Group 11 - Tay Kaiying, Roydon (A0271742E), Ngu Jia Hao (A0272789H), Chia Bing Xuan (A0259419R)

1. Abstract

Large volumes of healthcare knowledge are readily available on the internet. However, much of this information takes the shape of formal scientific literature, which often consists of medical jargon and highly technical terms. Attempting to understand such forms of academic writing can be daunting. Moreover, one might be unsure of how to access healthcare information that is presented in a more simplified and digestible manner. Therefore, the average person might find it more challenging to seek answers for any healthcare queries which he or she may have.

We are interested in the aforementioned problem of healthcare Question Answering (QA), because it has only been addressed to a certain extent. Large language models (LLMs) of higher complexity can indeed answer queries with a high degree of accuracy, whilst summarising and simplifying content for a non-technical audience. However, due to the large number of parameters, training these LLMs is computationally expensive. Therefore, we aim to explore the potential of leveraging more lightweight LLMs for healthcare QA, which would be more cost-effective.

There are currently two limitations for these smaller models. Firstly, most of them may not be fine-tuned for the healthcare use case, thus lacking the specific domain knowledge required for healthcare QA. Secondly, these models may not be aligned with layperson preferences for natural, jargon-free responses to healthcare queries. We would like to fine-tune these models to perform better in these two areas.

2. Dataset

We are using a subset of the Medical Question Answering Dataset (MedQuAD), which can be accessed via Kaggle. MedQuAD was created from 12 National Institutes of Health (NIH) websites. It consists of 47457 question-answer pairs related to medicine and healthcare (Abacha & Demner-Fushman, 2019). In the subset that we are using, there are 14984 unique question-answer pairs, covering a total of 5127 focus areas.

3. Methodology

We can choose a lightweight LLM from the selection of SmolLMs offered by Hugging Face. Model training can be done in two iterative stages - instruction fine-tuning and reinforcement learning (RL).

3.1. Instruction Fine-Tuning

Using the MedQuAD subset, we will firstly conduct instruction fine-tuning on the LLM for the healthcare QA task. This improves the conversation model's knowledge in the healthcare domain, whilst aligning it with specific QA instructions. We intend to use Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA) adapters for training, so as to reduce computational cost and memory requirements. We could also apply quantization on the LLM, converting the pretrained weights and activations to values of a lower precision - this would further boost memory efficiency (Jacob et al., 2017).

3.2. Reinforcement Learning (RL)

Secondly, we wish to carry out RL to improve the formatting and style of the LLM, enabling it to respond to healthcare questions with minimal jargon and in an easily comprehensible manner. We plan to use

Group Relative Policy Optimisation (GRPO), which does not require a value model or a large set of labelled human preference data (Guo et al., 2025).

The process of GRPO is as follows:

1. We use the instruction-tuned LLM from the first stage to generate answers for a question set. For each question, we generate multiple different answers with temperature sampling.
2. We need to assign a reward value to each of these question-answer pairs. To assign rewards, we can adopt one of two approaches:
    a. Use a larger LLM-as-a-judge as an implicit reward model (Cao et al., 2024). Techniques like few-shot prompting and chain-of-thought prompting can be used to ensure that an answer is rewarded appropriately, given its corresponding question.
    b. Define a simple rule-based reward function that is computationally inexpensive, such as regex matching to find and reward traces of reasoning (Guo et al., 2025).
3. We calculate the advantage of each response by standardising the rewards for the multiple outputs generated from one question. We repeat this for all questions.
4. We use KL-Divergence to ensure model outputs do not drift too far from the instruction-tuned model weights.

4. Expected Outcomes

4.1. Data Availability

For instruction fine-tuning, we are using an existing dataset which is readily accessible on Kaggle - there is no need to spend additional time collecting data. In addition, our proposed RL workflow only requires crafting some few-shot examples and system prompts for an evaluator LLM, at the maximum.

4.2. Evaluation

To evaluate our work, we plan to use a larger LLM-as-a-judge for the model outputs. We aim to use few-shot prompting to help the larger LLM understand our scoring criteria when we use it as a judge. We will also perform qualitative evaluation, comparing sample outputs from the instruction-tuned model with those from the post-RL model.

4.3. Ablation Studies

We intend to perform several ablation studies to investigate the effect of quantization on the instruction-tuned model's performance. We plan to use varying degrees of quantization on the model (e.g. float16, int8, etc.), and evaluate them after the instruction fine-tuning stage. We may also test the effect of LoRA on the instruction-tuned model by altering the LoRA hyperparameters (e.g. rank, alpha, target modules (which layers to apply LoRA to), etc.).

5. References

1. Abacha, A. B., & Demner-Fushman, D. (2019). A Question-Entailment Approach to Question Answering. *BMC Bioinformatics*, *20*(511). https://doi.org/10.1186/s12859-019-3119-4
2. Cao, Y., Zhao, H., Cheng, Y., Shu, T., Chen, Y., Liu, G., Liang, G., Zhao, J., Yan, J., & Li, Y. (2024). Survey on Large Language Model-Enhanced Reinforcement Learning: Concept, Taxonomy, and Methods. *IEEE Transactions on Neural Networks and Learning Systems, 36*(6), 9737–9757. https://arxiv.org/abs/2404.00282
3. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., … Zhang, Z. (2025). Deepseek-R1: Incentivizing reasoning capability in LLMS via reinforcement learning. arXiv.org. https://arxiv.org/abs/2501.12948
4. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2017). Quantization and training of neural networks for efficient integer-arithmetic-only inference. arXiv.org. https://arxiv.org/abs/1712.05877