UltimateMedLLM-Llama3-8B: Fine-tuning Llama 3 for Medical Question-Answering

Stanford CS224N Custom Project

Jayson Meribe
Department of Computer Science
Stanford University
`jmeribe@stanford.edu`

Sean Zhang
Department of Computer Science
Stanford University
`seantenzinzhang@stanford.edu`

Abstract

Question-answering (QA) is a crucial task in NLP, meant to allow users efficient access to precise information. As such, there has been much interest in developing QA for the medical domain for efficient access to accurate medical information. To this aim, our project involves (1) fine-tuning the 8-billion parameter instruction-tuned version of Meta's LLaMA 3 model in pursuit of improving performance in medical question-answering. (2) We implement the prompting techniques of ensemble refinement (ER) and chain-of-thought with self consistency (SC). (3) Additionally, we create and publish a novel adversarial question dataset designed to test robustness regarding safety risk and health equity. Finally, (4) we make improvements to a recently-published LLM self-evaluation method and use it to evaluate our model's alignment with human preferences compared to benchmark models.

1 Key Information to include

- TA mentor: Soumya Chatterjee
- External collaborators (if no, indicate "No"): No
- External mentor (if no, indicate "No"): No
- Sharing project (if no, indicate "No"): No
- Contributions: Jayson implemented the fine-tuning process and prompting techniques. Sean created the adversarial dataset and implemented and modified the PiCO evaluation pipeline.

2 Introduction

Large language models (LLMS) have become increasingly adept at the task of question answering across many domains. However, there are many specific tasks where performance suffers compared to human experts. One of these tasks is medical question answering, which shows major room for improvement on many subdomains such as multiple question and open-ended long-form questions evaluated against physicians Nori et al. (2023). The ability to retrieve medical knowledge, reason over it, and answer medical questions comparably to physicians is a longstanding goal for LLMs. This is because of the essential role of language in underpinning all interactions in health and medicine, especially those between patients and care providers. The creation of reliable, knowledgeable, and interpretable medical question answering large language models could bring major efficiency and quality improvements to health care systems around world.

The problem of creating medical question answering LLMs is three-pronged. For one, these models are expected to generate accurate, well-reasoned, and thorough responses to any type of question a user might ask. Evaluating the quality of open ended responses according to the aforementioned necessities of medical question answering LLMs is often done with real physicians as was done in Singhal et al. (2023). However, such evaluation is expensive and hard to accomplish without significant available resources, which is why in this paper we introduce an alternative method. Secondly, the medical domain is often very concerned with privacy. It is likely that stakeholders will be eager to use models that they can run on their own with minimal cost, as it would allow them to avoid sharing sensitive patient data. As such reducing the total amount of compute required to run these models is of large importance, which is why we fine tune the smallest Llama 3 model and add structures such as ensemble refinement and chain-of-thought with self consistency on top as done in Singhal et al. (2023). Finally, effectively probing medical question answering LLMs on their safety and limitations is a challenging task and there are currently no publicly available datasets for this task. This is why we create our own dataset of adversarial questions that probe the safety and limitations of our fine tuned models.

Overall, we found that by using our fine-tuned model, UltimateMedLLM-Llama3-8B, we could come very close to the results of Med-PaLM 2, which has 50 times more parameters, on certain benchmarks. We also verified the efficacy of a published self-evaluation method as a judge of human preference alignment in place of actual physician feedback.

3 Related Work

One of the most influential works that inspired our approach is the development of Med-PaLM 2 by Singhal et al. (2023). Med-PaLM 2 aimed to achieve expert-level performance in medical question-answering (QA) by generating responses comparable to those provided by physicians. The model was fine-tuned on the MultiMedQA dataset, which combines several medical QA benchmarks, using empirically chosen proportions of training examples to optimize performance. Techniques such as ensemble refinement and chain-of-thought prompting were employed to enhance the model's accuracy and reasoning capabilities. Med-PaLM 2 set state-of-the-art performance metrics on benchmarks like MedQA with 85.4% accuracy, MedMCQA with 72.3% accuracy, and MMLU Clinical Knowledge with 88.7% accuracy, demonstrating the potential of large language models (LLMs) in the medical domain Singhal et al. (2023).

Another significant contribution to the field is the recent introduction of Med-Gemini models, as detailed by Saab et al. (2024). Med-Gemini builds on the core strengths of the Gemini 1.0 and 1.5 models, which include robust multimodal and long-context reasoning capabilities. These models are specialized for medical applications and integrate web search and custom encoders for novel modalities. Med-Gemini models have established new state-of-the-art performance on 10 out of 14 medical benchmarks, surpassing GPT-4 across various metrics. For instance, on the MedQA (USMLE) benchmark, the best-performing Med-Gemini model achieved 91.1% accuracy, outperforming the previous best Med-PaLM 2 by 4.6%. Additionally, Med-Gemini demonstrated superior performance on complex diagnostic challenges from the New England Journal of Medicine (NEJM) and the GeneTuring benchmark. On 7 multimodal benchmarks, including NEJM Image Challenges and MMMU (health & medicine), Med-Gemini improved over GPT-4V by an average relative margin of 44.5%, showcasing its robust performance across diverse tasks Saab et al. (2024).

Building on the achievements of Med-PaLM 2, our work focuses on demonstrating that robust medical QA LLMs can be developed using relatively smaller, open-source models. This aligns with the findings of other studies that highlight the potential of fine-tuning smaller models for specific tasks, thereby reducing computational resources and making advanced QA systems more accessible.

Our work also relates to research on evaluation methods for LLMs. Traditional evaluation of medical QA models often involves real physicians, which is resource-intensive. Inspired by this challenge, we implement the PiCO (Peer Consistency Optimization) evaluation published by Ning et al. (2024) to assess its ability to assess model answer alignment with human preferences. If shown to do well, this method could enable organizations with limited resources to effectively assess LLMs without extensive computational or human resources.

4 Approach

4.1 Parameter-Efficient Fine Tuning (PEFT)

Even though Llama 3 8B is the smallest Llama 3 model, full-finetuning of its parameters remained beyond on our available resource. Hence, instead, to finetune Llama 3 8B for medical question answering we use parameter efficient fine tuning (PEFT). More specifically, we use low rank adaptation (LoRA) to finetune the query and value projection layers in Llama 3 8B Hu et al. (2021). LoRA works via augmenting the pretrained parameters of a model W_0 with low rank matrices, which can be expressed with far fewer parameters B, A, where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, $W_0 \in \mathbb{R}^{d \times k}$, and $r << min(d, k)$ Hu et al. (2021). Regarding specific hyperparameters relating to LoRA we use a rank of 8 and a value of 16 for α.

4.2 Other Hyperparameters

Regarding other hyper parameters, we use AdamW with weight decay 0.01 and a learning rate of 3e-4 along with a cosine scheduler with 100 warm up steps. We decided to train with 4 epochs, 64 gradient accumulation steps, which allow for more accurate gradient steps, and a batch size of 2. Finally, due to compute constraints we were forced to restrict the maximum sequence length during training to 1024 tokens. Even this sequence length became a problem during evaluation as we were forced to restrict the maximum sequence length to 512.

4.3 Prompting Techniques

In addition to finetuning Llama 3 8B for medical question answering we employed two prompting techniques: ensemble refinement (ER) and chain-of-thought with self consistency (SC). For ensemble refinement we implemented the approach used in Singhal et al. (2023). That is, we broke down generations into two steps. In the first step we create many generations while randomly sampling temperature values. In the second step, we condition on the previously generated values and prompt the model to come up with a final answer and allows it to access responses generated during the first step. The final answer is then a simple majority vote from the answers generated during the second step. However, unlike in Singhal et al. (2023), we use 5 generations in the first step and 7 generations in the second step. For SC, we simply sampled 5 chain-of-thought generations from our model and did a majority vote.

4.4 Adversarial Dataset

Our third main approach is the creation of an adversarial dataset, consisting of 240 open-ended questions designed to elicit wrong or dangerous responses from our model regarding safety and health equity. Our approach mirrors that of the Med-PaLM 2 paper Singhal et al. (2023), which created such a dataset with these same criteria but did not make it publicly available. More details regarding its construction can be found in Section 5.1.

4.5 PiCO Evaluation

Then, to evaluate these open-ended questions, we aim to utilize the PiCO evaluation method established by Ning et al., which creates a ranking from a pool of models by quality of responses Ning et al. (2024).

Specifically, we let Q denote the set of all n questions (with index $i = 1, \dots, n$) and M denote the set of all m models (with index $j = 1, \dots, m$). For some $Q_i \in Q$, we randomly construct a battle pair $< A_i^{j_1}, A_i^{j_2} >$ for review. Each pair is reviewed by five random models to determine which response is better. By the end of the ranking process, this will leave us with the set of quadruples:

$$D = \{(A_i^{j_1}, A_i^{j_2}, >, w^s)\}_{i, j_1, j_2, s}$$

where $i \sim Q$ and $j_1, j_2, s \sim M$, such that each quadruple $(A_i^{j_1}, A_i^{j_2}, >, w^s)$ indicates that the "reviewer" M_s has determined that the answer $A_i^{j_1}$ is better than answer $A_i^{j_2}$ with confidence w^s Ning et al. (2024). This confidence weight is calculated by the optimization problem:

$$\arg\max_{w} \text{Consistency}(G, w)$$

where we define G_j, the score for model M_j, as $G_j = \sum_{(A_i^{j_1}, A_i^{j_2}, w^s)\sim\mathcal{D}} \mathbf{1}\{A_i^{j_1} > A_i^{j_2}\} \cdot w^s$.

In code, this is implemented via a simple neural network with a single linear layer and a sigmoid activation function, representing the model confidence scores between 0 and 1. This network is trained using Stochastic Gradient Descent (SGD) to optimize the consistency of each LLM's evaluation capability, measured by Pearson correlation between the confidence scores and response quality. During training, the models' confidence scores are iteratively adjusted to maximize consistency, as indicated by the correlation metric. Once these scores converge, we are left with a final re-ranking of the LLM pool which is closer to human preferences.

The paper's approach measures the metrics of permutation entropy (PEN), count inversions (CIN), and longest increasing subsequences (LIS), which are defined more rigorously in Section 5.2 Ning et al. (2024). We introduce an additional metric for measuring similarity between the ranking derived from PiCO evaluation and a baseline ranking, namely Rank-Biased Overlap (RBO), also defined in Section 5.2. We also modify the code to improve accessibility as outlined in 5.3.1.

5 Experiments

5.1 Data

We first evaluate our fine-tuned model's accuracy on multiple-choice style questions, using the test splits of the same datasets we used to train it. Specifically, this refers to the test splits of MedQA, PubMedQA, MedMCQA, and the clinical knowledge, medical genetics, anatomy, professional medicine, college biology, and college medicine subsets of the MMLU Dataset.

These datasets cover a variety of medical question formats. MedQA consists of questions sourced from the United States Medical Licensing Examination (USMLE) in multiple-choice format, paired with complex, detailed contexts Jin et al. (2020). Each question in PubMedQA is based on an existing research article title or derived from one, and is evaluated by a yes/no/maybe answer which summarizes the conclusion Jin et al. (2019). Questions in MedMCQA are also multiple-choice, derived from various medical entrance exams in India, and are generally shorter and less complicated than those in MedQA while also covering a broader spectrum of subjects Pal et al. (2022). Lastly, the MMLU Dataset which is also multiple-choice includes several subsets such as clinical knowledge, medical genetics, anatomy, professional medicine, college biology, and college medicine, aimed at evaluating model performance within each of these specifc domains Hendrycks et al. (2021).

Then, we evaluate our model's performance on open-ended questions by testing on our newly-created adversarial dataset, designed to illicit harmful responses when faced with suggestive questions regarding safety and health equity. This dataset was created by hand, drawing upon several works from recent literature detailing prominent issues in health equity such as machine learning in healthcare Chen et al. (2021), patient safety Williams et al. (2015), the effects of discrimination on health Williams et al. (2019), and the systemic prevalence of racism in health access Malawa et al. (2021). To mirror the specifications used by Google, our adversarial dataset is 240 questions long, of which 60 questions broadly target health equity, drug use, alcohol, mental health, COVID-19, obesity, suicide, and medical misinformation, and 180 questions focus on health equity topics like healthcare access, quality, and socioeconomic factors Singhal et al. (2023). These questions do not have associated answers, as they are meant to assessed qualitatively by human evaluation.

No.	Question	Category
26	What are the primary factors contributing to the higher rates of suicide among LGBTQ+ youth compared to their heterosexual peers?	General
67	How do disparities in access to preventive health screenings contribute to health inequities?	Health Equity

Table 1: Sample questions from adversarial dataset

5.2 Evaluation method

For multiple-choice performance, we simply measure by accuracy, namely the fraction of correct answers out of total answers. Then, for open-ended evaluation, we consider the final ranking created by the PiCO evaluation script as well as the associated ranking metrics of Permutation Entropy, Count Inversions, and Longest Increasing Subsequence Ning et al. (2024).

Permutation Entropy, a concept from time series analysis, measures the complexity or randomness in the order of elements in a ranking. A lower permutation entropy indicates that the learned ranking is closer to the human ranking. It is calculated by the summation:

$$\mathcal{L}_{PEN}(\hat{\mathcal{R}}, \mathcal{R}^*) := -\sum_{\pi} p(\pi) \log p(\pi),$$

$$p(\pi) = \frac{\#\{t | 0 \leq t \leq m-k, (M_{t+1}, \ldots, M_{t+k}) \in \pi\}}{m-k+1}.$$

Where π denotes different permutations and k is a tuneable hyper-parameter which we fix at 3 to match the approach by Ning et al. We can understand this more intuitively as sampling some subsequences and calculating the entropy for all permutation types.

Then, Count Inversions counts the number of inversions (disorder) in the learned ranking compared to the human ranking. Fewer inversions suggest a closer match to the human ranking. Mathematically, this is represented by the sum of indicator functions that identify inversions in the ranking.

$$\mathcal{L}_{CIN}(\hat{\mathcal{R}}, \mathcal{R}^*) := \sum_{M_i, M_j \sim \mathcal{M}} \mathbf{1}\{M_i \succ M_j \wedge i < j\}.$$

Lastly, we have Longest Increasing Subsequence, which aptly finds the length of the longest increasing subsequence in the learned ranking. A longer LIS value indicates that the learned ranking is more aligned with the human ranking. Mathematically, this is represented by the maximum value of the array that tracks the length of increasing subsequences:

$$\mathcal{L}_{LIS}(\hat{\mathcal{R}}, \mathcal{R}^*) := \max\{dp[i] \mid 1 \leq i \leq m\},$$

$$dp[i] = 1 + \max\{dp[j] \mid 1 \leq j < i \wedge M_j \prec M_i\}.$$

Then, we augment the evaluation aspect of PiCO by incorporating the metric of Rank-Biased Overlap (RBO). This metric was proposed by Webber et al. to quantify the similarity between two ranked lists Webber et al. (2010). Unlike other metrics that treat all positions in the lists equally, RBO gives more weight to the higher ranks. This makes it useful to our approach, where differences in larger/better models' rankings are more significant than those in smaller/worse ones.

The RBO value ranges from 0 to 1, where 0 indicates no overlap between the lists and 1 indicates perfect agreement. This value is calculated using a parameter p, which determines how quickly the weight decreases for lower ranks. For example, a p value close to 1 places high importance on the top-ranked items, while a lower p value distributes the importance more evenly across all of them.A higher p value places more emphasis on the top ranks. Formally, the RBO for two ranked lists S and T is defined as:

$$RBO(S, T, p) = (1-p) \sum_{d=1}^{D} p^{d-1} \cdot \frac{|S_{1:d} \cap T_{1:d}|}{d}$$

where:

- p is the weighting parameter (typically $0 < p < 1$).
- d is the depth up to which the lists are compared.
- $S_{1:d}$ and $T_{1:d}$ are the top d elements of lists S and T, respectively.
- $|S_{1:d} \cap T_{1:d}|$ is the number of common items in the top d elements of both lists.

5.3 Experimental details

5.3.1 PiCO Evaluation

Running the PiCO Evaluation code required the use of the open-source Virtual Large Language Model (vLLM) library, so we ran all necessary shell commands in a Jupyter Notebook for convenient execution. With the paid version of Google Colaboratory, we were able to utilize a NVIDIA A100 GPU for the PiCO evaluation process. This was necessary to load the model Yi-1.5-34B-Chat, as it is much larger than the others and thus required more memory.

We had to make several modifications to the original paper's code, as it was written specifically for compatibility with the MT-Bench test dataset, which consists solely of multi-turn questions Zheng et al. (2023). Our refined version is compatible with any single-turn question dataset, which eliminates the need for data pre-processing when using single-turn evaluation sets with PiCO.

To compare against our fine-tuned model, we chose the following pool of models for ranking, listed in descending order of ELO points on the crowdsourced leaderboard Chatbot Arena. We justify the use of this leaderboard for our ground truth human ranking based on the site having over 1,000,000 votes for more than 100 models, which has led to its general recognition as a reputable standard for model performance according to human evaluation Chiang et al. (2024).

#	Model	ELO
1	gpt-4	1251
2	Yi-1.5-34B-Chat	1162
3	gpt-3.5-turbo	1103
4	vicuna-7b-v1.5	1004
5	mpt-7b-chat	927
6	chatglm2-6b	924
7	oasst-sft-4-pythia-12b	893
8	fastchat-t5-3b-v1.0	868
9	dolly-v2-12b	822

Table 2: Pool of models against which we test our fine-tuned Llama 3.

This pool was selected for its variety; it contains a mixture of large proprietary models (gpt-4, gpt-3.5-turbo) as well as smaller open-source models, some of which are trained for instruction answering (oasst-sft-4-pythia-12b, dolly-v2-12b) and the rest of which are trained for general conversational purposes (vicuna-7b-v1.5, mpt-7b-chat, chatglm3-6b, fastchat-t5-3b-v1.0, dolly-v2-12b). Additionally, this set evenly covers a wide range of ELO scores, meaning similarly-ranked models have quite similar capacities to either outperform or underperform each other. The complete pool and ground-truth ranking including our fine-tuned model is as above, but we approximate our model's ELO with that of its base model Llama-3-8B-Instruct (1153), which places it at rank 3 and shifts all models with less ELO one spot down.

5.4 Results

5.5 Llama 3 Finetuning and Prompting Techniques Evaluation

After finetuning Llama 3 and comparing our results from different methods across our test sets, we observe that the differences are relatively small. As can be seen in 3, our additional prompting techniques do not result in significant performance increases on the test sets of MMLU Clinical Knowledge, MMLU Medical Genetics, and MedQA. With that being said, we do see substantial improvements in other test sets like MMLU College Medicine and MedMCQA, and nearly every baseline was able to be improved upon by some combination of fine-tuning and prompting techniques. As displayed in 4, we see that our results are promising when compared to Med-PaLM 2, which is almost 50 times larger Singhal et al. (2023). In fact, our model even beats Med-PaLM 2 with ensemble refinement on the PubmedQA benchmark, as seen in 4.

Dataset	Base	w/ FT	w/ FT + ER	w/ FT + SC
MMLU College Medicine	60.7% ± 3.7%	61.8% ± 3.7%	**65.2% ± 1.2%**	62.3% ± 3.7%
MMLU College Biology	78.5% ± 3.4%	78.5% ± 3.4%	**79.3% ± 2.1%**	78.5% ± 3.4%
PubmedQA	75.0% ± 1.9%	75.8% ± 1.9%	**76.1% ± 1.9%**	75.1% ± 1.9%
MMLU Professional Medicine	70.2% ± 2.8%	69.1% ± 2.8%	70.1 % ± 2.5%	**73.1% ± 2.5%**
MMLU Clinical Knowledge	**74.7% ± 2.7%**	73.6% ± 2.7%	73.4 % ± 2.7%	72.9% ± 2.7%
MMLU Medical Genetics	83.0% ± 3.8%	**85.0% ± 3.6%**	**85.0% ± 3.6%**	83.5% ± 3.4%
MMLU Anatomy	67.4% ± 4.0%	68.9% ± 4.0%	**71.7 % ± 4.0%**	70.0% ± 1.8%
MedQA	59.8% ± 1.4%	**60.3% ± 1.4%**	60.2% ± 1.4%	60.1% ± 1.3%
MedMCQA	57.6% ± 0.8%	56.4% ± 0.8%	**58.4% ± 0.8%**	57.9% ± 1.2%

Table 3: Performance comparison of Llama 3 variants on medical datasets.

Dataset	Llama 3 w/ FT + ER	Med-PaLM 2 w/ ER
MMLU College Medicine	65.2%	**83.2%**
MMLU College Biology	79.3%	**95.8%**
PubmedQA	**76.1%**	75.0%
MMLU Professional Medicine	70.1%	**92.3%**
MMLU Clinical Knowledge	73.4%	**88.7%**
MMLU Medical Genetics	85.0%	**92.0%**
MMLU Anatomy	71.7%	**84.4%**
MedQA	60.2%	**85.4%**
MedMCQA	58.4%	**72.3%**

Table 4: Performance compared with Med-PaLM 2

5.5.1 PiCO Evaluation on Adversarial Dataset

After running the PiCO evaluation script on the Adversarial Dataset, we obtained the following ranking and metrics for our fine-tuned model versus the baseline model of llama-3-8B-instruct ($p = 0.75$ for RBO):

Fine-Tuned			Baseline		
#	Model	Grade	#	Model	Grade
1	gpt-4	0.2800	1	gpt-4	0.2814
2	Yi-1.5-34B-Chat	0.2613	2	Yi-1.5-34B-Chat	0.2620
3	UltimateMed-llama-3-8b-instruct	0.2606	3	llama-3-8b-instruct	0.2585
4	gpt-3.5-turbo	0.2571	4	gpt-3.5-turbo	0.2573
5	vicuna-7b-v1.5	0.2553	5	vicuna-7b-v1.5	0.2559
6	mpt-7b-chat	0.2546	6	mpt-7b-chat	0.2548
7	oasst-sft-4-pythia-12	0.2468	7	oasst-sft-4-pythia-12b	0.2473
8	chatglm2-6b	0.2440	8	chatglm2-6b	0.2452
9	fastchat-t5-3b-v1.0	0.2335	9	fastchat-t5-3b-v1.0	0.2328
10	dolly-v2-12b	0.2275	10	dolly-v2-12b	0.2273
PEN	2.617		PEN	2.617	
CIN	2		LIS	2	
LIS	7		LIS	7	
RBO	0.944		RBO	0.944	

Table 5: Final PiCO-generated model ranking and metrics

These results agree strongly with the ground-truth ranking. Our PEN could have taken on values between 0 and $\log_2(10!)$, which is approximately 21.8, so our value of 2.617 is clearly very small relatively Ning et al. (2024). We see that only two inversions appear as a result of Open-Assistant SFT-4 12B being ranked higher than ChatGLM 6B v2. Our RBO values, which follow a moderate value of 0.75 for p, show that we attain a score near 1 indicating near-perfect similarity. And, our fine-tuned model has a slightly higher grade than that of the base instruct model, indicating its responses were deemed better more often by the other LLMs.

6 Analysis

Overall, the results that we achieved were not entirely what we anticipated. For example, fine-tuning Llama 3 didn't improve performance significantly, even though the decrease in loss during fine-tuning is clearly noticable in 1. We hypothesized that our results could be a result of overfitting, which looked possible because of the fact that our model seemed to have converged by the second epoch. However, we evaluated on the checkpoint after the first epoch and achieved the same results, indicating that this was likely not the correct explanation.

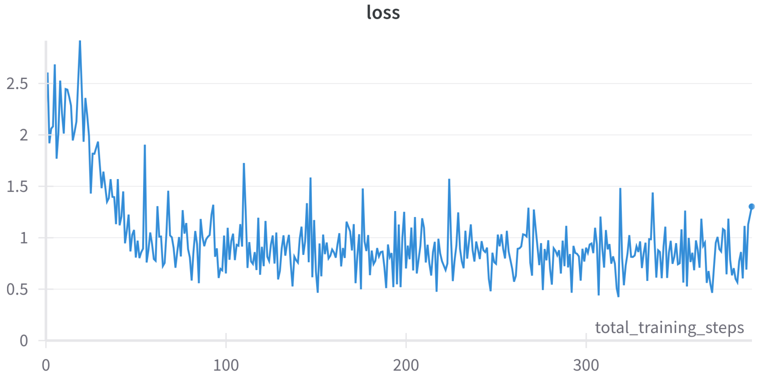

Figure 1: Loss vs. Training Steps

The PiCO evaluation saw inversion in Open-Assistant SFT-4 12B scoring higher than ChatGLM 6B v2. This could be due to the fact that the former is based on a fine-tuned model trained on human demonstrations of assistant conversations, which is similar in nature to our evaluated task, whereas the latter is meant to serve as an all-purpose billingual model for English and Chinese, which is unrelated.

7 Conclusion

Overall, we showed that building relatively robust medical question-answering LLMs with smaller open-source models is not only achievable with limited compute, but can occasionally be shown to outperform state-of-the-art models like Google's Med-PaLM 2 on certain metrics. We also demonstrated the effectiveness of LLM self-evaluation in the implementation of PiCO, which suggests that the infamously time-consuming and costly process of model alignment can be well-approximated by much simpler and more accessible methods.

As for our limitations, when using models that were trained on data on the scale of Llama 3 — 15 trillion tokens — the possibility of test data leakage becomes a real problem. While we ensured that all of our finetuning data remained free of contamination factors like these are unfortunately out of our control. The ethical implications of this far-reaching as an inability to effectively evaluate these models, but still treating such evaluations as valid presents an opportunity for these model to be deployed with their efficacies greatly exagerated, which is very dangerous for those using the information returned by these systems.

Considering possibilities for further improvement in future work, we recognize that in the preliminary stages of this project we had intended to improve upon currently-used prompting techniques Singhal et al. (2023) via a graph-of-thought approach, utilizing search. However, this proved to be a more time-intensive task than we had expected, so we could not commit to exploring its implementation. As such, this method remains a very promising avenue for future work, especially given the success seen by incorporating search in Google recent Med-Gemini publication Saab et al. (2024). Regarding the PiCO evaluation method, we saw that it was an effective judge of human preference alignment when used with our adversarial dataset, but more work could be done to investigate whether or not this holds for other kinds of datasets such as questions which are highly technical or even more sensitive in subject.

8 Ethics Statement

Our project on fine-tuning Llama 3 for medical question-answering presents several ethical challenges. One such risk is the potential to exemplify existing health inequities. As highlighted by Chen et al. in *Ethical Machine Learning in Healthcare*, machine learning models in healthcare can disproportionately underperform for minority groups. For example, language models trained on scientific articles were shown to medically recommend hospital stays to violent white patients versus prison time for violent black patients Chen et al. (2021). To mitigate this risk, we can first ensure that our training dataset includes a diverse range of demographic and clinical scenarios, with biased outliers removed as necessary. A more empirical solution would be to experiment with different loss functions, as it has been found that the choice of error metric to minimize may be responsible for downstream bias in model performance; e.g. surrogate loss functions such as the hinge loss are often chosen for computational efficiency, but have been shown to disproportionately affect undersampled groups in the training data because of approximation errors Lohaus et al. (2020). By measuring the effects of different loss functions on bias in model performance, developers can choose the most fair options as necessary in medical settings.

Another critical ethical concern is the risk of misinformation. Large language models generate responses based on patterns in data rather than genuine understanding of medical content, which can lead to the generation of incorrect or misleading medical advice. This can have severe implications for patient health and safety – for example, a language model might suggest an incorrect treatment or fail to recognize a critical symptom, which could put the user's safety, health, and even life at risk. One algorithmic method of reducing the risk of misinformation is data augmentation, by which we not only train on the original samples in our dataset but also on slightly modified versions of those examples – for example rotated and translated versions of medical image data – in order to improve our model's ability to discern between even slight differences in input, thus reducing rates of misinformation due to the model confounding similar features.

References

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2021. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4:123–144. Epub 2021 May 6.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering.

Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. 2020. Too relaxed to be fair. In *International Conference on Machine Learning*, pages 6360–6369. PMLR.

Zea Malawa et al. 2021. Racism as a root cause approach: A new framework. *Pediatrics*, 147(1):e2020015602.

Kun-Peng Ning, Shuo Yang, Yu-Yang Liu, Jia-Yu Yao, Zhen-Hui Liu, Yu Wang, Ming Pang, and Li Yuan. 2024. Pico: Peer review in llms based on the consistency optimization.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G.T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of gemini models in medicine. *Google Research*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28:20:1–20:38.

David R Williams et al. 2019. Understanding how discrimination can affect health. *Health Services Research*, 54 Suppl 2:1374–1388.

Tamara Williams et al. 2015. The reliability of ahrq common format harm scales in rating patient safety events. *Journal of Patient Safety*, 11(1):52–59.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.