Assessing the efficacy of large language models in generating accurate teacher responses

Yann Hicke, Abhishek Masand, Wentao Guo, Tushaar Gangavarapu
Cornell University

Abstract

(Tack et al., 2023) organized the shared task hosted by the 18th Workshop on Innovative Use of NLP for Building Educational Applications on generation of teacher language in educational dialogues. Following the structure of the shared task, in this study, we attempt to assess the generative abilities of large language models in providing informative and helpful insights to students, thereby simulating the role of a knowledgeable teacher. To this end, we present an extensive evaluation of several benchmarking generative models, including GPT-4 (few-shot, in-context learning), fine-tuned GPT-2, and fine-tuned DialoGPT. Additionally, to optimize for pedagogical quality, we fine-tuned the Flan-T5 model using reinforcement learning. Our experimental findings on the Teacher-Student Chatroom Corpus subset indicate the efficacy of GPT-4 over other fine-tuned models, measured using BERTScore and DialogRPT.

We hypothesize that several dataset characteristics, including sampling, representativeness, and dialog completeness, pose significant challenges to fine-tuning, thus contributing to the poor generalizability of the fine-tuned models. Finally, we note the need for these generative models to be evaluated with a metric that relies not only on dialog coherence and matched language modeling distribution but also on the model's ability to showcase pedagogical skills.

1 Introduction

The advent of powerful open-source generative language models such as GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2020), OPT (Zhang et al., 2022), BLOOM (Scao et al., 2022), Flan-T5 (Chung et al., 2022) or LLAMA (Touvron et al., 2023) has led to significant developments in conversational agents, opening avenues for various applications in education (Wollny et al., 2021). Such AI-driven educational dialogues offer the potential for skill improvement and personalized learning experiences, with intelligent tutoring systems increasingly gaining traction (Bibauw et al., 2022). However, deploying AI-based teachers in the educational ecosystem demands the careful modeling and evaluation of these agents to ensure their capability to address critical pedagogical concerns.

(Tack and Piech, 2022) created the AI teacher test challenge which follows the recommendations from (Bommasani et al., 2021) (pp. 67-72) stating that, if we want to put generative models into practice as AI teachers, it is imperative to determine whether they can (a) speak to students like a teacher, (b) understand students, and (c) help students improve their understanding.

Taking inspiration from the AI teacher test challenge which asks whether state-of-the-art generative models are good AI teachers, capable of replying to a student in an educational dialogue this paper seeks to investigate the applicability of reinforcement learning (RL) techniques in the generation of AI teacher responses within educational dialogues. The AI teacher test challenge emphasizes the need for a systematic evaluation of generative models to ensure that they can effectively communicate with students, comprehend their needs, and facilitate their academic improvement. Can we guide the language generator with RL to help it focus on these pedagogical requirements?

(Tack et al., 2023) organized the shared task hosted by the 18th Workshop on Innovative Use of NLP for Building Educational Applications on generation of teacher language in educational dialogues. Following the structure of the shared task, in this study, we aim to evaluate the potential of combining state-of-the-art generative language models with reinforcement learning algorithms to generate AI teacher responses in the context of real-world educational dialogues sourced from the Teacher Student Chatroom Corpus (Caines et al., 2020, 2022). The natural baselines for the task at hand are SOTA closed-source models such as GPT-4, and

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 745–755

fine-tuned open-source pre-trained models such as GPT-2 (Radford et al., 2019). We will evaluate these natural baselines before evaluating fine-tuned pre-trained models using RL techniques, that optimize for pedagogical quality.

By exploring the role of reinforcement learning in guiding the generation of AI teacher responses, we aim to advance the discourse on the utilization of conversational agents in educational settings and contribute innovative ideas to the ongoing shared task on the generation of teacher language in educational dialogues at the 18th Workshop on Innovative Use of NLP for Building Educational Applications.

The rest of this paper is structured as follows. Section 2 offers a comprehensive review of relevant literature in the areas of AI-driven educational dialogues and reinforcement learning-based language generation. Section 3 discusses the analysis and processing of the dataset prior to conducting any language modeling tasks. In Section 4, the proposed model and its methodology for generating AI teacher responses in educational interactions are introduced. Section 5 evaluates the effects of our approach on the quality and relevance of the generated AI teacher responses and highlights key observations. Finally, Section 6 concludes the paper and explores potential directions for future research.

2 Related Work

A variety of related literature exists in the realm of conversational teaching between a student and a teacher. In this section, we review several notable works addressing aspects of teacher-student dialogues, foundation models, and conversational datasets, which have contributed to the progress and understanding of generative models in educational contexts.

Teacher-Student Dialogues

One prominent resource in educational dialogues is the National Council of Teachers of English (NCTE) dataset (Kane, 2015). It includes numerous examples of teacher-student interactions, which can serve as a valuable resource for the training and evaluation of generative models in an educational context.

The SimTeacher dataset (Cohen et al., 2020) is an assemblage of information obtained through a "mixed-reality" simulation platform. This unique environment aids beginner educators in honing essential skills for classroom settings by employing student avatars managed by human actors. All aspiring teachers from a prominent public university participate in several brief simulation sessions throughout their educational preparation program, focusing on improving their ability to encourage more profound textual understanding among students. The original researchers annotated a variable called "quality of feedback" within the transcript, determining how effectively teachers proactively assist students.

In (Chen et al., 2019), we can find a dataset collected from an education technology company that provides on-demand text-based tutoring for math and science. With a mobile application, a student can take a picture of a problem or write it down and is then connected to a professional tutor who guides the student to solve the problem. The dataset represents, after some selection, 108 tutors and 1821 students. Each session is associated with two outcome measures: (1) student satisfaction scores (1-5 scale) and (2) a rating by the tutor manager based on an evaluation rubric (0-1 scale).

Foundation Models

(Bommasani et al., 2021) provided a comprehensive analysis of the opportunities and risks of foundation models, including insights into their use in educational applications. They identified potential benefits, such as personalized learning and accessibility, while also highlighting the major risks, such as unfair biases and the generation of harmful content. This work establishes the need for carefully crafted benchmarks and evaluations to assess the potential of generative models in education.

The AI Teacher Test (Tack and Piech, 2022) builds on this idea by examining the performance of generative models such as GPT3 (Brown et al., 2020) and Blender (Roller et al., 2020) in generating appropriate and informative responses in a teacher-student dialogue.

Kasneci et al. (Kasneci et al., 2023) conducted an investigation to understand the effectiveness of ChatGPT (Team, 2022) as a tool for educational support. They analyzed the model's performance in a student-tutoring context, examining its ability to provide accurate, relevant, and engaging responses for learners. By identifying the strengths and weaknesses of ChatGPT in this specific setting, they contributed to a better understanding of how

generative models can be successfully deployed in educational applications. Our work builds on these foundations by evaluating the potential of combining reinforcement learning with generative models to enhance the performance of AI teacher agents in educational dialogues.

Conversational Uptake

(Collins, 1982) introduced the concept of uptake as a way to comprehend the effectiveness of conversational responses in a teacher-student dialogue. It laid the groundwork for the evaluation of generative models in dialogues by taking into account the relevance and appropriateness of model-generated responses.

Demszky et al. (Demszky et al., 2021) further explored the concept of Conversational Uptake by proposing metrics to assess the success of responses in maintaining and advancing a conversation. By applying these metrics to AI-generated responses, their work contributes to the evaluation of models in realistic conversation settings, including teacher-student dialogues. Our work attempts to guide the language generation process with similar goals in mind. We hope to find proxies of pedagogical quality through NLP metrics such as BERTScore combined with DialogRPT.

We continue by reviewing the literature utilizing reinforcement learning as a guide for language generation.

Reinforcement Learning for language generation

Policy gradient-based algorithms and their variants have been widely used in text generation to optimize sequence-level metrics (Ranzato et al., 2015; Shen et al., 2015; Norouzi et al., 2016; Pasunuru and Bansal, 2018). Off-policy Reinforcement Learning (RL) is also commonly used in dialogue applications where online interaction with users is expensive (Serban et al., 2017; Jaques et al., 2019). The main difference in our work is that we take advantage of demonstrations and design generic reward functions for generation tasks. We extend this concept to educational contexts by employing reinforcement learning to guide the generation of AI teacher responses in educational dialogues. We focus on optimizing the responses of fine-tuned generative models based on a reward system designed to enhance the pedagogical quality of the generated responses. Recently, Ramamurthy et al. (Ramamurthy et al., 2022) explored the efficacy of using RL to optimize language models in several natural language processing tasks, including text classification, sentiment analysis, and language generation. They developed a library, RL4LMs, which provides a generic framework for deploying RL-based language models for various tasks. We build on top of the RL4LMs framework by adding a new task to its existing array of tasks which we hope can be added as a standard for any future RLHF benchmark.

3 Data

The shared task for BEA 2023 is based on the Teacher-Student Chatroom Corpus (TSCC) (Caines et al., 2020). This corpus comprises data collected from 102 chatrooms where English as a Second Language (ESL) teachers interact with students to work on language exercises and assess the students' English language proficiency.

3.1 Data Extraction and Format

From each dialogue in the TSCC, several shorter passages were extracted. Each passage is at most 100 tokens long, consisting of several sequential teacher-student turns (i.e., the preceding dialogue context) and ending with a teacher utterance (i.e., the reference response). These short passages are the data samples used in this shared task.

The data samples are formatted using a JSON structure inspired by the ConvoKit (Chang et al., 2020). Each training sample is represented as a JSON object with three fields:

- **id**: a unique identifier for the sample.
- **utterances**: a list of utterances corresponding to the preceding dialogue context. Each utterance is a JSON object with a "text" field containing the utterance and a "speaker" field containing a unique label for the speaker.
- **response**: a reference response, which corresponds to the final teacher's utterance. This utterance is a JSON object with a "text" field containing the utterance and a "speaker" field containing a unique label for the speaker.

Each test sample is represented as a JSON object that uses the same format as the training sample but excludes the reference response. As a result, each test sample has two fields:

- **id**: a unique identifier for the sample.
- **utterances**: a list of utterances, which corresponds to the preceding dialogue context. Each utterance is a JSON object with a "text" field containing the utterance and a "speaker" field containing a unique label for the speaker.

3.2 Data Distribution and Characteristics

The TSCC corpus is divided into three sets: train, dev, and test, each comprising 2747, 305 and 273 of the samples, respectively. The corpus has 3325 samples, and each sample has an average length of 7.52 turns, with about 7.33 tokens per turn on average. Table 1 presents a summary of the statistics of the TSCC corpus across the training, development, and testing sets.

The TSCC corpus exhibits several characteristics that are specific to educational dialogues and pose challenges to natural language generation models. For instance, the dialogues often include technical vocabulary and idiomatic expressions related to English language learning. Additionally, the dialogues can be highly varied in terms of topic, complexity, and participant proficiency. Finally, the dialogues can include challenging responses which are based on out-of-context information, posing challenges for conversational agents. These characteristics must be taken into consideration when selecting and evaluating generative models for the TSCC corpus.

3.3 Data Overlap and Challenges

It is worth noting that the released development and training sets in the TSCC dataset have some overlaps, as individual conversation samples within these sets have been generated by creating chunks from larger conversations. This overlap may lead to potential biases and overfitting when training and evaluating models on this dataset. However, the test set for the BEA 2023 shared task is free of overlaps, allowing for a more accurate assessment of the model's performance in generating AI teacher responses.

The presence of overlaps in the development and training sets posed a challenge, as models inadvertently learned to predict teacher responses based on the similarities between the samples rather than genuinely understanding the context and dynamics of the teacher-student interaction. It is essential to be aware of this issue and devise strategies to mitigate the risks associated with such overlaps and ensure that the models are robust and capable of handling diverse and unseen scenarios.

To ensure the validity of our model on the validation set, we employed an iterative inclusion process to create a train-val split without any overlap between them. This process involved carefully selecting and excluding samples from the training set that had any similarity or overlap with the samples in the development set. This approach aimed to minimize the risk of data leakage and ensure that our model was evaluated on a truly unseen set of dialogues.

4 Methods

The primary objective of our study is to investigate the potential of using in-context learning, supervised fine-tuning, and reinforcement learning to generate AI teacher responses in educational dialogues. Our proposed methods will be evaluated using the Teacher Student Chatroom Corpus (TSCC) dataset. In this section, we provide an overview of the three main parts of our methodology: in-context learning using GPT-4, supervised fine-tuning with existing models such as GPT-2 and DialoGPT, and supervised fine-tuning with Reinforcement Learning using the RL4LMs library (Ramamurthy et al., 2022).

4.1 In-context Learning

4.1.1 GPT-4

As a preliminary step, we investigate the potential of in-context learning using GPT-4, a state-of-the-art language model. It generates educational dialogues based on its pre-trained knowledge, which has been acquired from a vast corpus of text data during its training process (the pre-training data might have included the test set; we will address this issue in the discussion section).

To evaluate the performance of GPT-4, we prompted GPT-4 in a few-shot fashion. We retrieved 5 most similar teacher-student conversations from the TSCC dataset and provided them to the model in addition to the current conversation and instructions about the model's role as a teacher. Details about the prompt construction that helps guide the model toward generating suitable responses as a teacher can be found in the Appendix A.

Table 1: Summary of the statistics for the TSCC corpus across the train, dev, and test sets.

Dataset	Train	Dev	Test
Num Samples	2747	305	273
Avg Turns	7.7	7.92	5.23
Avg Tokens Per Turn	7.29	7.21	8.27

4.2 Supervised Fine-tuning

To further adapt pre-trained language models to the specific educational context and generate more accurate and context-aware teacher responses, we explore supervised fine-tuning using GPT-2 and DialoGPT models.

4.2.1 GPT-2

GPT-2 (Radford et al., 2019) is a decoder-only large language model pre-trained on WebText, and we used GPT-2 Large, which has 24 transformer decoder blocks with 774 million parameters.

We fine-tune the GPT-2 model (Radford et al., 2019) using the Huggingface Library on the Teacher Student Chatroom Corpus (TSCC) dataset. For each educational dialogue, we concatenated all dialogue turns into a single string with additional information of speaker roles i.e. students or teachers. As a result, the input to the GPT-2 model consists of a sequence of text representing the conversation history, culminating in the teacher's response. We then finetuned GPT-2 Large (Radford et al., 2019) with a casual language modeling task. Details of the exact hyperparameters used during the fine-tuning process can be found in the Appendix.

After the fine-tuning process, we evaluated the fine-tuned GPT-2 model's performance on the test set by comparing its generated teacher responses to reference responses, assessing the model's ability to generate context-aware and educationally relevant responses in line with the teacher's role in the TSCC dataset.

4.2.2 DialoGPT

DialoGPT (Zhang et al., 2019) is a dialogue model based on the GPT-2 architecture, specifically designed for generating conversational responses. DialoGPT is trained with 147M conversation pieces extracted from Reddit (Zhang et al., 2019), and it is trained with casual language modeling objectives with multi-turn dialogue. We adapt our training dataset with the same format as that of DialoGPT during pretraining and then prompt the DialoGPT to generate an educational dialogue of teachers in the validation set. After training, we follow the same methodology for evaluation as GPT-2 which we discussed in the earlier section.

4.3 Supervised Fine-tuning with Reinforcement Learning

4.3.1 Flan-T5 Fine-tuned with RL4LMs

To optimize the generative models for pedagogical quality, we explore the use of reinforcement learning techniques in the fine-tuning process. We employ the RL4LMs library (Ramamurthy et al., 2022), which provides an efficient and scalable framework for reinforcement learning-based language model fine-tuning.

The RL4LMs library incorporates Proximal Policy Optimization (PPO) (Schulman et al., 2017) as the reinforcement learning algorithm, which is known for its stability and sample efficiency. The library also supports the integration of custom reward functions, allowing us to design rewards that encourage the generation of pedagogically sound teacher responses.

To implement the reinforcement learning-based fine-tuning, we first fine-tune the Flan-T5 (Chung et al., 2022) model on the TSCC dataset using supervised learning, as described in the previous section. Next, we utilize the RL4LMs library to fine-tune the model further using the PPO algorithm. We use an equal division of the F1 as calculated by the roberta-large version of BERTScore and DialogRPT-updown as the reward function. More Details about the reinforcement learning fine-tuning process can be found in the Appendix.

The subsequent evaluation of the fine-tuned Flan-T5 model reveals the benefits of incorporating reinforcement learning into the fine-tuning process, contributing to more context-aware, relevant, and pedagogically effective AI teacher responses.

5 Results

In this section, we present the results and discuss the performance of GPT-4, fine-tuned GPT-2, and fine-tuned DialoGPT models on the TSCC dataset. We analyze the strengths and weaknesses of each approach and provide insights into their potential

applications and limitations in an educational context.

5.1 GPT-4

The GPT-4 model, without fine-tuning on the TSCC dataset, demonstrates a relatively strong performance in generating educational dialogues. The generated teacher responses are generally fluent and contextually relevant, indicating that GPT-4 has a good understanding of the educational context based on its pre-trained knowledge. However, some limitations are observed in the model's ability to generate accurate and pedagogically sound responses consistently.

The carefully crafted prompt provided to the model plays a crucial role in guiding GPT-4 toward generating suitable responses as a teacher. Although the model is capable of generating contextually relevant and linguistically correct responses, it may not always produce the most pedagogically sound or helpful responses for the students. This limitation highlights the importance of fine-tuning the model on a specific educational dataset, such as TSCC, to further enhance its performance in generating AI teacher responses.

Additionally, due to the nature of the dataset, where conversations were often cut off, the model sometimes lacked the full context needed to generate meaningful responses that accurately represented the ground truth. Despite this limitation, GPT-4's responses were generally sensible and appropriate given the available context.

5.2 Finetuned GPT-2

We observe that compared with DialoGPT, GPT-2 usually generates longer and more formal responses, even with the same generation hyperparameters.

5.3 Finetuned DialoGPT

We observe that DialoGPT usually generates shorter and more vernacular responses. It fits better in a conversational setting, but sometimes the educational uptakes are not satisfactory since the responses are not guiding students to learn the language.

5.4 Finetuned Flan-T5 w/ RL

We observe that the results of Flan-T5 w/ RL on the validation set are really good suggesting that the model was able to hack the metrics designed as the reward. On the contrary, it is performing poorly on the test set suggesting that it overfits the validation set. We hypothesize two reasons for this to be the case: the way conversations are split into chunks in the training dataset or the difference in distribution between the training set and the test set.

6 Discussion

Conversational agents have the potential to revolutionize the teaching landscape by addressing several challenges and enhancing the overall learning experience for both students and educators (Wollny et al., 2021). However, developing conversational agents that can behave like human teachers requires addressing several challenges (Tack and Piech, 2022).

Data challenges. As noted in the subsections above, the generations from the GPT-4 model outperformed all the fine-tuned models, with and without reinforcement learning. To this end, we put forward the proposition that an array of dataset features plays a crucial role in posing significant challenges to the fine-tuning process of generative models. These features include several dataset characteristics, including sampling, representativeness, prompt and response lengths, and dialogue completeness—upon manual inspection, we identified several dialogues to be cut off—pose serious challenges in achieving superior performance with fine-tuning. Furthermore, upon random inspection of the generations from the fine-tuned models, we identified that these models seem to have learned simple, generic, often inappropriate yet correct responses such as "thank you" and "okay." While more recent language models have been shown to have high few-shot performance, we believe that fine-tuned models could adapt better to provide domain-specific responses in comparison. To achieve this, we emphasize the need for extending the current dataset to include longer prompts with more context.

It is important to acknowledge that these models might not be as effective as desired in their response generation due to these intricacies. The current efforts made by the research community to collect and build quality datasets encompassing enough information about the educational task to enable AI teacher generative models to fully generalize in any context is what we assess to be the main focus that the community should adopt

Table 2: Validation set results

Model	BERTScore	DialogRPT
GPT-4	0.82	0.69
Finetuned GPT-2 Large	**0.94**	0.63
Finetuned DialoGPT Large	**0.94**	0.64
Finetuned Flan-T5 w/ RL	0.89	**0.71**

Table 3: Test set results

Model	BERTScore	DialogRPT
GPT-4	**0.8**	**0.70**
Finetuned Flan-T5 w/ RL	0.66	0.34

[student]	someone plugged the charger in
[teacher]	that's bad, charger must be ___?
[student]	umm . . .
[model]	(a) plugged in ← score: 0.91 (b) disconnected ← score: 0.90
[reference]	plugged out

Figure 1: An example dialog demonstrating that two opposing responses, (a) and (b), ranked alike using the BERTScore metric.

(Jarratt, 2023).

Evaluation metrics. In addition, we emphasize that to truly gauge the efficiency of these AI-powered teaching models, it is vital to go a step further and examine their ability to comprehend the unique nuances in the students' queries and cater to their particular educational requirements. This implies the need for a pedagogically meaningful evaluation metric. We believe that it is crucial for the research community to embrace this as the second primary focus. While common evaluation metrics such as BERTScore and DialogRPT are commonly used in several language and dialog modeling tasks, it is important to note that these metrics were not fundamentally designed to capture the level of pedagogical meaningfulness in the generated responses. As an example, consider the dialog shown in Figure 1—depending on the given context, only one of the responses (option (b): disconnected) is correct, while both the responses are ranked as equally correct by the BERTScore metric. Commonly-used domain-agnostic metrics often serve as a proxy for how coherent and human-like the generated responses are. However, for more goal-oriented tasks such as modeling teacher-student conversational dialogues, these metrics seem to fall short. This generalization gap becomes more apparent on analyzing the results from the fine-tuned Flan-T5 model with a feedback loop based on BERTScore and DialogRPT scores—despite the model performing significantly well on training and validation sets, it failed to generalize on unseen test data. In an effort to advance research on this front, we note the need for auxiliary training-level metrics, including the faithfulness of the generation to the true response, to ensure that the generations are context-aware and factually accurate (e.g., correct option (b) vs. incorrect option (a) in Figure 1).

GPT-4 unknown pre-training data. We understand that the use of GPT-4 as a baseline in our study presents challenges due to its unknown training data. Yet, whether GPT-4 has seen parts of the TSCC dataset during its pre-training or not, the improvement of performance compared to the reference with regard to the DialogRPT scores and human evaluation scores attached to the leaderboard of the shared task suggests that the potential of using such high-performing models in this domain warrants further exploration.

7 Conclusion

In this paper, we explored the potential of using large pre-trained language models and reinforcement learning for generating AI teacher responses in an educational context. We first presented a few-shot approach using the GPT-4 model, which demonstrated promising results in generating contextually relevant and fluent responses, but with limitations in generating pedagogically sound responses consistently. We then fine-tuned GPT-2 and DialoGPT on the TSCC dataset and evaluated their performance using BERTScore and DialogRPT metrics. We also proposed an approach using RL to optimize directly for pedagogical values. We hypothesized that several dataset characteristics (e.g., dialog completeness, sampling) pose challenges to achieving superior performance with fine-tuning. To this end, we recommend the extension of the dataset to include longer prompts with extended context. Finally, we also draw attention to the need for more domain-specific metrics (in both evaluation and reward-based training) in enabling the generation of accurate, context-aware, and factually correct teacher responses.

References

Serge Bibauw, Thomas François, and Piet Desmet. 2022. Dialogue systems for language learning: Chatbots and beyond. In *The Routledge handbook of second language acquisition and technology*, pages 121–135. Routledge.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Andrew Caines, Helen Yannakoudakis, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2022. The teacher-student chatroom corpus version 2: more lessons, new annotation, automatic detection of sequence shifts. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 23–35.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The teacher-student chatroom corpus. *arXiv preprint arXiv:2011.07109*.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246*.

Guanliang Chen, Rafael Ferreira, David Lang, and Dragan Gasevic. 2019. Predictors of student satisfaction: A large-scale study of human-human online tutorial dialogues. *International Educational Data Mining Society*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Julie Cohen, Vivian Wong, Anandita Krishnamachari, and Rebekah Berlin. 2020. Teacher coaching in a simulated environment. *Educational evaluation and policy analysis*, 42(2):208–231.

James Collins. 1982. Discourse style, classroom interaction and differential treatment. *Journal of reading behavior*, 14(4):429–437.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. *arXiv preprint arXiv:2106.03873*.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.

Daniel Jarratt. 2023. Chatgpt: The double-edged sword of ai in education.

Thomas Kane. 2015. *National Center for Teacher Effectiveness Main Study*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Mohammad Norouzi, Samy Bengio, Navdeep Jaitly, Mike Schuster, Yonghui Wu, Dale Schuurmans, et al. 2016. Reward augmented maximum likelihood for neural structured prediction. *Advances In Neural Information Processing Systems*, 29.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. *arXiv preprint arXiv:1804.06451*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. 2017. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2015. Minimum risk training for neural machine translation. *arXiv preprint arXiv:1512.02433*.

Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, page to appear, Toronto, Canada. Association for Computational Linguistics.

Anaïs Tack and Chris Piech. 2022. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. *arXiv preprint arXiv:2205.07540*.

OpenAI Team. 2022. Chatgpt: Optimizing language models for dialogue.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. 2021. Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, 4:654924.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

A Appendix

A.1 GPT-4 Prompt Construction

To evaluate the performance of GPT-4, we provided it with a few-shot prompt that includes a selection of similar teacher-student conversations from the TSCC dataset. This approach helps guide the model toward generating suitable responses as a teacher. The prompt is constructed as follows:

- We direct the system role to act as a teacher and encourage learning by using the prompt as given below.
- Retrieve the 5 most similar teacher-student conversations from the TSCC dataset. This is done by computing the cosine similarity between the input conversation context and the current conversation context in the dataset using embeddings generated by the text-embedding-ada-002 model.
- Concatenate the selected conversations with the input conversation, separated by special tokens to indicate the beginning and end of a new sample conversation.

This prompt construction aims to provide GPT-4 with the necessary context and guidance to generate accurate and pedagogically relevant responses in the context of teacher-student dialogues. The prompt is designed as follows:

> You are acting as a teacher, and you are helping a student learn. Be patient, helpful, and kind. Don't be superimposing; give short responses to encourage learning. Make the student feel comfortable and confident, and help them learn. Now, join the following conversation: <conversation context>

The prompt is designed using the following directives in mind:-

- We instruct the system with several indicators to act as a teacher and provide helpful advice to the student.
- To mitigate the challenge of generating teacher-like responses, we advise the model to be patient, kind, and helpful to the student.
- Through the directive to keep responses short and encouraging, we guide the model toward generating suitable responses that might help the student learn effectively.
- The model is also instructed to make the student feel comfortable and confident in their learning process, providing an overall supportive environment for the student.
- Finally, the conversation context is provided to the model to set the context for the given student query, allowing the model to generate appropriate responses given the conversation context.

Combining all these aspects together, we aim to guide the model toward generating contextually relevant and pedagogically meaningful responses in the given teacher-student dialogue.

We use the following hyperparameters for querying the GPT-4 model:

- Model: gpt-4-0314
- Temperature: 1
- Max Tokens: 100
- Top p: 1

A.2 Fine-tuning Exact Parameters

For our supervised fine-tuning experiments, we used the following hyperparameters:

A.2.1 GPT-2

- Learning rate: 1e-5
- Batch size: 32
- Epochs: 10
- Max sequence length: 1024
- Optimizer: AdamW
- Scheduler: linear learning rate scheduler

A.2.2 DialoGPT

- Learning rate: 1e-5
- Batch size: 32
- Epochs: 10
- Max sequence length: 1024
- Optimizer: AdamW
- Scheduler: linear learning rate scheduler

A.3 Supervised Fine-tuning with Reinforcement Learning Details

To implement the reinforcement learning-based fine-tuning using the RL4LMs library, we first fine-tuned the Flan-T5 model on the TSCC dataset using supervised learning. After this initial fine-tuning step, we utilized the RL4LMs library to fine-tune the model further using reinforcement learning. We used an equal division of the BERTScore and DialogRPT as the reward function to optimize the model for pedagogical quality. The following hyperparameters were used for the reinforcement learning fine-tuning process:

- Learning rate: 1e-6
- Batch size: 64
- Epochs: 5
- Max prompt length: 512
- Max episode length: 100
- Optimizer: AdamW
- Scheduler: linear learning rate scheduler

The YAML file for the RL4LMs scipt is as follows:

```
tokenizer:
del_name: google/flan-t5-small
dding_side: left
uncation_side: left
d_token_as_eos_token: False
rd_fn:
: dialog_rpt_bert
gs:
BERTScore_coeff: 0.5
DialogRPT_coeff: 0.5
pool:
: bea
uncate: False
gs: {}

envs: 1
gs:
max_prompt_length: 100
max_episode_length: 20
terminate_on_eos: True
context_start_token: 0
prompt_truncation_side: "right"
: ppo_separate
gs:
n_steps: 20
batch_size: 64
verbose: 1
learning_rate: 0.000001
clip_range: 0.2
n_epochs: 1
value_update_epochs: 3
# batchify: False
gae_lambda: 0.95
gamma: 0.99
ent_coef: 0.01
_div:
coeff: 0.001
target_kl: 2.0
licy:
id: seq2seq_lm_actor_critic_policy
args:
  model_name: google/flan-t5-small
  apply_model_parallel: True
  prompt_truncation_side: "right"
  generation_kwargs:
    do_sample: True
    top_k: 0
    min_length: 9
    max_new_tokens: 20
n_evaluation:
al_batch_size: 64
iters: 200
al_every: 20
ve_every: 10
trics:
- id: bert_score
  args:
    language: en
- id: dialog_rpt
  args:
    model_name: "microsoft/DialogRPT
                -updown"
    label_ix: 0
    batch_size: 1
# - id: uptake
#   args:
#     model_name: None
#     label_ix: 0
#     batch_size: 1
neration_kwargs:
num_beams: 5
min_length: 9
max_new_tokens: 20
```