasks requiring explanation. This work advances the understanding of AI-driven learner evaluation under the lens of fine-tuning versus few-shot prompting on the nuanced task of equity training, contributing to more effective training solutions and assisting practitioners in choosing adequate assessment methods.

Keywords: open-response grading; feedback; equity training; LLMs; BERT; GPT-4

1. Introduction

Assessing learners' responses to open-ended questions in ill-defined domains is an area of ongoing research and a novel educational task for language models. In education, ill-defined domains include situational judgment testing, involving assessing complex scenarios that lack clear right or wrong answers. These domains are constructs where the "correct" answer to a question is not well-defined and often require nuanced reasoning and decision-making, such as identifying the most appropriate response to students experiencing inequities or motivational challenges (Thomas et al., 2023). Traditional assessment methods often involve humans scoring responses based on complex rubrics given the context-dependent nature of

such tasks, creating a need for novel scalable, automated approaches to evaluation (Vercellotti, 2021).

Advancements in natural language processing (NLP) offer promising solutions for automating assessments in these contexts. Earlier-generation models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), RoBERTa (Liao et al., 2021), and T5 (Carmo et al., 2020), provide efficient, low-cost, and computationally resource-light options, while large language models (LLMs) like GPT-4 models and Gemini excel in capturing contextual nuances but are comparatively resource-intensive (Brown, 2020). As state-of-the-art LLMs, OpenAI GPT models have access to a wide range of real-world knowledge represented in their training corpus, which could show satisfactory accuracy with human-like reasoning abilities when prompted with few examples (Brown, 2020).

Despite the promise of LLMs, their performance and cost-effectiveness in real-world educational contexts remain underexplored, particularly in tasks requiring subtle interpretation of human responses. Legacy models like BERT, on the other hand, are more limited in real-world knowledge but can be fine-tuned offline based on a small set of human-annotated examples, being comparatively computationally lightweight and open-source, meaning that BERT models can be downloaded, inspected, and used by researchers and practitioners with stable versioning. This study addresses these gaps by comparing the performance of BERT, GPT-4o, and GPT-4-turbo in assessing responses to situational challenges in advocacy-focused training lessons. GPT-4.o was used as it is substantially more cost-effective compared to GPT-4, although we aimed to examine both these models on their human-like reasoning abilities within the ill-defined domain of equity training. Using a dataset of 291 lesson completions from 243 college students learning to become middle school tutors, we evaluate these models on their ability to assess their performance by capturing the nuances of open-ended responses. This research contributes to understanding how AI tools can enhance evaluation in ill-defined domains, offering practical insights into selecting appropriate models for diverse educational applications. To this end, we investigate the following research question:

RQ1: What is the comparative performance between legacy natural language processing models such as BERT and contemporary models like GPT-4o and GPT-turbo in assessing learners' open responses within scenario-based equity training?

2. Related Work

2.1. Smaller-scale Early-generation Language Models in Education

Early-generation language models have been extensively used in evaluating learner performance through open response grading tasks, evaluating responses that do not have predefined answers. These models have focused on prioritizing efficiency and task-specific optimization after learning from vast pools of training data from the open web (Kaddoura et al., 2022). BERT has played a key role in applying NLP to assessment. BERT has shown its capacity to analyze open-ended responses and other unstructured forms of data due to its bidirectional text understanding, gaining an understanding of all the words in a text body, rather than only analyzing text backward as N-Gram models do (Devlin et al., 2018). BERT models have excelled in assessment tasks such as automated grading in science education, where they have been fine-tuned to evaluate student understanding of scientific

phenomena and the application of experimental principles (Zhai, 2024). Variant models including RoBERTa, XLnet, and T5 have offered incrementally improved performance (Liao et al., 2021; Carmo et al., 2020). Employing these machine learning-based models in the process of assessing human responses saves time and reduces costs (Zhai et al., 2020). Due to increasing portability to be deployed on different computational resources, these models have become more suitable for local deployments, making them ideal for use in environments with limited educational infrastructure. For instance, T5 has allowed NLP tasks to evolve into a text-to-text framework, simplifying its use in assessing tutor responses (Bird et al., 2023).

2.2. LLMs in Education

Large-language models include more parameters than traditional transformer-based models like BERT, allowing for substantially increased model training and a knowledge base from its large training corpus to grade responses. LLMs such as GPT-4 have strong reasoning abilities, which hold potential for use in contextual judgment tasks previously only accessible to humans (Kalyan, 2023). Emerging evidence points to these LLM reasoning abilities allowing for handling nuanced educational tasks such as real-time feedback generation and timely tutor evaluation of specific competencies such as reacting to student errors (Lin et al., 2024; Kakarla et al., 2024). Prompt engineering strategies — the ability to craft prompts for generating adequate responses for tasks — can enhance additional flexibility within LLM responses. These improvements, however, come at the costs of computationally heavy training, API use costs, changing model versions over time, and external model deployment issues, such as temporary downtime. Therefore, we are comparing the lightweight, legacy BERT model with GPT models by OpenAI.

3. Method

Our dataset comprised open-response text responses of undergraduate tutors in the Northeastern United States, employed as paid tutors for a remote tutoring organization aiding middle-school students with corresponding human-labeled scores from two lessons: *Avoiding Unconscious Assumptions* (N = 79 human-labeled responses) and *Helping Students Manage Inequity* (N = 164 human-labeled responses). These lessons and their questions were created by a university research team aligned with research-supported, real-world competencies of effective tutor equity training (Thomas et al., 2024). The lesson was delivered to tutors via an online learning platform. Label imbalance in these annotated samples ranged from 41% to 43% across prediction tasks, which is moderate, with the exception of generating dialog in the Helping Students Manage Inequity lesson, standing at 85%, which we address through stratification of cross-validation folds. Both sample sizes were determined based on convenience and data availability at the time of writing this manuscript. For each lesson, tutors are tasked to predict and explain the appropriate research-supported approach. Explained responses require learners to explain why they would say something to a specific student, while predicted responses require learners to generate open-ended text to represent real dialogue they would have with students during tutoring. The predict question for a sample scenario within the *Helping Students Manage Inequity* lesson is shown in Figure 1. The Explain question, shown in Figure 2, tasks tutors to explain the rationale behind their response.

Each response was graded by human coders in a binary fashion, meaning that the response was either correct and adequate or incorrect and inadequate. A sample response for the situation depicted in Figure 1 on responding to student inequity is *"I would provide him with other possible ways to access the internet that helps him do the homework equally. For example, he could do the homework at school."* Human coders graded this response as incorrect, given that it is not mindful of inequitable access differences to resources (see digital appendix linked below in this section for full rubrics). Two human coders achieved an inter-rater reliability of Cohen's κ of 0.88 and 0.87 for the *Avoiding Unconscious Assumptions* lesson (based on 79 tutor responses) and 0.86 and 0.93 for the *Helping Students Manage Inequity* lesson (based on 164 tutor responses), which indicates excellent reliability (Warrens, 2015).

Figure 1: Predict Scenario for the *Helping Students Manage Inequity* Lesson.

Figure 2: Explain Scenario for the *Helping Students Manage Inequity* Lesson

For evaluation, human-labeled data were split into training and validation subsets according to stratified 5-fold cross-validation. Given the small overall sample size, no hold-out test set evaluation was performed. Rather, we report model performance averages across cross-validation folds.

Responses were tokenized using the BERT Wordpiece tokenizer with a maximum token length of 256. The tokenization process included truncation and padding to standardize input sizes. The BERT model was initialized using the standard `bert-base-uncased` pretrained weights and fine-tuned for classification tasks. The training utilized the AdamW optimizer with a learning rate of 0.00002. Batches were processed using a batch size of 16, and the model was trained for five epochs per fold based on monitoring training loss.

Predictions from two GPT-based models (GPT-4 Turbo and GPT-4o) were included for each response and evaluated in the same manner and on the same observations as the BERT model. To generate GPT-based predictions, we engineered a prompt with 2-3 few-shot examples of how each category response should look. Based on these specifications the models were instructed to output a 0 or 1 depending on whether they believed a given learner response was adequate or not given these criteria. Unlike the BERT classifier, this approach does not allow for an output of a probability. Hence, AUC (area under the ROC-curve) could not be defined for these models. Instead, for all models we computed, the average accuracy and $F1$ score for each prediction task across folds. Finally, for each metric and model, we created an average across all prediction tasks (average across all folds) to comprehensively compare each model.

The human annotation rubrics, dataset, lesson content, generative AI model prompts, and code for the BERT model, a contribution of this work, can be found at the following GitHub repository: https://github.com/conradborchers/bert-llm-open-response.

4. Results

RQ1 related to the comparative performance of BERT, GPT-4o, and GPT-turbo in assessing learners' open responses within scenario-based equity training. The results presented in Table 1 demonstrate that the BERT model, fine-tuned on about 100 examples, consistently outperformed both few-shot prompted GPT models across all four prediction tasks. The Area Under the Curve (AUC) scores generally fell within a satisfactory to excellent range, spanning approximately 0.8 to 0.9. In contrast, the accuracy of the GPT models proved less consistent, with GPT-4.0 achieving higher accuracy in three out of four prediction tasks.

Table 1: Comparative Performance between GPT and BERT models.

Model	Task	Accuracy	F1	AUC
BERT	Explain - *Avoiding Unconscious Assumptions*	0.854	0.903	0.8779
GPT-4o	Explain - *Avoiding Unconscious Assumptions*	0.550	0.523	n/a
GPT-4 Turbo	Explain - *Avoiding Unconscious Assumptions*	0.798	0.805	n/a
BERT	Predict - *Avoiding Unconscious Assumptions*	0.720	0.797	0.7936
GPT-4o	Predict - *Avoiding Unconscious Assumptions*	0.754	0.755	n/a
GPT-4 Turbo	Predict - *Avoiding Unconscious Assumptions*	0.506	0.480	n/a
BERT	Explain - *Helping Students Manage Inequity*	0.896	0.909	0.9527
GPT-4o	Explain - *Helping Students Manage Inequity*	0.868	0.866	n/a
GPT-4 Turbo	Explain - *Helping Students Manage Inequity*	0.598	0.526	n/a
BERT	Predict - *Helping Students Manage Inequity*	0.880	0.935	0.8573
GPT-4o	Predict - *Helping Students Manage Inequity*	0.733	0.737	n/a
GPT-4 Turbo	Predict - *Helping Students Manage Inequity*	0.293	0.390	n/a
BERT	Average Across Tasks	0.837	0.886	0.8704
GPT-4o	Average Across Tasks	0.726	0.720	n/a
GPT-4 Turbo	Average Across Tasks	0.549	0.550	n/a

$F1$ scores indicated moderate to satisfactory performance for GPT-4.0, except for the task related to explaining how to avoid unconscious assumptions. Overall, few-shot prompting with the newer and commercial GPT architectures did not surpass the performance of

the resource-efficient and offline BERT fine-tuning approach on these four prediction tasks.

5. Discussion

Advances in LLMs promise to open the door to assessing increasingly complex and nuanced open-response assessment tasks in education. However, our results showcase that fine-tuning early-generation models such as BERT not only led to superior accuracy than few-shot LLM prompting in the ill-defined domain of equity training, but is also more resource-efficient.

5.1. Efficiency and Resource Utilization

The average performance of BERT at 0.837 was higher than both GPT models at 0.726 and 0.549 respectively. BERT demonstrated competitive performance using an offline approach that is more resource-efficient compared to deploying commercial GPT models. Moreover, BERT might be more stable in performance, too, because it does not face the continuous new model versioning often common for closed-source models such as GPT, which may alter model behavior and performance on specific tasks.

The lower performance of the GPT models across both tasks could be attributed to these models primarily relying on few-shot prompt engineering techniques to assess responses. This approach does reduce the need for constant retraining; however, it may struggle to adequately capture nuanced response patterns, especially in complex, multi-faceted tasks where grading is often performed according to specific rubrics. In this task, GPT's reliance on few-shot prompting relies heavily on the ability of the model to generalize given a few examples, which can increase variability within the responses in ill-defined domains.

5.2. Limitations

The lack of probability outputs from GPT models limited metric diversity, excluding AUC as a measure for GPT models as this metric relies on probabilities, not binarized predictions. Future research may use LLM embedding-based prediction models to circumvent this issue (Zhang et al., 2024), and consider using a larger set of open-source language models compared to the LLMs used here.

Another limitation is related to our sample size, which constrained our ability to evaluate model performance on hold-out test sets beyond cross-validation. Such evaluations may pose a better estimate of the generalizability of our classifiers. Our work studies higher education student populations completing equity training, however, an extended tutor sample size in different cultural and instructional contexts (i.e., hybrid learning) and more diverse samples could allow us to make more robust conclusions for this form of the classification task.

5.3. Implications for Tutoring Systems

BERT's robust performance suggests its potential for integration into assessment and learning systems for automated evaluation of learner responses at scale. Applications of such assessment models are manifold but include real-time evaluation of tutorial dialog during problem-solving practice, which may enhance learning (Borchers et al., 2024). The ability of these models to grade open responses underscores their potential for use in challenging educational domains, such as equity training studied here.

GPT models, while promising in some tasks, may require additional adaptation or fine-tuning to improve consistency and task-specific performance compared to relying solely on few-shot examples and explanations that carry significant variability.

6. Future Work and Conclusion

This study put the abilities of contemporary LLMs to the test of understanding nuanced situations of reacting to students during tutoring in an equitable fashion. Our findings suggest that the assessment task studied here requires fine-tuning beyond few-shot prompting, whereby fine-tuned BERT demonstrated superior performance across all prediction tasks across the two lessons. This performance emphasizes the value of fine-tuning task-specific datasets, even with small sample sizes. These results underscore the limitations of few-shot prompting for GPT in handling tasks within ill-defined educational domains that require nuanced understanding and contextual reasoning, such as equity training. Resource-efficient, offline methods like BERT fine-tuning, on the other hand, offer a viable, open-source, and cost-efficient alternative to commercial AI models for educational applications.

Future work should explore increasing dataset sizes, enhancing GPT model prompt engineering approaches, exploring costs related to different model sizes and computation, and evaluating other fine-tuning strategies (such as fine-tuning GPT) to improve model performance on complex tutoring tasks. Our approach comparing GPT performance to BERT can also be extended to various other lessons on tutoring competencies, such as addressing microaggressions and narrowing opportunity gaps along with other educational domains beyond equity training. Our open availability of study materials, data, and code will facilitate replication, transparency, and future advancements in automated assessment of open-responses in scenario-based training.

Acknowledgments

References

Jordan J Bird, Anikó Ekárt, and Diego R Faria. Chatbot interaction with artificial intelligence: human data augmentation with t5 and language transformer ensemble for text classification. *Journal of Ambient Intelligence and Humanized Computing*, 14(4): 3129–3144, 2023.

Conrad Borchers, Kexin Yang, Jionghao Lin, Nikol Rummel, Kenneth R Koedinger, and Vincent Aleven. Combining dialog acts and skill modeling: What chat interactions enhance learning rates during ai-supported peer tutoring? In *Proceedings of the 17th International Conference on Educational Data Mining*, 2024.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Bidirectional encoder representations from transformers. *arXiv preprint arXiv:1810.04805*, page 15, 2018.

Sanaa Kaddoura, Daniela Elena Popescu, and Jude D Hemanth. A systematic review on machine learning models for online learning and examination systems. *PeerJ Computer Science*, 8:e986, 2022.

Sanjit Kakarla, Danielle R Thomas, Jionghao Lin, Shivang Gupta, and Kenneth R Koedinger. Using large language models to assess tutors' performance in reacting to students making math errors. *arXiv preprint arXiv:2401.03238*, 2024.

Katikapalli Subramanyam Kalyan. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, page 100048, 2023.

Wenxiong Liao, Bi Zeng, Xiuwen Yin, and Pengfei Wei. An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta. *Applied Intelligence*, 51:3522–3533, 2021.

Jionghao Lin, Zifei Han, Danielle R Thomas, Ashish Gurung, Shivang Gupta, Vincent Aleven, and Kenneth R Koedinger. How can i get it right? using gpt to rephrase incorrect trainee responses. *International Journal of Artificial Intelligence in Education*, pages 1–27, 2024.

Danielle R Thomas, Xinyu Yang, Shivang Gupta, Adetunji Adeniran, Elizabeth Mclaughlin, and Kenneth Koedinger. When the tutor becomes the student: Design and evaluation of efficient scenario-based lessons for tutors. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 250–261, 2023.

Danielle R Thomas, Conrad Borchers, Sanjit Kakarla, Jionghao Lin, Shambhavi Bhushan, Boyuan Guo, Erin Gatz, and Kenneth R Koedinger. Do tutors learn from equity training and can generative ai assess it? *arXiv preprint arXiv:2412.11255*, 2024.

MaryLou Vercellotti. Beyond the rubric: Classroom assessment tools and assessment practice. *Tesl-ej*, 25(3):n3, 2021.

Matthijs J Warrens. Five ways to look at cohen's kappa. *Journal of Psychology & Psychotherapy*, 5, 2015.

Xiaoming Zhai. Ai and machine learning for next generation sci-ence assessments. *Machine Learning, Natural Language Processing, and Psychometrics*, page 201, 2024.

Xiaoming Zhai, Yue Yin, James W Pellegrino, Kevin C Haudek, and Lehong Shi. Applying machine learning in science assessment: a systematic review. *Studies in Science Education*, 56(1):111–151, 2020.

Jiayi Zhang, Conrad Borchers, Vincent Aleven, and Ryan S Baker. Using large language models to detect self-regulated learning in think-aloud protocols. In *Proceedings of the 17th International Conference on Educational Data Mining (EDM)*, 2024.