LexiLogic at BEA 2025 Shared Task: Fine-tuning Transformer Language Models for the Pedagogical Skill Evaluation of LLM-based tutors

Souvik Bhattacharyya, Billodal Roy, Niranjan Kumar M, Pranav Gupta
Lowe's
Correspondence: {souvik.bhattacharyya, billodal.roy, niranjan.k.m, pranav.gupta}@lowes.com

Abstract

While large language models show promise as AI tutors, evaluating their pedagogical capabilities remains challenging. In this paper, we, team LexiLogic presents our participation in the BEA 2025 shared task on evaluating AI tutors across five dimensions: Mistake Identification, Mistake Location, Providing Guidance, Actionability, and Tutor Identification. We approach all tracks as classification tasks using fine-tuned transformer models on a dataset of 300 educational dialogues between a student and a tutor in the mathematical domain. Our results show varying performance across tracks, with macro average F1 scores ranging from 0.47 to 0.82, achieving rankings between 4th and 31st place. Such models have the potential to be used in developing automated scoring metrics for assessing the pedagogical skills of AI math tutors.

1 Introduction

While significant progress has been made in making today's large language models helpful, aligned, and responsible (Tan et al., 2023; Ji et al., 2023; Feng et al., 2024), their full potential in academic settings remains underutilized. Despite growing interest in using LLM-based AI tutors for academic support, traditional evaluation benchmarks tend to focus more on knowledge, factual accuracy, and reasoning (DeepSeek-AI et al., 2025; Abdin et al., 2025) rather than on the ability of these dialogue systems to function effectively in the role of a tutor. In educational contexts, there is a pressing need for systems and evaluation metrics specifically designed to assess complex pedagogical qualities. Therefore, it is essential to not only develop intelligent tutoring systems but also to evaluate them in terms of their ability to provide sufficient, helpful, and factually accurate guidance.

The shared task organized as part of the BEA workshop (Kochmar et al., 2025) focuses on educational dialogues between a student and a tutor in the mathematical domain, specifically addressing student mistakes or confusion. The goal of the AI tutor is to help remediate these issues. The tutor responses, generated by the task organizers, come from a range of state-of-the-art LLMs with varying sizes and capabilities, including GPT-4 (OpenAI et al., 2023), Gemini (Reid et al., 2024), Sonnet (Anthropic, 2025), Mistral (Jiang et al., 2023), Llama 3.1 (Grattafiori et al., 2024) and Phi-3 (Abdin et al., 2024a). In addition to the generated responses, the development set includes annotations evaluating their quality across several pedagogically motivated dimensions: Mistake Identification, Mistake Location, Providing Guidance, Actionability, and Tutor Identification.

Across all tracks of the shared task, we approached the problems as classification tasks and followed a fine-tuning approach using several transformer-based encoder and decoder models. Table 1 summarizes the performance of our submitted models compared to the top-performing entries in terms of macro average F1 score in each task.[1]

Track	Our Score	Best Score
Track 1	0.65	0.72
Track 2	0.48	0.60
Track 3	0.47	0.58
Track 4	0.69	0.71
Track 5	0.82	0.95

Table 1: Performance of our models compared to the best scores in each track.

2 Related Work

With the widespread use of large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023; Team et al., 2025) as conversational systems in educational contexts, several studies have evaluated

[1]The code for this work is available at https://github.com/prannerta100/acl-bea2025-workshop-st

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1180–1186

their pedagogical capabilities. There are numerous LLM evaluation metrics such as BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020), ROUGE (Lin, 2004), DialogRPT (Gao et al., 2020), etc., which are not necessarily designed to assess an LLM's educational or pedagogy-related capabilities (Jurenka et al., 2024) and shown to have relatively low correlation with human judgments (Liu et al., 2023). This highlights the need for alternative methods to evaluate LLM performance in educational settings. One such approach is to use human annotators to rate LLM responses based on various criteria (Collins et al., 2023; Shen and Wu, 2023; Lee et al., 2024). While human evaluators can consider context, tone, and pedagogical effectiveness, offering qualitative insights that go beyond quantitative metrics, they are also prone to bias, and the process tends to be time-consuming and relatively expensive.

At the other end of the spectrum, there is growing interest in automated evaluation systems and LLM-as-a-judge approaches (Jurenka et al., 2024). Chen et al. (2023)'s experimental results show that ChatGPT is capable of evaluating text quality effectively from various perspectives without reference, and it demonstrates superior performance compared to most existing automatic metrics. Macina et al. (2025) developed `MATHTUTORBENCH` to score the pedagogical quality of open-ended teacher responses and also trained several LLM-based reward models, showing that these models can distinguish expert from novice teacher responses with high accuracy. `TUTOREVAL`, a diverse question-answering benchmark, was released by Chevalier et al. (2024), who evaluated the capabilities of several open-weight and proprietary LLMs using GPT-4 as the evaluator. Maurya et al. (2025a) introduced `MRBench`, which includes a large set of student–tutor conversations from seven state-of-the-art LLM-based and human tutors, and evaluated them across various dimensions using a different set of LLMs. Jurenka et al. (2024) also introduced `LearnLM-Tutor`, a fine-tuned model that was consistently preferred over base models for various academic tasks as judged by LLM-based critics.

3 Task Description and Methodology

The dataset provided for the shared task (Maurya et al., 2025b) consisted of conversation history between a tutor and a student along with a final response from the tutor based on which the various pedagogical capability label is to be predicted. There were a total of 300 distinct conversations out of which we chose 50 to include in our test set, which resulted in 2067 training data points and 409 test data points. The same train-test split is used in all our experiments.

3.1 Track 1 - Mistake Identification

Track 1 of the shared task aims to develop systems that can identify whether a tutor's response acknowledges mistakes in a student's answer. The distribution of three categories in this track is detailed in Table 2. Each data point consists of a conversation history between a tutor and a student, along with a final response from the tutor. Participants are required to assess whether the tutor's reply explicitly recognizes the student's mistake within the conversation.

Tutor	Yes	No	To some extent
GPT4	234	15	1
Gemini	215	21	14
Sonnet	212	20	18
Phi-3	68	176	6
Mistral	223	10	17
Llama318B	202	31	17
Llama31405B	239	7	4
Expert	188	15	47
Novice	28	11	28
Total	1609	306	152

Table 2: Distribution of instances across categories for each tutor in the dataset in Track 1

For this task, our experiments involved fine-tuning various encoder and decoder models. The input sequence was formed by concatenating the conversation history with the final response, and we replaced the model's un-embedding layer with a classification head for the three target classes. The models we used included FLAN-T5 (Chung et al., 2022), ModernBert (Warner et al., 2024), Llama 3.2 (Grattafiori et al., 2024), Phi-4 (Abdin et al., 2024b), and Qwen-2.5 (Qwen et al., 2025). All models were trained for 10–15 epochs with an initial learning rate between 5e-5 and 1e-4, using an exponential learning rate scheduler, a batch size of 8–10, and a gradient accumulation step of 2. On the test set, `FLAN-T5-large` performed the best, achieving a macro average F1 score of 0.65 and placing us 22nd among 44 submissions on the official leaderboard. The training and test set performance of all models is presented in Table 3 (with

the train set F1 scores corresponding to the epoch with the highest test set performance).

Model	Train F1	Test F1
FLAN-T5-large	0.94	**0.65**
ModernBERT-large	0.98	0.61
Llama-3.2-3B	0.99	0.62
Phi-4-mini-instruct	1.0	0.63
Qwen2.5-7B-Instruct	0.73	0.55

Table 3: Strict macro average F1 scores of different models on training and test datasets of Track 1

3.2 Track 2 - Mistake Location

In subtask 2, the objective is to develop a system capable of identifying whether a tutor's response effectively locates the mistake in the student's answer and provides a clear explanation of the error. This includes assessing whether the tutors' responses accurately point to a genuine mistake and its location in the students' responses. The distribution of each labels across different categories for each tutor in the training dataset is shown in Table 4.

Tutor	Yes	No	To some extent
GPT4	242	37	13
Gemini	176	93	31
Sonnet	207	60	33
Phi-3	73	223	4
Mistral	216	52	35
Llama318B	161	108	31
Llama31405B	252	33	15
Expert	197	58	45
Novice	19	60	2
Total	1543	724	209

Table 4: Distribution of instances across categories for each tutor in the dataset for Track 2

The final response is concatenated with the conversation history and fed as input into our model. Our experimental setup predominantly focused on transformer-based encoder and decoder models. In both encoder-decoder and large language model (LLM) configurations, we modify the original models by removing the final un-embedding layer and replacing it with a classification head. Among the encoder-based models, we evaluated ModernBert (Warner et al., 2024), and MathBERT (Peng et al., 2021). For large language models, we conducted experiments with Llama 3.2 (Grattafiori et al., 2024), Phi-4 (Abdin et al., 2024b), and Qwen-2.5 (Qwen et al., 2025).

We fine-tuned all models for a maximum of 10 epochs, with an initial learning rate in the range of 2e-2 to 5e-5, an exponential learning rate scheduler with gamma set between 0.9 and 0.9375 with a batch size between 4 and 12, with gradient accumulation steps set to 2. During training we minimized the categorical cross-entropy loss. In Table 5, we report the strict Macro average F1 scores of various models. The reported training set F1 scores correspond to the epoch with the highest F1 score on the test set. On the held-out test set, our submission based on `Phi-4-mini-instruct` achieved an F1 score of 0.48 on the unseen test dataset placing us at the 23rd position out of total 31 submissions.

Model	Train F1	Test F1
MathBERT	0.67	0.5
ModernBERT-large	0.72	0.52
Llama-3.2-3B	0.73	0.55
Llama-3-8B	0.71	0.53
Phi-4-mini-instruct	0.78	**0.68**
Qwen2.5-7B-Instruct	0.67	0.55

Table 5: Strict Macro average F1 scores of different models on training and test datasets

3.3 Track 3 - Providing Guidance

Track 3 focuses on evaluating whether a tutor's response provides effective guidance to help students understand and correct their mistakes. This task goes beyond simply identifying and locating errors to assess the pedagogical quality of the tutoring response. The system must determine if the tutor offers constructive feedback, explanations, or suggestions that would help the student learn from their mistakes. Similar to the previous tracks, the task includes three categories: 'Yes', 'No', and 'To some extent', with their distribution across different tutors shown in Table 6.

Our approach for this track followed a similar methodology to the previous tasks, where we concatenated the conversation history with the final tutor response and fed it as input to our classification models. The experimental setup involved fine-tuning various transformer-based models to classify the quality of guidance provided in tutor responses.

We evaluated several model architectures including both encoder-only and decoder-only models. Among the encoder-based models, we experimented with ModernBERT (Warner et al., 2024),

Tutor	Yes	No	To some extent
GPT4	228	41	31
Gemini	168	47	85
Sonnet	184	52	64
Phi-3	51	189	60
Mistral	189	47	64
Llama318B	134	65	101
Llama31405B	238	16	46
Expert	205	47	48
Novice	10	62	4
Total	1407	566	503

Table 6: Distribution of instances across categories for each tutor in the dataset for Track 3

while for large language models, Phi-4 (Abdin et al., 2024b), and FLAN-T5 (Chung et al., 2022). All models were modified by replacing the final un-embedding layer with a three-way classification head corresponding to our target categories. The train and test F1 values are in Table 7.

The training configuration involved fine-tuning for 8-12 epochs with learning rates ranging from 1e-5 to 8e-5, using an exponential learning rate scheduler with gamma values between 0.85 and 0.95. We employed batch sizes of 6-14 with gradient accumulation steps of 2, and optimized using categorical cross-entropy loss. The performance of different models on both training and test sets is presented in Table 7, where the training F1 scores correspond to the epoch achieving the highest test set performance.

Model	Train F1	Test F1
FLAN-T5-large	0.92	0.36
ModernBERT-large	0.89	0.39
Phi-4-mini-instruct	0.97	**0.45**

Table 7: Strict Macro average F1 scores of different models on training and test datasets for Track 3

Our best performing model, `Phi-4-mini-instruct`, achieved a macro average F1 score of 0.47 on the test set, securing the 31st position out of 35 total submissions on the official leaderboard. The relatively lower performance across all models suggests that evaluating the quality of pedagogical guidance is inherently more challenging than simple mistake identification, as it requires understanding the nuanced aspects of effective tutoring strategies and educational support.

3.4 Track 4 - Actionability

In Track 4, the goal is to develop system to identify whether the tutor's response is clear in regards to what the student should do next, i.e., whether or not the tutor response was vague, unclear or a conversation stopper. Table 8 shows the distribution of instances across different categories for each tutor in the training dataset provided.

Tutor	Yes	No	To some extent
GPT4	116	125	9
Gemini	142	52	56
Sonnet	141	74	35
Phi-3	27	215	8
Mistral	168	43	39
Llama318B	106	93	51
Llama31405B	182	40	28
Expert	200	18	32
Novice	3	52	12
Total	1085	673	309

Table 8: Distribution of instances across categories for each tutor in the dataset in Track 4

We use as an input the sequence of tokens after the final response from the tutor is appended with the original conversation. We experimented with multiple transformer based encoder and decoder models in this task as well. In all the experiments, we remove the final un-embedding layer from the original models and replace it with a classification head producing three dimensional logits corresponding to the three available classes in this task. Among the encoder models we have experimented with FLAN-T5 (Chung et al., 2022), ModernBert (Warner et al., 2024) and MathBERT (Peng et al., 2021) and among the LLMs we tried Llama 3.2 (Grattafiori et al., 2024), Phi-4 (Abdin et al., 2024b) and Qwen-2.5 (Qwen et al., 2025).

We fine-tune all the models for 15–20 epochs, using an initial learning rate in the range of 5e-5 to 1e-4, with an exponential learning rate scheduler (gamma set to 0.9). We use a batch size between 8 and 12, gradient accumulation steps of 2, and minimize the categorical cross-entropy loss. In Table 9, we report the Strict macro average F1 scores of various models. Note that the reported training set F1 scores correspond to the epoch with the highest test set F1 score. In the held-out test set, our submission based on `Phi-4-mini-instruct` scored an F1 score of 0.69 securing us the 4-th place among 29 submissions in the official leaderboard.

Model	Train F1	Test F1
FLAN-T5-base	0.76	0.59
MathBERT	0.98	0.58
ModernBERT-large	1.0	0.67
Llama-3.2-3B	0.97	0.61
Llama-3-8B	0.74	0.55
Phi-4-mini-instruct	1.0	**0.71**
Qwen2.5-7B-Instruct	1.0	0.65

Table 9: Strict macro average F1 scores of different models on training and test datasets of Track 4

Model	Train F1	Test F1
FLAN-T5-base	0.76	0.59
ModernBERT-large	0.99	0.84
Phi-4-mini-instruct	1.0	0.78
Llama-3.2-3B	1.0	**0.85**
Longformer	-	0.83*
BigBird Roberta Large	-	0.79*
MathBERT	-	0.79*

Table 10: Macro average F1 scores of different models on training and test datasets of Track 5
∗: test set drawn from the same distribution but might differ from the other models

3.5 Track 5 - Tutor Identification

The goal of track 5 was to predict the identity of the tutor for a given response, from a set of 9 identities, such as Sonnet, Llama3.1 8B, Llama 3.1 405B, GPT4 to name a few. We mainly fine-tuned various transformer models for this task with a similar setup to the previous tasks, and have reported our scores in Table 10. We observed that for many models the per-class F1 score for Novice, Llama 3.1 405B and 8B was lower than other classes. For the Novice class, a possible cause could be the lack of enough Novice examples in the dataset. We did not investigate the cause for the low performance for Llama 3.1 8B and 405B in detail, but when we looked at the test set confusion matrix for one of the models, we found that there was significance confusion between Llama 3.1 8B and 405B. It would be interesting to investigate how much of these similarities are task-specific and how much are specific to the base model. A recent preprint (Smith et al., 2025) suggests similar patterns in cosine similarities between the outputs of various LLMs. Note that these metrics are reported on our hold out sets and not the leaderboard test sets. Our best leaderboard test set performance was 0.82, and our final leaderboard position was 16th according to the macro average F1 metric.

4 Conclusion

In this work, we presented our experiments using a fine-tuning-based approach with several encoder-based and large language models to evaluate the pedagogical capabilities of AI tutors. We observed that different LLMs yield varying performance levels, highlighting model-specific behavior. Some class labels in the training data had very few examples, which may have impacted performance. Future work could explore data augmentation and sampling techniques to address this imbalance and potentially improve results. It would also be worthwhile to investigate prompt-based classification methods for evaluating tutor responses in zero-shot or few-shot settings, and explore the use of the models reported in this paper as reward models for post-training or performing test-time scaling on LLMs for improving their pedagogical skills. Additionally, future research could examine the potential of using the same set of AI tutors to reflect on and revise their responses to better align with the goals of effective and helpful AI tutoring systems.

5 Limitations

Automated scoring metrics for evaluating the pedagogy of AI math tutors and AI tutors in general come with their own limitations. Bias introduced by the finetuned model and the underlying pretrained model can lead certain behaviors to be reinforced and certain demographics to be highlighted over other demographics. Cultural considerations also play an important part in pedagogy. A lack of rigorous theoretical guarantees on the mathematical and conceptual accuracy of LLM models can propagate incorrect concepts among students and lead to unwanted friction with instructors. Accessibility of AI tutoring tools could be a barrier for some students with limited resources and internet access, given the resource-expensive nature of LLMs. Moreover, AI tutoring tools typically require students to access internet on their phone or computer, enhancing their risk of being exposed to other websites and social media, causing the risks to outweigh the benefits.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen,

Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, and 4 others. 2025. Phi-4-reasoning technical report. *Preprint*, arXiv:2504.21318.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024b. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Anthropic. 2025. Claude 3.7 sonnet. Available at `https://www.anthropic.com/claude/sonnet`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. *Preprint*, arXiv:2304.00723.

Alexis Chevalier, Jiayi Geng, Alexander Wettig, Howard Chen, Sebastian Mizera, Toni Annala, Max Jameson Aragon, Arturo Rodríguez Fanlo, Simon Frieder, Simon Machado, Akshara Prabhakar, Ellie Thieu, Jiachen T. Wang, Zirui Wang, Xindi Wu, Mengzhou Xia, Wenhan Xia, Jiatong Yu, Jun-Jie Zhu, and 3 others. 2024. Language models as science tutors. *Preprint*, arXiv:2402.11111.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Katherine M. Collins, Albert Q. Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B. Tenenbaum, William Hart, Timothy Gowers, Wenda Li, Adrian Weller, and Mateja Jamnik. 2023. Evaluating language models for mathematics through interactions. *Preprint*, arXiv:2306.01694.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *Preprint*, arXiv:2402.00367.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. *Preprint*, arXiv:2009.06978.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 24678–24704. Curran Associates, Inc.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Irina Jurenka, Markus Kunesch, Kevin R. McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, Ankit Anand, Miruna Pîslar, Stephanie Chan, Lisa Wang, Jennifer She, Parsa Mahmoudieh, Aliya Rysbek, Wei-Jen Ko, Andrea Huber, and 55 others. 2024. Towards responsible development of generative ai for education: An evaluation-driven approach. *Preprint*, arXiv:2407.12687.

Ekaterina Kochmar, Kaushal Kumar Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the bea 2025 shared task on pedagogical ability assessment of ai-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. 2024. Evaluating human-language model interaction. *Preprint*, arXiv:2212.09746.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. *Preprint*, arXiv:2502.18940.

Kaushal Kumar Maurya, KV Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025a. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. *Preprint*, arXiv:2412.09416.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025b. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

OpenAI and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. Mathbert: A pre-trained model for mathematical formula understanding. *Preprint*, arXiv:2105.00377.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Machel Reid, Nikolay Savinov, Denis Teplyashin, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hua Shen and Tongshuang Wu. 2023. Parachute: Evaluating interactive human-lm co-writing systems. *Preprint*, arXiv:2303.06333.

Brandon Smith, Mohamed Reda Bouadjenek, Tahsin Alamgir Kheya, Phillip Dawson, and Sunil Aryal. 2025. A comprehensive analysis of large language model outputs: Similarity, diversity, and bias. *Preprint*, arXiv:2505.09056.

Xiaoyu Tan, Shaojie Shi, Xihe Qiu, Chao Qu, Zhenting Qi, Yinghui Xu, and Yuan Qi. 2023. Self-criticism: Aligning large language models with their understanding of helpfulness, honesty, and harmlessness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 650–662, Singapore. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.