EYE-Llama, an in-domain large language model for ophthalmology

Authors: Tania Haghighi[1,2], Sina Gholami[1], Jared Todd Sokol[3], Enaika Kishnani[1], Adnan Ahsaniyan[2], Holakou Rahmanian[2], Fares Hedayati[2], Theodore Leng[3], and Minhaj Nur Alam[1,*]

[1] Department of Electrical Engineering, University of North Carolina at Charlotte, Charlotte, NC, United States
[2] Department of Computer Science, Baha'i Institute for Higher Education, Tehran, Iran
[3] Department of Ophthalmology Stanford University School of Medicine, Stanford, CA, United States

Correspondence*:
9201 university city boulevard, Charlotte, NC 28223
minhaj.alam@charlotte.edu

Funding:
Supported by NEI R15EY035804 (MNA) and UNC Charlotte Faculty Research Grant (MNA)

Abstract

Background:

Training Large Language Models (LLMs) with in-domain data can significantly enhance their performance, leading to more accurate and reliable question-answering (QA) systems essential for supporting clinical decision-making and educating patients.

Methods:

This study introduces LLMs trained on in-domain, well-curated ophthalmic datasets. We also present an open-source substantial ophthalmic language dataset for model training. Our LLMs (EYE-Llama), first pre-trained on an ophthalmology-specific dataset, including paper abstracts, textbooks, EyeWiki, and Wikipedia articles. Subsequently, the models underwent fine-tuning using a diverse range of QA datasets. The LLMs at each stage were then compared to baseline Llama 2, ChatDoctor, and ChatGPT (GPT3.5) models, using four distinct test sets, and evaluated quantitatively (Accuracy, F1 score, and BERTScore) and qualitatively by two ophthalmologists.

Results:

Upon evaluating the models using the American Academy of Ophthalmology (AAO) test set and BERTScore as the metric, our models surpassed both Llama 2 and ChatDoctor in terms of F1 score and performed equally to ChatGPT, which was trained with 175 billion parameters (EYE-Llama: 0.57, Llama 2: 0.56, ChatDoctor: 0.56, and ChatGPT: 0.57). When evaluated on the MedMCQA test set, the fine-tuned models demonstrated a higher accuracy compared to the Llama 2 and ChatDoctor models (EYE-Llama: 0.39, Llama 2: 0.33, ChatDoctor: 0.29). However, ChatGPT outperformed EYE-Llama with an accuracy of 0.55. When tested with the PubmedQA set, the fine-tuned model showed improvement in accuracy over both the Llama 2, ChatGPT, and ChatDoctor models (EYE-Llama: 0.96, Llama 2: 0.90, ChatGPT: 0.93, ChatDoctor: 0.92).

Conclusion:

The study shows that pre-training and fine-tuning LLMs like EYE-Llama enhances their performance in specific medical domains. Our EYE-Llama models surpass baseline Llama 2 in all evaluations, highlighting the effectiveness of specialized LLMs in medical QA systems. (Funded by NEI R15EY035804 (MNA) and UNC Charlotte Faculty Research Grant (MNA).)

Introduction

LLMs have revolutionized the field of Natural Language Processing (NLP) with their ability to generate human-like text and excel in various tasks. In the medical domain, LLMs can simplify the responsibilities of healthcare professionals by automating routine tasks, enabling them to concentrate more on patient care. Additionally, LLMs can enhance personalized treatment and patient engagement by providing educational materials and self-monitoring tools.

In recent years, numerous researchers have assessed LLMs in fields that demand extensive knowledge, such as medicine, by evaluating their capacity to answer multiple-choice (MC) and open-ended questions [1, 2, 3]. Among the general-purpose models, ChatGPT [4] has demonstrated exceptional performance in answering both types of questions. Llama 2 [5] has also shown proficiency in answering queries, despite having fewer parameters. It is also important to note that the performance of these models can be enhanced by pre-training and fine-tuning on in-domain data. For instance, ChatDoctor [6], BioGPT [1] and PMC_LLaMA [7] are LLMs trained on medical data specifically to answer medical questions. While these models have demonstrated impressive performance in QA, they still require further refinement to provide more accurate answers in specialized fields like medical subspecialties, including ophthalmology.

This study presents Eye-Llama LLMs, pre-trained and fine-tuned on in-domain ophthalmic data. Subsequently, the performance of our models was compared with existing baseline models designed for general (Llama 2 and ChatGPT models) and biomedical (ChatDoctor model) domains. We have also made the code and dataset available to others to facilitate further research based on our benchmark results.

The contributions of our study are as follows:

- We have developed and open-sourced a comprehensive dataset in ophthalmology. The data is composed of roughly 744k samples collected from PubMed abstracts, 22k samples from nearly 570 textbooks, articles from EyeWiki, and the ophthalmology category on Wikipedia. Additionally, our dataset includes supervised data, comprising approximately 18k QA pairs from medical datasets, 1.5k QA pairs from medical forums, as well as 15k QA pairs generated using GPT3.5.
- We have proposed a framework for developing an open-source model that was trained on in-domain datasets and tailored to specific areas within medicine, such as ophthalmology. (https://github.com/QIAIUNCC/EYE-Llama).
- We have investigated and validated our model against baseline LLMs, to prove the feasibility of our hypothesis, that is: in-domain dataset enhances LLM performance, even in smaller LLMs with reduced size and trainable parameters.

Method

A schematic description of our training framework is provided in Figure **1**. We implemented a two-stage training approach on the Llama 2 (Llama 2-7b-chat) model to tailor it to the ophthalmology field. This involved pre-training and fine-tuning phases, utilizing data from the ophthalmology domain. In the initial stage, the model was pre-trained over ophthalmology text corpus resulting in the EYE-Llama_p model. In the subsequent step, the model underwent further refinement, utilizing a curated dataset comprising a diverse set of ophthalmic QA datasets. This process yielded the fine-tuned models, EYE-Llama_qa and EYE-Llama_gqa. Compared to EYE-Llama_qa, EYE-Llama_gqa was fine-tuned using a more extensive QA dataset, which included an additional 15k QAs.

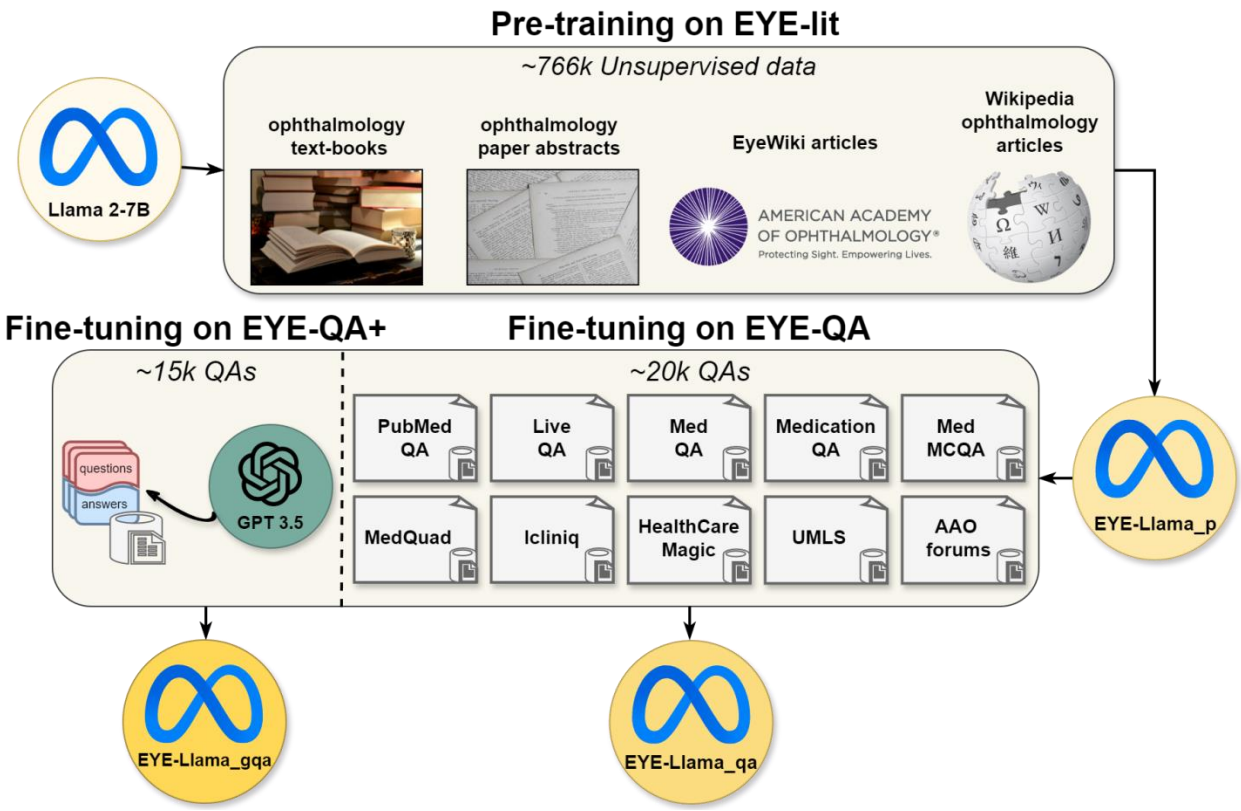

Figure 1 Schematic representation of the Llama 2 model's pre-training and fine-tuning process for specialization in ophthalmology. EYE-Llama_p was trained using approximately 766k unsupervised data. It was then fine-tuned with 20k QA from various medical datasets, the AAO forum and 15k questions and answers generated by ChatGPT.

Dataset:

Unsupervised data:

We collected approximately 744k ophthalmic abstracts from the PubMed webpage (https://pubmed.ncbi.nlm.nih.gov). To facilitate this, we employed specific ophthalmic keywords and

leveraged the BioPython Entrez database search system to interact with NCBI's PubMed database. Our aim was to retrieve abstracts that included any specific pertinent ophthalmic keywords. The comma-separated keywords used for retrieving both PubMed abstracts and the subset of supervised questions were the following: *ophthalmology, eye, Conjunctiva, Retina, Sclera, Cone cell, Choroid, Macula, Cataract, Strabismus, Optic nerve, Macular degeneration, Ciliary body, Ophthalmologist, Pinkeye, Astigmatism, Diabetic eye disease, diabetic retinopathy, age-related macular degeneration, glaucoma, iris, cornea, optometry, geographic atrophy*

Our dataset was expanded using 570 textbooks and full articles from EyeWiki (https://eyewiki.org) and Wikipedia (https://en.wikipedia.org/wiki/Category:Ophthalmology) articles specifically categorized under ophthalmology.

Data cleaning:

The abstracts obtained from PubMed were initially in XML format. Using XML files, we extracted raw texts and removed any special characters or phrases that did not provide additional information, such as "in this paper". It is crucial to highlight that the texts from textbooks, EyeWiki, and Wikipedia articles were divided into smaller data chunks as we needed to set a maximum size of text when feeding them to our model. We set the maximum size to 512 tokens and split the text into paragraphs. We ensured that each data chunk contained relevant information and that sentences were complete within each chunk. Notably, the abstracts retrieved from PubMed did not require this segmentation process, as they consisted of brief paragraphs. We named this diverse collection of data, the EYE-lit dataset. The dataset formed the foundation for our exploration into the pre-training of our model and has been made available at *huggingface.co/datasets/QIAIUNCC/EYE-lit.*

Supervised data:

Supervised fine-tuning was undertaken by selecting a QA subset from the comprehensive medical dataset employed in [7]. This was combined with subsets from the Icliniq and HealthCareMagic datasets, as introduced in [6], and a subset from the MedQuad dataset [8]. This selection process involved leveraging specific ophthalmic keywords, the same keywords employed during the initial collection of PubMed abstracts. The resultant subset comprised 18k samples, with roughly 6k samples sourced from each of the PubMedQA [9] and MedMCQA [10] datasets for the training phase. In addition, approximately 800 samples from each of the PubMedQA and MedMCQA datasets were utilized for testing purposes.

Following the initial steps, we prepared a QA dataset from the AAO forum webpage (https://www.aao.org/eye-health/ask-ophthalmologist). This dataset, containing approximately 1,527 samples, was integrated with the 18k samples from the supervised dataset that was created in the preceding step for training purposes. We also created a test set using an additional 1,307 samples from the same forum.

We merged all these different datasets into a more extensive and comprehensive dataset, called the EYE-QA dataset.

EYE-QA dataset:

- Subset of medical QA datasets (18k samples)
    - PMC_LLaMA_instructions dataset (contains subsets of several medical datasets)
    - Icliniq and HealthCareMagic datasets
    - MedQuad dataset
- QA pairs from the AAO forum (1.5k samples)

To improve the model's ability to respond to more specialized questions, such as general disease inquiries or those an ophthalmologist might ask an AI assistant, we generated approximately 15k ophthalmic QAs using the ChatGPT (GPT3.5- turbo). This data was then combined with the existing EYE-QA dataset, thereby creating an expanded dataset that we refer to as EYE-QA+ (*available at huggingface.co/datasets/QIAIUNCC/EYE-QA-PLUS*). To generate this data we used ophthalmic textbooks and prompted ChatGPT to generate a question and an answer to that question given one page of a textbook (see the prompts in Appendix A).

Model Training:

We trained a model using Llama 2, a substantial model containing 7 billion parameters, as a foundation. The training process was executed in two distinct stages, employing a maximum length of 512, a learning rate of 2e-4, and an effective batch size of 256. The initial stage involved pre-training the model on the EYE-lit dataset resulting in the EYE-Llama_p model. Subsequently, in the second stage, we fine-tuned the model using the EYE-QA dataset to develop the EYE-Llama_qa model and EYE-QA+ to develop the EYE-Llama_gqa model.

Quantized Low-Rank Adaptation (QLoRA):

To train this model efficiently on limited resources, we employed a technique called QLoRA [11]. This approach has proven to be effective in optimizing the use of available resources during the training process. QLoRA addresses the computational limitation challenge by quantizing the weights of an LLM and employing low-rank adapters to adapt the LLM to a new task. In this study, the LoRA_r value, representing the rank of the low-rank approximation, was set to 16. The LoRA_alpha value, a scaling factor adjusting the learning rate for the low-rank parameters, was set to 64. These hyper parameters were chosen to optimize the balance between model performance and computational efficiency.

With the QLoRA method, we utilized two 49G GPUs to train the model optimizing approximately 12 million trainable parameters from the total number of 7 billion parameters of the Llama 2 model. The pre-training phase involved 24,400 steps. Each model, EYE-Llama_qa and EYE-Llama_gqa, underwent a fine-tuning phase. The EYE-Llama_qa model required 300 steps for fine-tuning and the EYE-Llama_gqa model required 500 steps.

Evaluation:

We assess the results produced by LLMs using a variety of test sets that encompass both MCQAs and open-ended QAs. Our test sets include a 4-option QA dataset with 780 questions (MedMCQA), a yes/no QA dataset comprising 800 questions (PubMedQA), a collection of 1307 patient/ophthalmologist QAs from the AAO forum, and a set of 20 questions (20-Questions dataset) specifically designed by an ophthalmologist to qualitatively assess the LLMs.

MCQA Evaluation

The performance of our LLMs in answering MC questions was evaluated using the MedMCQA and PubMedQA datasets, with accuracy and F1 scores as the metrics. The supervised data and the PubMedQA and MedMCQA test sets may contain some questions that are not entirely related to ophthalmology. This is because we collected the supervised data using certain ophthalmic keywords, which might have resulted in the inclusion of some non-ophthalmic content.

It is important to note that each LLM responds to each MCQ that includes not just the chosen option, but also a justification for that choice or why other options were considered wrong. The model earns 1 point for correct answers, indicating task comprehension, and 0 points for incorrect, multiple answers, or no answer, indicating misunderstanding or errors. This system effectively measured model performance but was conducted by non-ophthalmologists, which may introduce inaccuracies.

For evaluating the MCQA test sets, we used the following prompt for the EYE-Llama_qa, EYE-Llama_gqa, and ChatGPT models:

System prompt: {instruction}

User prompt: {question}

And the following prompt for the ChatDoctor:

"{instruction}\n\nPatient: {question}\n\nChatDoctor: "

Open-ended QA Evaluation

The AAO patient/ophthalmologist QA dataset and 20-Questions dataset were used to evaluate LLM's proficiency in generating open-ended answers. The AAO dataset, which includes 1307 QA pairs from an ophthalmology forum, provided a broad range of real-world patient inquiries and expert responses. The model's performance on this dataset was quantified using the BERTScore metric [12].

The 20-Questions dataset, designed by an ophthalmologist, served as a more focused test, with the model's responses qualitatively evaluated by two ophthalmologists. This dataset contains questions that patients do not typically ask to an ophthalmologist. Instead, they are the type of questions that an ophthalmologist would pose to an AI assistant. They do not describe a patient's symptoms or discomforts; rather, they are designed to assess a problem from a medical perspective. Examples of these types of questions include: "what is the best therapy to start on a patient with a diagnosis of primary open-angle glaucoma?", and "what is the best method for retinal detachment repair?".

For the open-ended QA datasets, we used the recommended prompt for ChatDoctor and the following prompt for evaluating the EYE-Llama_qa, EYE-Llama_gqa and ChatGPT models on this dataset:

System prompt: "Given your profession as an ophthalmologist, please provide a detailed and comprehensive response to the question."

User prompt: {question}

Results

In this section, we provide an overview of our results as summarized in tables.

MCQA

Table 1 Performance comparison of various models on the MedMCQA dataset, with ChatGPT leading in both accuracy and F1-score.

MedMCQA		
Model	**Accuracy**	**F1-score**
Llama 2-7b-chat	0.33	0.19
EYE-Llama_p	0.34	0.21
EYE-Llama_qa	0.39	0.23
EYE-Llama_gqa	0.35	0.22
ChatDoctor	0.29	0.17
ChatGPT (GPT3.5)	**0.55**	**0.37**

Table 2 Evaluation of different models on the PubMedQA dataset, with EYE-Llama_qa showing the highest accuracy and F1-score.

PubMedQA		
Model	**Accuracy**	**F1-score**

Llama 2-7b-chat	0.9	0.45
EYE-Llama_p (without supervised fine-tuning)	0.74	0.34
EYE-Llama_qa	**0.96**	**0.74**
EYE-Llama_gqa	0.93	0.5
ChatDoctor	0.92	0.69
ChatGPT (GPT3.5)	0.93	0.52

A noticeable performance decline is observed in the EYE-Llama_p model on the PubMedQA dataset. This can be attributed to the fact that the Llama 2 model, being chat-based, loses its task-specific knowledge (answering a question based on the provided context) after being fine-tuned on unsupervised data. Nevertheless, the model relearns this task during the second fine-tuning phase, resulting in an improvement over its previous performance.

Open-ended QA

AAO dataset

BERTScore Metric

We used BERTScore to evaluate the model's performance with the AAO dataset. We used the ophthalmologists' answers from the AAO forum as the reference answer and the answers generated by each model as the candidate answer. BERTScore is an automatic evaluation metric for text generation that uses cosine similarity between the reference and candidate sentences to match words and computes the precision, recall, and F1 score.

Table 3 Performance comparison of various models reveals that the EYE-Llama_qa and EYE-Llama_gqa models demonstrate a balanced performance with F1 scores.

AAO			
Model	**Precision**	**Recall**	**F1**
Llama 2-7b-chat	0.5194	0.6189	0.5630
EYE-Llama_qa	0.5420	0.6152	0.5742
EYE-Llama_gqa	0.5483	0.6108	**0.5758**
ChatDoctor	**0.5933**	0.5434	0.5648

ChatGPT	0.5295	**0.6352**	0.5757

A Dataset of 20-Questions

This dataset is used for a qualitative investigation aimed at understanding the preferences of ophthalmologists. Despite the presence of automated evaluation metrics, the emphasis lies on the subjective ranking of various models by these specialists within their domain. The 20-Questions dataset was both designed and evaluated by two ophthalmologists. To evaluate each answer in this dataset, the ophthalmologist answered the following questions:

- Is the answer aligned with the scientific community (alignment)?
- Is the question understood (Grasped)?
- Mention of relevant and/or correct facts for answering (relevant)?
- Presence of inaccurate or irrelevant information (irrelevant)?
- How much harm can the answer cause from 0 to 5(harm)?
- What is the answer's overall score from 0 to 10?

While the last two questions were scored directly, the responses to the first four questions were rated by the ophthalmologists using qualitative terms. These terms correspond to specific scores: 'Excellent' or 'yes' correspond to a score of 1, 'Mostly' equates to a score of 0.75, 'Somewhat' or 'Some' are equivalent to a score of 0.5, 'Mild' corresponds to a score of 0.35, 'Minimal' is equivalent to a score of 0.25, and 'No' corresponds to a score of 0.

Table 4 This table shows the summation of all twenty LLM scores provided by ophthalmologists.

20-Questions Set						
Model	**Score /200**	**Alignment /20**	**Grasped /20**	**Irrelevant /20**	**Relevant /20**	**Harm /100**
Llama 2	100	12	20	19	17.25	54
EYE-Llama_qa	81	8.5	19.5	18	15.5	61
EYE-Llama_gqa	116	14	20	11.45	17.25	51
ChatDoctor	89.5	10	20	15	14.5	51
ChatGPT (GPT3.5)	**160**	**19**	**20**	**6**	**19.75**	**28**

Discussion

In this study, we trained and evaluated our models, EYE-Llama_qa and EYE-Llama_gqa, in comparison with baseline LLMs such as ChatGPT, ChatDoctor, and Llama 2. The evaluation was conducted using MCQA and open-ended QA datasets. Our results revealed that, upon fine-tuning, both EYE-Llama_qa and EYE-Llama_gqa models exhibited improvements over the respective base model. Specifically, EYE-Llama_gqa excelled in open-ended QA, while EYE-Llama_qa led in MCQA datasets. This enhancement allowed for higher accuracy in answering subspecialty MC and open-ended questions.

In our quantitative assessment of LLMs, EYE-Llama_qa, and EYE-Llama_gqa outperformed both Llama 2 and ChatDoctor, achieving superior F1 scores and accuracy on the PubMedQA and MedMCQA datasets. Notably, EYE-Llama_qa outperformed ChatGPT in terms of accuracy and F1 score on the PubMedQA dataset but fell slightly behind ChatGPT on the MedMCQA dataset. For the AAO dataset, BERTScore was employed as an evaluation metric. ChatDoctor excelled in Precision, outperforming other models, but its lower recall suggests that the higher precision may be attributed to the brevity of its responses compared to others. When evaluating the F1 score, both EYE-Llama_qa and EYE-Llama_gqa surpassed ChatDoctor and Llama 2, with their performance being roughly equivalent to ChatGPT.

When qualitatively assessing the model's responses to the 20-Questions dataset, it is evident that EYE-Llama_gqa performs on par with or better than ChatDoctor, Llama 2, and EYE-Llama_qa across all axes. However, ChatGPT outperforms all other models (Table **4**), with several factors contributing to these results. The high parameter count of ChatGPT (175B) played a significant role in its strong performance across all four test sets. The ChatGPT model is considerably larger (GPT3.5 has 175B and Llama 2 has 7B parameters) and requires extensive computing resources. However, EYE-Llama_qa and EYE-Llama_gqa, with only 7B parameters, showed marked improvement over the Llama 2 model and outperformed the ChatDoctor model across all three datasets. ChatDoctor used Llama 2-7b and patient dialogue data, while we used the chat-optimized Llama 2 model and ophthalmic literature and QA pairs. Different base models coupled with different diverse data types significantly influenced the performance of each model. Explicitly, EYE-Llama_qa and EYE-Llama_gqa were trained on ophthalmic data, while ChatDoctor was trained on general medical data. Future work could include expanding datasets, refining metrics for ophthalmic queries, and involving more ophthalmologists for evaluation. While this study focused on the ophthalmic domain, applying the approach to other medical domains could yield interesting results.

Studies have evaluated the performance of LLMs like ChatGPT, a widely recognized model, in ophthalmology [13, 14]. Findings have indicated that GPT-4 achieved an impressive score of approximately 81% on the Ophthalmic Knowledge Assessment Program (OKAP) examination, surpassing the average human score of approximately 74% [15]. Bernstein et al. [16] Evaluated the quality of ophthalmology advice from an LLM chatbot compared to ophthalmologists. It found that a panel of eight ophthalmologists could distinguish between human and chatbot responses to 200 eye care questions with 61% accuracy.

Various strategies have been employed to adapt general-purpose LLMs to the medical domain, such as Med-Palm 2, BioGPT, Med-Alpaca, and PMC_LLaMA. Research has also been conducted in medical subspecialties. For instance, [17] demonstrated the superior performance of GPT-4 and Claude 2 over

open-source LLMs on a MCQA dataset within the nephrology domain. Despite this research that leverages open-source LLMs, our study aimed to further refine these models through additional pre-training and fine-tuning over newly generated ophthalmic datasets similar to Ophtha-llama2 [18] and OphGLM [19].

It is crucial to acknowledge the limitations of this study, one of which is the limited availability of a supervised QA dataset in the field of ophthalmology. A more extensive dataset would likely produce more precise outcomes. Moreover, a significant portion of the QA data used in this study was synthetically generated using GPT-3.5. While this model produced reasonable QAs, it may not perfectly reflect real-world data. The second constraint was resource limitations. We employed QLoRA to train large models with limited resources. Future research could benefit from access to more extensive computational resources to avoid this trade-off. Lastly, the feedback received from ophthalmologists on the responses generated by our models was limited. Therefore, incorporating more expert feedback would likely enhance the quality of our findings.

In conclusion, further exploration into the capabilities of LLMs within the medical domain is warranted and may lead to significant advancements to improve patient care and efficiency. Looking ahead, we foresee the application of LLMs to various aspects of the healthcare field, including underserved medicine, medical trainee education, remote patient monitoring, and personalized patient care. LLMs have a promising future in the medical field and will hopefully enable clinicians to better care for patients and improve patient outcomes.

References

[1] Renqian Luo, Liai Sun, Yingce Xia, et al, "BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining," *Briefings in Bioinformatics, Volume 23, Issue 6,* vol. 23, no. 6, 2022.

[2] Karan Singhal, Tao Tu, Juraj Gottweis, et al, "Towards Expert-Level Medical Question Answering with Large Language Models," *arXiv preprint arXiv:2305.09617v1,* 2023.

[3] Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou et al, "MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data," *arXiv preprint arXiv:2304.08247,* 2023.

[4] OpenAI, "ChatGPT," *https://chat.openai.com,* 2023.

[5] Hugo Touvron, Louis Martin, Kevin Stone, et al, "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv preprint arXiv:2307.09288,* 2023.

[6] Z Yunxiang Li, Zihan Li , Kai Zhang , Ruilong Dan , Steve Jiang , You Zhang, "ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge," *Cureus,* vol. 15, no. 6, 2023.

[7] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, Weidi Xie, "PMC-LLaMA: Towards Building Open-source Language Models for Medicine," *arXiv preprint arXiv:2307.09288,* 2023.

[8] Asma Ben Abacha, Dina Demner-Fushman, "A Question-Entailment Approach to Question

Answering," *{BMC} Bioinform.,* vol. 20, pp. 511:1--511:23, 2019.

[9] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, Xinghua Lu, "PubMedQA: A Dataset for Biomedical Research Question Answering," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* pp. 2567--2577, 2019.

[10] Ankit Pal, Logesh Kumar Umapathi, Malaikannan Sankarasubbu, "MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering," *PMLR,* vol. 174, pp. 248-260, 2022.

[11] Dettmers, Tim and Pagnoni, Artidoro and Holtzman, "QLoRA: Efficient Finetuning of Quantized LLMs," *arXiv preprint arXiv:2305.14314,* 2023.

[12] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi, "BERTScore: Evaluating Text Generation with BERT," in *International Conference on Learning Representations (ICLR)*, 2020.

[13] Ting DSJ, Tan TF, Ting DSW, "ChatGPT in ophthalmology: the dawn of a new era?," *Eye,* vol. 38, pp. 4-7, 2023.

[14] Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, Renaud Duval, "Evaluating the Performance of ChatGPT in Ophthalmology," *Ophthalmology Science,* vol. 3, no. 4, 2023.

[15] Sean Teebagy, Lauren Colwell, Emma Wood, Antonio Yaghy, Misha Faustina, "Improved Performance of ChatGPT-4 on the OKAP Exam: A Comparative Study with ChatGPT-3.5," *Journal of Academic Ophthalmology,* vol. 15, pp. e184-e187, 2023.

[16] Isaac A. Bernstein, Youchen (Victor) Zhang, Devendra Govi, et al, "Comparison of Ophthalmologist

and Large Language Model Chatbot Responses to Online Patient Eye Care Questions," *JAMA Network Open,* vol. 6, no. 8, 2023.

[17] Sean Wu, Michael Koo , Lesley Blum , Andy Black, et al, "Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology," *NEJM AI,* vol. 1, no. 2, 2024.

[18] Huan Zhao, Qian Ling, Yi Pan, et al, "Ophtha-LLaMA2: A Large Language Model for Ophthalmology," *arXiv preprint arXiv:2312.04906v1.*

[19] Weihao Gao, Zhuo Deng, Zhiyuan Niu, et al, "OphGLM: Training an Ophthalmology Large Language-and-Vision Assistant based on Instructions and Dialogue," *arXiv preprint arXiv:2306.12174v2.*