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

Jérémy Chaverot | 315858 | jeremy.chaverot@epfl.ch
ZEPHYR

1 Summary

The paper "Assessing the Efficacy of Large Language Models in Generating Accurate Teacher Responses" (Hicke et al., 2023) delves into the evaluation of large language models (LLMs) concerning their ability to generate accurate, informative and pedagogical responses. As the integration of artificial intelligence (AI) technologies into educational settings becomes more ubiquitous, understanding the capabilities and limitations of LLMs in supporting teaching and learning processes is of paramount importance.

The study assesses the abilities of several models, including GPT-4 with in-context learning. However, it not only examines the evaluation of LLM for such a purpose but also engages in careful modeling. This involves fine-tuning GPT-2, fine-tuning DialoGPT, and finally fine-tuning the Flan-T5 model, utilizing reinforcement learning in addition.

The dataset utilized throughout for training is the Teacher-Student Chatroom Corpus (TSCC), as well as the BERTScore and DialogRPT for comparisons between the models to gauge response accuracy and coherence as outlined in the paper's aim.

After extensive evaluation, it finally appears that OpenAI's GPT-4, with its presumed 1 trillion parameters, outperforms all fine-tuned models despite their extensive training.

As an attempt to explain this rather disappointing result, the authors suggest the inherent limitations imposed by the TSCC dataset used during fine-tuning. This dataset poses a true challenge due to its small number of tokens per sample and its lack of representativeness and dialog completeness. Additionally, the two metrics, BERTScore and DialogRPT, may not fully capture the inherent challenges of student-teacher interactions. Specifically, the Flan-T5 model, fine-tuned and enhanced with RL, whose reward function depends on both metrics, does not lead to a significant improvement in results on the test set.

The paper stands out for its comprehensive explanation, detailing every step from dataset preprocessing to model evaluation, aimed at creating a functional and relevant AI teaching assistant. Moreover, it explores the unique and singular approach of applying RL techniques to generate AI teacher responses.

The findings reveal nuanced insights into the performance of LLMs in generating teacher responses. While the models demonstrate impressive capabilities in certain domains, such as factual recall and language fluency, they also exhibit limitations, particularly in interpreting nuanced prompts and providing contextually accurate answers. Variations in performance are observed across different LLM architectures, underscoring the importance of considering model characteristics in educational applications.

In the context of the literature review, my stance on the paper is somewhat mixed. While the authors provide relevant insights, there are moments of disappointment due to the lack of novelty, the shallow nature of the solutions proposed in addressing the challenges, and the absence of published source code behind the results.

2 Strengths

In this section, we will focus on the strengths of the paper. In the related works, a wide range of existing resources is presented, including datasets, foundation models, conversational uptake, and RL for language generation. This comprehensive overview is a significant asset for summarizing and gaining a better understanding of the foundation of their study.

The comprehensive preprocessing outlined in the Data section, applied during the formation of samples from the TSCC dataset, is appreciated. Not only does it enhance reproducibility, but it

also aids readers in their comprehension. Given the utilization of both fine-tuning with and without RL, clarity on the data format is crucial to avoid confusion.

Regarding the challenges with the data, the authors highlight the overlap between different sets of the TSCC dataset and describe their consequences on potential biases and overfitting. Subsequently, they undertake a remarkable effort to mitigate these risks and construct robust models capable of generalization.

A positive aspect regarding reproducibility is that the prompting techniques used for in-context learning with GPT-4 are described in great detail, point by point. Not only do they provide the prompt sentences, but they also explain their purpose and why they are needed conceptually.

3 Weaknesses

Let's now explore the downsides and shortcomings in the literature. Throughout the article, there is a notable absence of innovation. The selected LLMs (GPT-2, GPT-4, DialoGPT, and Flan-T5) are well-established, the TSCC dataset is used without substantial modification, RL technique is implemented using the RL4LMs library, and the evaluation of model effectiveness (despite being the paper's title focus) is conducted using BERTScore and DialogRPT, both of which are not particularly innovative metrics.

In July 2023, the paper on Direct Preference Optimization (DPO) (Rafailov et al., 2023) had already been published. This revolutionary and simpler technique addresses the primary drawback of RLHF, specifically its unstable reward model. Given its independence from metrics, it would have been pertinent to consider employing DPO. Unfortunately, the authors solely focus on RL in their approach.

Regarding the metrics, it is mentioned that "common evaluation metrics such as BERTScore and DialogRPT are commonly used in several language and dialog modeling tasks [...]", yet no further explanation is provided regarding their selection (for instance the contextual embedding using the BERT transformer provided by BERTScore). Moreover, it is acknowledged multiple times that these two metrics alone fail to fully capture the pedagogical challenges of an AI assistant. However, the paper lacks a genuine attempt to construct a new metric, except for suggesting that it should encompass the faithfulness of the generation to the true response to ensure contextual awareness.

In the methods section, it is stated, "We use an equal division of the F1 as calculated by the roberta-large version of BERTScore and DialogRPT-updown as the reward function." However, no justification for choosing the average of the two metrics is provided, despite its critical role in the reward model. It would have been appreciated to conduct more experiments with different weightings of the two metrics and assess their impact on accuracy on the test set.

In the results section, it is observed that the qualitative assessment of generations from the fine-tuned GPT-2 and DialoGPT models is superficial, shallow and very brief (4 to 6 lines). To gain a better understanding, it would have been necessary to make the generations accessible, providing readers with a more learning experience.

While the paper provides insightful comparisons among the studied LLMs, it completely overlooks comparing baseline models to their fine-tuned versions to grasp the utility of additional training.

The results for the validation set are provided for all four models using the two metrics employed. However, in the table detailing the test set results, scores for fine-tuned GPT-2 and DialoGPT are conspicuously absent, with no explanation provided. This lack of transparency is criticized.

Additionnaly, even if more detailed information on hyperparameter choices can be found in the appendix, it is noted that the source code has not been published jointly to the article, which greatly impedes reproducibility.

The paper completely overlooks ethical considerations regarding LLMs, a critical area deserving deeper exploration. An in-depth examination of potential safeguards or policies to mitigate the risk of misuse of generated text would offer readers a better understanding of this crucial concern.

In summary, while (Hicke et al., 2023) provides a thorough and comprehensive examination of LLMs in educational contexts, it lacks novelty and innovation. The absence of detailed comparisons and reproducibility impedes its depth. Ethical considerations regarding LLMs are overlooked. Despite these shortcomings, the study offers valuable insights into LLM performance in generating teacher responses, serving as a stepping stone for further research and refinement in the integration of AI technologies in education.

References

Yann Hicke, Abhishek Masand, Wentao Guo, and Tushaar Gangavarapu. 2023. Assessing the efficacy of large language models in generating accurate teacher responses. pages 745–755.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.