Ophtimus-LLM: Development of a Specialized Large Language Model for Ophthalmology

Seung Ju Baek,[2] Kuk Jin Jang,[1] Sooyong Jang,[1] Hyonyoung Choi, [1] Minwook Kwon, [2] Yong Seop Han, [3] Seongjin Lee, [2] Jin Hyun Kim, [2] Insup Lee [1]

[1] PRECISE Center & Department of Computer and Information Science, University of Pennsylvania,
[2]Department of AI Convergence Engineering, Gyeongsang National University, [3]Changwon Hanmaeum Hospital
{jangkj, sooyong, hyonchoi, lee}@seas.upenn.edu,
seungju41625886@gmail.com, medcabin@hanmail.net, {minwook1125, insight, jin.kim}@gnu.ac.kr

Abstract

The development of domain-specific language models has become increasingly important in healthcare, where the complexity and precision of medical knowledge often exceed the capabilities of general-purpose large language models (LLMs). This study introduces Ophtimus-LLM, a compact 8-billion-parameter LLM tailored for ophthalmology. Key findings demonstrate that scalability laws observed in larger models also hold true for smaller, domain-specific LLMs, suggesting that well-designed compact models can achieve high performance. Additionally, the study highlights the critical role of data quality in boosting model accuracy, with significant gains observed when training on domain-relevant content. Ophtimus-LLM exemplifies the potential of specialized LLMs to provide efficient, accessible, and high-performing tools for advancing medical AI while addressing challenges of scalability and equity in healthcare technology.

1 Introduction

The proliferation of large language models (LLMs) has unlocked opportunities to transform various domains, including healthcare (Anil et al. 2023; Jiang et al. 2024; Yang et al. 2022). As the capabilities of state-of-the-art (SOTA) general-purpose LLMs expand, concerns regarding accessibility, equity, and practicality in their deployment continue to grow (Pierson et al. 2025; Weidinger et al. 2022).

Access to computational resources and equitable technology distribution significantly influence healthcare outcomes (Nambisan and Nambisan 2017; Bajwa et al. 2021). Healthcare systems worldwide face persistent disparities in access to quality care, driven by socioeconomic, geographic, and systemic barriers (Mullins et al. 2005; Tzenios 2019). AI has the potential to bridge some of these gaps by providing decision-support tools, automating repetitive tasks, and improving diagnostic accuracy (Elhaddad and Hamam 2024; Aminizadeh et al. 2024). Yet, the computational demands and high costs of deploying state-of-the-art general-purpose models could make them inaccessible to resource-constrained healthcare settings, particularly in underserved and low-income regions (Wahl et al. 2018; Hu et al. 2025).

Moreover, despite their impressive versatility, general-purpose LLMs frequently fail to capture the specialized knowledge required for domains like ophthalmology (Zhao et al. 2023; Haghighi et al. 2024). The lack of domain-specific expertise in these models can lead to suboptimal or incorrect clinical recommendations, undermining trust in AI. Furthermore, privacy concerns and exposing sensitive information to third-party services must be addressed for trustworthy deployment (Wu, Duan, and Ni 2024; Yan et al. 2024; Yao et al. 2024). The costs of running large-scale models and the concerns about trustworthiness highlight the need for an alternative solution.

In this situation, smaller, more efficient, and domain-specific models emerge as a compelling solution (Sanh 2019; Jiao et al. 2019; Sun et al. 2020; Zhang et al. 2024). The focus on the targeted application allows these models to deliver high-performance outcomes with significantly lower computational overhead. Additionally, the smaller size allows for less data to achieve comparable performance to their counterparts. However, several challenges remain in determining the best approach for developing such domain-specific solutions. First, there is a high cost for curating quality datasets. Creating high-quality datasets for healthcare AI requires significant resources, as domain-specific data often needs to be sourced from paid academic journals, clinical guidelines, and textbooks (Chia et al. 2024). Beyond acquisition, ensuring data accuracy and relevance demands expert validation, meticulous annotation, and rigorous filtering. Privacy regulations, such as de-identifying patient data, further increase the time and financial investment needed to prepare datasets suitable for training reliable medical models. Second, there may be high computational costs for training and deploying the model (Samsi et al. 2023). Training language models involves significant computational costs, requiring specialized hardware such as GPUs or TPUs to process large datasets and optimize billions of parameters. This demands substantial energy consumption and infrastructure, increasing both financial and environmental costs. Deployment also carries computational overhead, especially for real-time applications, as serving large models requires high-performance servers with consistent uptime (Griggs et al. 2024). These costs can be prohibitive, particularly for smaller organizations or resource-constrained settings, limiting widespread accessibility and scalability.

In this study, we develop and present *Ophtimus-LLM*, a set of LLMs tailored for Ophthalmology. Ophtimus-LLM was trained on high-quality domain-specific data and fine-tuned to optimize its performance on ophthalmology-related tasks. We explore various approaches for curating domain-specific data and computationally efficient methods for fine-tuning. We summarize the main contribution of this study:

1. We present Ophtimus-LLM, a set of lightweight LLMs tailored for ophthalmology with empirical results demonstrating our approach's effectiveness that combines pre-training and fine-tuning.
2. We present an approach for collecting a high-quality dataset for fine-tuning and demonstrating the effectiveness of model performance.
3. Results from ablation studies demonstrating the importance of data quality and model parameter size for improved performance.

This work highlights the need for specialized AI models to improve inclusivity, efficiency, and equity in healthcare.

2 Related Work

General-purpose LLMs and Medical LLMs. Many general-purpose large language models (LLMs) have been introduced in recent years (Achiam et al. 2023; Anil et al. 2023; Team et al. 2024; Dubey et al. 2024). These models have demonstrated success across a wide range of tasks, such as question answering (Anil et al. 2023) and code generation (Jiang et al. 2024). Their applications are now extending into the medical domains (Lehman et al. 2023). Many medical LLMs have been developed, demonstrating promising results in tasks such as medical question answering, generating chest X-ray reports, and performing on the United States Medical Licensing Exam (Yang et al. 2022; Singhal et al. 2023a,b; Kung et al. 2023; Ayers et al. 2023; Tu et al. 2024; Han et al. 2023; Saab et al. 2024). A systematic review of medical LLMs is available in (Thirunavukarasu et al. 2023).

LLMs for Ophtahlmology. There is a growing demand for smaller, domain-specific LLMs tailored to particular fields within medicine, for example, in ophthalmology. In experiments on the Ophthalmic Knowledge Assessment Program (OKAP), the United States Medical Licensing Examination (USMLE), and the Board of Ophthalmology Written Qualifying Examination, general-purpose models lack in ophthalmology-specific performance (Antaki et al. 2023; Haddad, Saade et al. 2024; Shemer et al. 2024). To address this need, various ophthalmology-specialized LLMs have been proposed (Tan et al. 2024; Zhao et al. 2023; Haghighi et al. 2024; Singer et al. 2024; Chen et al. 2024b; Gilson et al. 2024; Chen et al. 2024a). These models are generally obtained by fine-tuning general LLMs and are designed for different tasks, such as disease diagnosis, knowledge-based question answering (QA), and long-form QA. Further reviews on ophthalmology LLMs can be found in (Betzler et al. 2023; Sevgi et al. 2024; Wong et al. 2024).

Approaches and need for small-scale LLMs. Despite the capabilities of LLMs, several small language models (SLMs) have been developed due to their lower computational costs both during training and inference. Examples include DistilBERT (Sanh 2019), TinyBERT (Jiao et al. 2019), MobileBERT (Sun et al. 2020), and TinyLlama (Zhang et al. 2024). SLMs are typically created using techniques such as pruning, knowledge distillation, and quantization. For further details on the training methodologies, applications, and trustworthiness of SLMs, refer to comprehensive surveys on this topic (Wang et al. 2024).

3 Methods

3.1 Dataset Curation

In developing domain-specific language models, curating high-quality datasets is a crucial step that greatly influences the model's performance and usefulness. Unlike general-purpose language models that are trained on large and varied datasets, domain-specific models require datasets that are both comprehensive and precise to maintain relevance to their specialized fields. This is particularly crucial in medicine, where accuracy and contextual understanding are essential for clinical applications.

For Ophthimus-LLM, we aim to construct a dataset that encompasses the expertise of ophthalmology while ensuring **accuracy** and **diversity**, enabling the model to demonstrate **reliability** and **utility** in real-world clinical settings.

3.2 Base Model

The base models for the proposed Ophtimus-LLM were Meta LLaMA 3, LLaMA 3.1, and LLaMA 3.2. These were selected for their proven performance and efficient architecture. All versions provide strong foundational language capabilities while being computationally efficient, which is crucial for creating a highly accurate specialized model with manageable resource requirements. In this work, we explore combining a self-supervised pre-training phase along with a fine-tuning stage with an instruction question-answer (QA) dataset. We detail our method for both datasets in Sec. 3.3 (pre-training) and Sec. 3.4 (fine-tuning).

3.3 Pre-training Dataset

The pre-training dataset for the development of Ophtimus-LLM is curated through a systematic and rigorous process to ensure high-quality and domain-specific content, as illustrated in Fig. 2. The key steps involved in constructing the pre-training dataset are as follows:

1. **Source Data Collection.** A list of 25 ophthalmology-specific keywords was selected to capture a broad spectrum of topics within the field. These keywords included "glaucoma," "cataracts," "retinal detachment," and other crucial terms relevant to ophthalmic practice and research. Based on these keywords, a pre-training dataset is constructed using PubMed. This database was selected for its extensive collection of expert-reviewed biomedical literature, resulting in the collection of 11,487 open-access articles related to 25 specific keywords.

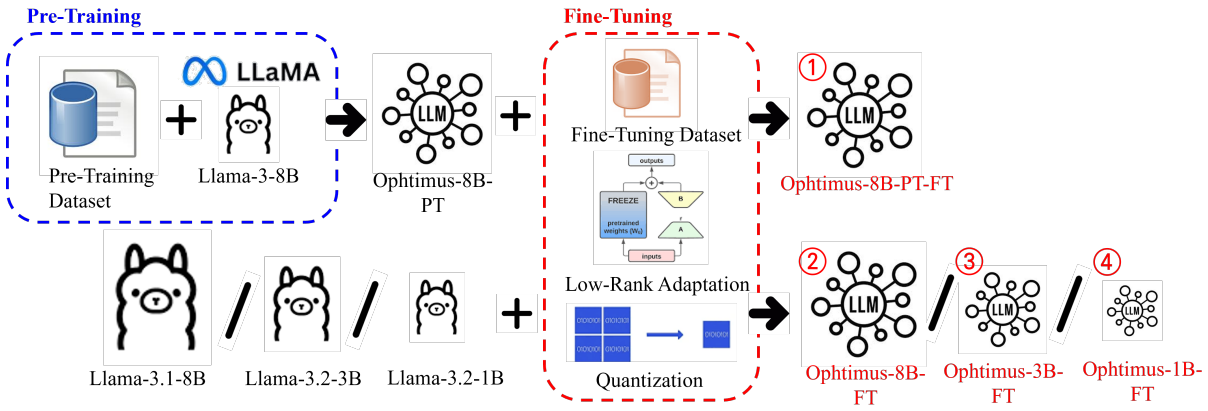

Figure 1: Overview of training process of Ophtimus-LLM

2. **Data Filtering.** Source data was filtered according to a language and relevance criterion. For language, only English-language papers were retained. For relevance, GPT-4o was used to remove papers unrelated to ophthalmology, despite containing relevant keywords. This process reduced the dataset to approximately 9,200 papers.
3. **Content Summarization.** The filtered papers are preprocessed using GPT-4o and the Map-Reduce method to extract key information on diseases, symptoms, treatments, and clinical cases, resulting in high-quality, domain-focused summaries.
4. **Additional Data Preprocessing.** Deduplication of the text was used to remove redundancy, retaining only unique content. In addition, Personally identifiable information (PII) was redacted. Ultimately, a corpus of approximately 12.2 million tokens was derived from around 9,200 carefully curated papers. This dataset provides a comprehensive foundation for pre-training, enhancing the model's performance in ophthalmology-specific tasks.

3.4 Fine-tuning Dataset

The fine-tuning dataset enhances domain expertise by adding specialized knowledge to the pre-trained model. It is also designed to ensure that the model provides answers in the correct format according to given instructions. A high-quality QA dataset was constructed as follows (See Fig. 4):

1. **Source Data Collection.** Using publicly accessible resources, we curated a collection of several ophthalmology textbooks totaling over 9,000 pages of source material. The manuscripts were selected to ensure high content accuracy and diversity. We extracted the textual information from each page of the manuscripts.
2. **Content Summarization.** Similar to the pretraining, the textbook manuscripts undergo preprocessing with GPT-4o to extract essential information while omitting unnecessary content such as publisher details, figure descriptions, and other irrelevant elements. This process produces high-quality, domain-specific summaries. This helps avoid creating data that may hinder training, such as domain-irrelevant or non-generalizable questions, during the Ophthalmology QA generation.
3. **Ophthalmology QA Generation.** Two types of QA pairs are generated using GPT-4o:
   - Descriptive/Essay QA (EQA): QAs that need descriptions of symptoms, treatments, and disease diagnoses.
   - Multi-Choice QA (MCQA): QAs in yes/no or multiple choice (with 4 or 5 options) formats. In addition, explanations were generated to enhance the model's accuracy and clarity.

   A total of 28,000 MCQA pairs and 47,000 EQAs were generated. The MCQAs and EQAs covered a wide range

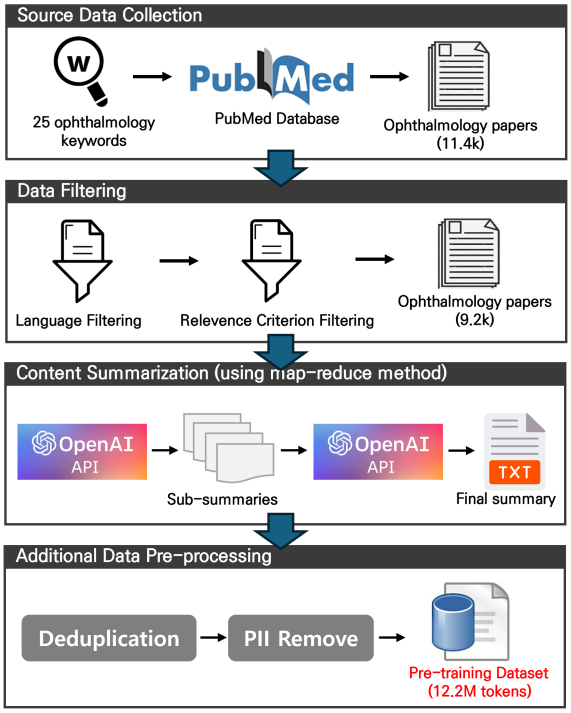

Figure 2: Pre-Training Dataset Building Process Pipeline

Dataset	Year	# of Ophthal QA	
		Original	Ophth-related
Multi-Choice QA (MCQA)			
(a) Ophtimus-Eval	-	-	2,156
(b) MedMCQA	2022	182k	6,932
(c) PubMedQA	2019	11k	298
Total	-	-	9,386
Essay QA (EQA)			
(a) MedQuAD	2019	47k	667
(b) Medical Flashcards	2023	34k	543
(c) Medical Wikwdoc	2023	10k	179
Total	-	-	1,389

Table 1: Evaluation Dataset

of topics for a comprehensive evaluation of the model's knowledge.

3.5 Fine-tuning Methodology

Fig. 1 illustrates the various configurations of pretraining and fine-tuning that were applied to develop our model. Four approaches were explored, differing primarily in the choice of the foundation model used for fine-tuning. In the first approach, we pre-train the Llama-3-8B base model with our ophthalmology-specific pretraining dataset (Sec. 3.3) (*Ophtimus-8B-PT-FT*). Second, to evaluate the effect of pretraining, we only fine-tune Llama-3.1-8B (*Ophtimus-8B-FT*). For the third and fourth models, in order to investigate model size, we fine-tune the 3B and 1B variants of Llama-3.2 (*Ophtimus-3B-FT*, *Ophtimus-1B-FT*, respectively).

Quantization and LoRA. Reducing computational costs is a key aim of the study. For this quantization and Low-Rank Adaptation (LoRA) techniques were applied to optimize efficiency during fine-tuning. Quantization from 16-bit to 4-bit precision reduced GPU memory by approximately 4X. LoRA was applied to all layers, with $\texttt{lora\_alpha} = 16$ and $\texttt{rank} = 32$.

4 Evaluation and Results

4.1 Evaluation Dataset

For the evaluation, MCQAs were curated from publicly available resources, independent of training material. A total of 2156 MCQA from 19 subfields of ophthalmology, which we call *Ophtimus-Eval*. Questions were reviewed for validity and overall quality. Note that although curated from publicly accessible sources, we will limit access until further validation and provide access with proper permissions. In addition, we selected ophthalmology-specific QAs from available medical benchmark datasets for evaluation, including MedMCQA, PubMedQA, MedQuAD, and other verified medical QA sources. Details on the datasets and number of data entries can be found in Table 1 and Sec. D.

4.2 Evaluation Metrics

For MCQA, answer accuracy was computed. For EQA, Rouge-L (Lin 2004), BLEU (Papineni et al. 2002), METEOR (Lavie and Agarwal 2007), and SemScore (Aynetdinov and Akbik 2024) was used. In particular, SemScore is an evaluation metric for assessing LLM outputs by measuring semantic similarity to reference responses, offering a closer alignment with human judgment compared to traditional metrics like BLEU or ROUGE. Additional description can be found in Sec. E.

4.3 Benchmark Comparison Models

Eye-Llama (Haghighi et al. 2024) is a specialized 7B-parameter LLM for ophthalmology, built on LLaMA 2 and pre-trained on domain-specific texts like PubMed abstracts, textbooks, and EyeWiki articles to capture ophthalmic knowledge. It was fine-tuned using diverse ophthalmology-focused QA datasets, including MedMCQA, PubMedQA, and patient-doctor discussions from the American Academy of Ophthalmology (AAO) forum.

PMC-Llama (Wu et al. 2024) is a 13B-parameter open-source LLM based on Meta's LLaMA, designed for medical applications through two training stages: knowledge injection with medical corpora pre-training and medical-specific instruction tuning using QA and conversational datasets. This enables it to outperform general-purpose models on medical QA benchmarks, making it a robust and scalable solution for clinical and research applications.

4.4 Results

Overall Results. Table 2 describes the overall results of the evaluation. In general, GPT-4o outperformed all models evaluated in this study. This is unsurprising, considering the scale of GPT-4o. Even so, on our specialized dataset (Ophtimus-Eval), GPT-4o had lower performance. This indicates the relative uniqueness of the questions curated.

Outside of GPT-4o, our model, Ophtimus-8B-PT-FT, showed the best performance on the Ophtimus-Eval dataset and comparable performances to other models on the other benchmark datasets. Eye-LLama and PMC-Llama were trained on the PubMedQA dataset, which is expected to lead to better performance. Nonetheless, the comparable performance of our models on ophthalmology-specific questions is noteworthy. The effect of pre-training appears minimal when comparing the 8B-FT and 8B-PT-FT. We suspect it is due to the limited diversity of the source data and insufficient data volume. Further exploration is needed.

Evidence of Scaling Laws. An interesting observation of our results is that the model performance seems to increase as the model size increases. Fig. 3 depicts an example of questions and the subsequent answers by the models of various sizes. It can be observed that the length and accuracy of the answer seem to improve as the model size increases. This scaling phenomenon has been observed in many instances throughout the literature. However, the performance of PMC-Llama-13B on our evaluation dataset was significantly lower, indicating that model size is not the only factor that can improve performance in specialized domains. These factors are explored in Section 4.5.

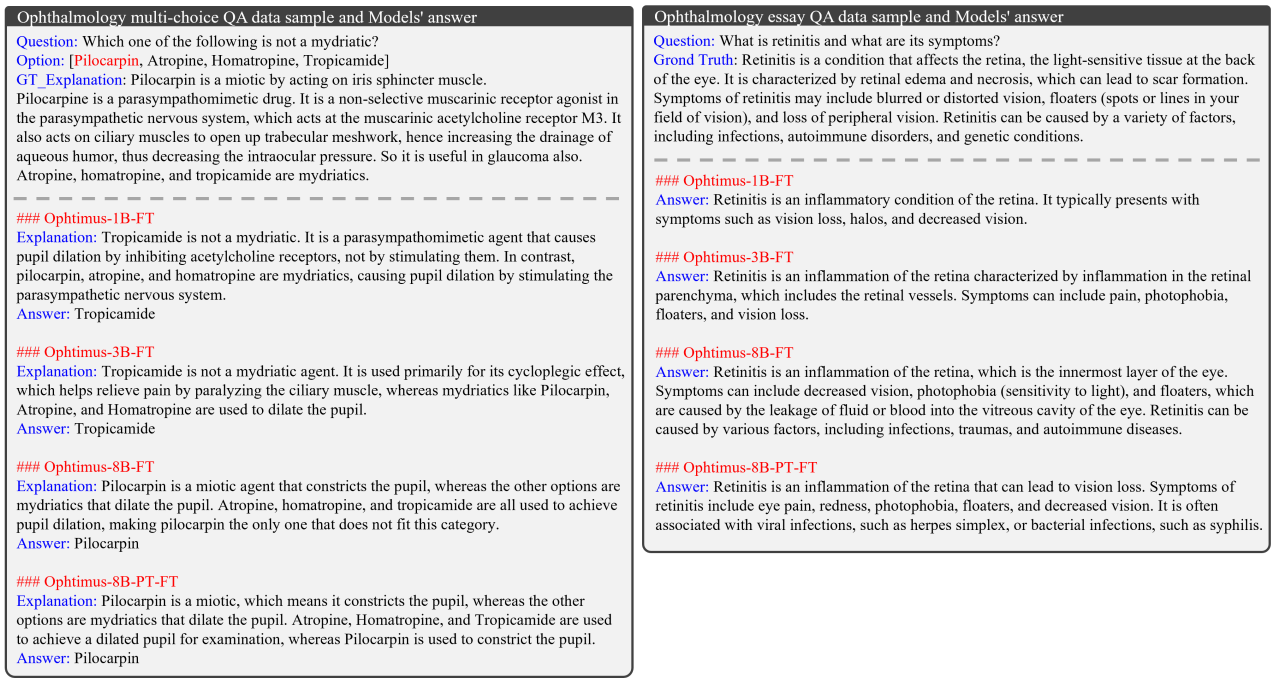

Figure 3: Comparison of Ophtimus-LLM models' answers to MCQA/EQA sample.

4.5 Ablation Studies

To better understand the factors influencing model performance, we present the results of several ablation studies.

Fine-tuning data quality. In order to illustrate the quality of our curated dataset, using the same methodology described in Sec. 3.4, 12,600 questions were derived from the PubMed pre-training dataset described in Sec. 3.3. In total, 6,300 MCQA pairs and 6,300 EQA pairs were created. The same number of questions were uniformly randomly chosen from the textbook-based fine-tuning QA dataset. The overall results can be seen in Table 3. The model trained on the textbook-based QA dataset demonstrated slightly better performance on our Ophtimus-Eval dataset and the ophthalmology questions from the MedMCQA dataset. Conversely, for the ophthalmology questions in the PubMedQA dataset, the opposite was true. This phenomenon is expected as the terminology and concepts in PubMed are likely have similar characteristics. Future studies could potentially quantify these effects. The smaller difference in the Ophtimus-Eval dataset can be attributed to the disproportionate number of questions on specific topics, such as general ophthalmology.

When analyzing the performance across the 19 topics in our Ophtimus-Eval dataset, as shown in Table 4, fine-tuning on the textbook dataset demonstrated equal or superior performance compared to the base Llama-3.1 model for 13 out of the 19 topics. In contrast, the model fine-tuned on the PubMed-based dataset only showed improved performance for 6 out of the 19 topics. Overall, either fine-tuned model outperformed the base model for most of the topics. These results demonstrate the quality of the dataset curated with the proposed approach in Sec. 3.1.

Improving Performance on Individual Topics. The performance comparison of the models across various topics is detailed in Appendix B.1. Among all the topics, the models performed the worst in the optics category. This lower performance may be attributed to a lack of QA samples in the fine-tuning dataset for that category compared to the others.

To explore this, we derived an additional 1090 MCQA samples from an text source not included in the initial fine-tuning dataset. The Ophtimus-8B-PT-FT model was trained for an additional five epochs on the additional samples. As can be seen in Fig. 5 in Appendix B.2 and Table 5, the performance on the topic increases to 40.3% after 4 epochs.

The additional training on the optics topic affected performance in other areas. A detailed breakdown of the changes by topic is available in Table 7 in Appendix B.2. According to Table 5, there was an overall performance decline, resulting in a score of 56.61%. This decline may be attributed to the limited capacity of smaller models. Further research is necessary to improve performance without compromising the knowledge that has already been learned.

Evaluation of early stopping. Training for multiple epochs is less common with large-scale datasets and LLMs. However, in our experiments, we found that when using a validation set that comprised 1% of the training data, performance improved after several epochs. To determine the optimal number of epochs, we trained each model for a maximum of 5 epochs and evaluated its performance on the evaluation dataset. Detailed results can be found in Table 8 in Appendix C. The findings suggest that while multiple epochs can be beneficial, implementing an early stopping criterion may be necessary to prevent overfitting.

Model	Multi-Choice Question			Essay Question			
	Ophtimus Eval	**MedMCQA (Ophth)**	**PubmedQA (Ophth)**	**RougeL**	**BLEU**	**METEOR**	**SemScore**
OpenAI GPT-4o	**71.95%**	**81.95%**	**89.90%**	0.193	0.082	**0.341**	**0.761**
Llama-3-8B-Instrct	48.60%	74.02%	63.97%	0.193	0.064	0.244	0.684
Llama-3.1-8B-Instrct	39.78%	57.96%	83.84%	0.177	0.054	0.215	0.641
Eye-Llama	32.56%	59.43%	66.11%	0.183	0.062	0.211	0.686
PMC-Llama-13B	48.28%	63.45%	72.48%	0.223	0.082	0.288	0.714
Ophtimus-1B-FT	34.77%	38.44%	68.46%	0.219	0.076	0.217	0.711
Ophtimus-3B-FT	46.01%	51.01%	69.80%	0.224	0.077	0.225	0.726
Ophtimus-8B-FT	57.78%	59.49%	72.48%	**0.226**	**0.083**	0.230	0.733
Ophtimus-8B-PT-FT	59.13%	58.82%	71.14%	0.222	0.079	0.224	0.735

Table 2: Overall results of evaluation for various models.

Dataset	Multi-Choice Question		
	Our Eval Dataset Ophtimus Eval	**MedMCQA (Ophth.)**	**PubMedQA (Ophth.)**
PubMed	48.09%	58.71%	78.11%
Textbook	50.18%	64.05%	54.54%

Table 3: Comparison of training data quality.

Topic	Model		
	Llama-3.1 8B-Instruct	**Ophtimus-8B (PubMed)**	**Ophtimus-8B (Textbook)**
ANTERIOR SEGMENT	26.7%	26.7%	**36.7%**
Cataract	37.5%	37.5%	37.5%
Conjunctiva	33.3%	66.7%	**71.4%**
Cornea	33.0%	36.4%	**42.0%**
Error of Refraction	36.8%	47.4%	**52.6%**
General Ophthalmology	42.4%	52.5%	**55.8%**
Glaucoma	40.5%	50.3%	50.3%
Neuro Ophthalmology	32.1%	46.7%	**47.3%**
Ocular Trauma	53.6%	**67.9%**	60.7%
Oculoplastic	36.4%	27.3%	**36.4%**
Optics	**34.7%**	32.2%	31.8%
ORBIT_LIDS_ADNEXA	47.9%	47.0%	**49.6%**
Pathology	47.8%	64.2%	**67.2%**
PEDIATRICS_STRABISMUS	34.5%	45.5%	45.5%
PHARMACOLOGY	55.0%	55.0%	**57.5%**
POSTERIOR SEGMENT	33.3%	40.0%	40.0%
Retina & vitreous	30.8%	30.8%	23.1%
Systemic diseases	**62.5%**	37.5%	50.0%
UVEITIS	17.5%	**50.0%**	45.0%

Table 4: Comparison of performance on various topics. Bold indicates the top score in a topic and underlines indicate the next best.

5 Discussion

Quantitative metrics for data quality. In this study, we demonstrated the importance of high-quality, diverse datasets for training models in specialized domains like ophthalmology, where dataset richness directly impacts performance. However, reliable methods for evaluating data quality remain underdeveloped. Developing quantitative metrics to evaluate dataset diversity and alignment with clinical requirements may result in curating datasets that more accurately reflect the complexities of target domains.

Topic	**Model**	
	Ophtimus-PT-FT	**w/ Optics FT**
Optics	**37.0%**	**40.3% (+3.3%)**
Overall Performance	59.13%	56.61%

Table 5: Comparison of performance after fine-tuning on additional Optics dataset.

Limitations in evaluation for clinical utility. While this study showcases the potential of domain-specific models like Ophtimus-LLM in ophthalmology, evaluating their clinical utility remains challenging. Standardized datasets and metrics often fail to capture the nuances of real-world clinical scenarios. Traditional benchmarks do not fully address critical factors such as interpretability, robustness, and workflow integration. To overcome these limitations, new evaluation frameworks must incorporate clinician feedback, simulate real-world conditions, and assess the broader impact on healthcare delivery.

6 Conclusion

In this work, we developed Ophtimus-LLM, a compact yet high-performing language model tailored to ophthalmology. By leveraging carefully curated datasets and efficient training techniques, we demonstrated robust performance without depending on resource-intensive, large-scale models. This highlights the importance of high-quality, diverse data, particularly in resource-limited settings.

Future work includes expanding the framework to other medical domains to address diverse healthcare challenges. This involves incorporating multimodal inputs like text and imaging to enhance prediction accuracy and conducting real-world clinical testing to validate utility and integration. By pursuing these directions and addressing challenges like computational costs and clinical evaluation metrics, models such as Ophtimus-LLM can advance accessible, equitable, and effective AI-driven healthcare.

Acknowledgements

This paper was supported in part by National Research Foundation of Korea (Ministry of Education, Science and Technology) NRF-2023R1A2C1006639, NIH 1R01EY037101, and ARO W911NF-20-1-0080.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aminizadeh, S.; Heidari, A.; Dehghan, M.; Toumaj, S.; Rezaei, M.; Navimipour, N. J.; Stroppa, F.; and Unal, M. 2024. Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service. *Artificial Intelligence in Medicine*, 149: 102779.

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Antaki, F.; Touma, S.; Milad, D.; El-Khoury, J.; and Duval, R. 2023. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmology science*, 3(4): 100324.

Ayers, J. W.; Poliak, A.; Dredze, M.; Leas, E. C.; Zhu, Z.; Kelley, J. B.; Faix, D. J.; Goodman, A. M.; Longhurst, C. A.; Hogarth, M.; et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6): 589–596.

Aynetdinov, A.; and Akbik, A. 2024. SemScore: Automated Evaluation of Instruction-Tuned LLMs based on Semantic Textual Similarity. ArXiv:2401.17072.

Bajwa, J.; Munir, U.; Nori, A.; and Williams, B. 2021. Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal*, 8(2): e188–e194.

Ben Abacha, A.; and Demner-Fushman, D. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20: 1–23.

Betzler, B. K.; Chen, H.; Cheng, C.-Y.; Lee, C. S.; Ning, G.; Song, S. J.; Lee, A. Y.; Kawasaki, R.; van Wijngaarden, P.; Grzybowski, A.; et al. 2023. Large language models and their impact in ophthalmology. *The Lancet Digital Health*, 5(12): e917–e924.

Chen, X.; Xu, P.; Li, Y.; Zhang, W.; Song, F.; He, M.; and Shi, D. 2024a. ChatFFA: An ophthalmic chat system for unified vision-language understanding and question answering for fundus fluorescein angiography. *Iscience*, 27(7).

Chen, X.; Zhao, Z.; Zhang, W.; Xu, P.; Gao, L.; Xu, M.; Wu, Y.; Li, Y.; Shi, D.; and He, M. 2024b. EyeGPT: Ophthalmic Assistant with Large Language Models. *arXiv preprint arXiv:2403.00840*.

Chia, M. A.; Antaki, F.; Zhou, Y.; Turner, A. W.; Lee, A. Y.; and Keane, P. A. 2024. Foundation models in ophthalmology. *British Journal of Ophthalmology*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Elhaddad, M.; and Hamam, S. 2024. AI-driven clinical decision support systems: an ongoing pursuit of potential. *Cureus*, 16(4).

Gilson, A.; Ai, X.; Xie, Q.; Srinivasan, S.; Pushpanathan, K.; Singer, M. B.; Huang, J.; Kim, H.; Long, E.; Wan, P.; et al. 2024. Language Enhanced Model for Eye (LEME): An Open-Source Ophthalmology-Specific Large Language Model. *arXiv preprint arXiv:2410.03740*.

Griggs, T.; Liu, X.; Yu, J.; Kim, D.; Chiang, W.-L.; Cheung, A.; and Stoica, I. 2024. M\'elange: Cost efficient large language model serving by exploiting gpu heterogeneity. *arXiv preprint arXiv:2404.14527*.

Haddad, F.; Saade, J. S.; et al. 2024. Performance of ChatGPT on ophthalmology-related questions across various examination levels: observational study. *JMIR Medical Education*, 10(1): e50842.

Haghighi, T.; Gholami, S.; Sokol, J. T.; Kishnani, E.; Ahsaniyan, A.; Rahmanian, H.; Hedayati, F.; Leng, T.; and Alam, M. N. 2024. EYE-Llama, an in-domain large language model for ophthalmology. *bioRxiv*.

Han, T.; Adams, L. C.; Papaioannou, J.-M.; Grundmann, P.; Oberhauser, T.; Löser, A.; Truhn, D.; and Bressem, K. K. 2023. MedAlpaca–An Open-Source Collection of Medical Conversational AI Models and Training Data. *arXiv preprint arXiv:2304.08247*.

Hu, S.; Oppong, A.; Mogo, E.; Collins, C.; Occhini, G.; Barford, A.; and Korhonen, A. 2025. Natural Language Processing Technologies for Public Health in Africa: A Scoping Review (Preprint).

Jiang, J.; Wang, F.; Shen, J.; Kim, S.; and Kim, S. 2024. A Survey on Large Language Models for Code Generation. *arXiv preprint arXiv:2406.00515*.

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W. W.; and Lu, X. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Kung, T. H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2): e0000198.

Lavie, A.; and Agarwal, A. 2007. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, 228–231. USA: Association for Computational Linguistics.

Lehman, E.; Hernandez, E.; Mahajan, D.; Wulff, J.; Smith, M. J.; Ziegler, Z.; Nadler, D.; Szolovits, P.; Johnson, A.; and

Alsentzer, E. 2023. Do we still need clinical language models? In *Conference on health, inference, and learning*, 578–597. PMLR.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Mullins, C. D.; Blatt, L.; Gbarayor, C. M.; Yang, H.-W. K.; and Baquet, C. 2005. Health disparities: a barrier to high-quality care. *American Journal of Health-System Pharmacy*, 62(18): 1873–1882.

Nambisan, S.; and Nambisan, P. 2017. How should organizations promote equitable distribution of benefits from technological innovation in health care? *AMA journal of ethics*, 19(11): 1106–1115.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260. PMLR.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Pierson, E.; Shanmugam, D.; Movva, R.; Kleinberg, J.; Agrawal, M.; Dredze, M.; Ferryman, K.; Gichoya, J. W.; Jurafsky, D.; Koh, P. W.; et al. 2025. Using Large Language Models to Promote Health Equity.

Saab, K.; Tu, T.; Weng, W.-H.; Tanno, R.; Stutz, D.; Wulczyn, E.; Zhang, F.; Strother, T.; Park, C.; Vedadi, E.; et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Samsi, S.; Zhao, D.; McDonald, J.; Li, B.; Michaleas, A.; Jones, M.; Bergeron, W.; Kepner, J.; Tiwari, D.; and Gadepally, V. 2023. From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. ArXiv:2310.03003.

Sanh, V. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sevgi, M.; Ruffell, E.; Antaki, F.; Chia, M. A.; and Keane, P. A. 2024. Foundation models in ophthalmology: opportunities and challenges. *Current Opinion in Ophthalmology*, 10–1097.

Shemer, A.; Cohen, M.; Altarescu, A.; Atar-Vardi, M.; Hecht, I.; Dubinsky-Pertzov, B.; Shoshany, N.; Zmujack, S.; Or, L.; Einan-Lifshitz, A.; et al. 2024. Diagnostic capabilities of ChatGPT in ophthalmology. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 1–8.

Singer, M. B.; Fu, J. J.; Chow, J.; and Teng, C. C. 2024. Development and evaluation of Aeyeconsult: a novel ophthalmology chatbot leveraging verified textbook knowledge and GPT-4. *Journal of Surgical Education*, 81(3): 438–443.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Sun, Z.; Yu, H.; Song, X.; Liu, R.; Yang, Y.; and Zhou, D. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*.

Tan, T. F.; Elangovan, K.; Jin, L.; Jie, Y.; Yong, L.; Lim, J.; Poh, S.; Ng, W. Y.; Lim, D.; Ke, Y.; et al. 2024. Fine-tuning Large Language Model (LLM) Artificial Intelligence Chatbots in Ophthalmology and LLM-based evaluation using GPT-4. *arXiv preprint arXiv:2402.10083*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; and Ting, D. S. W. 2023. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. 2024. Towards generalist biomedical AI. *NEJM AI*, 1(3): AIoa2300138.

Tzenios, N. 2019. The determinants of access to healthcare: a review of individual, structural, and systemic factors. *Journal of Humanities and Applied Science Research*, 2(1): 1–14.

Wahl, B.; Cossy-Gantner, A.; Germann, S.; and Schwalbe, N. R. 2018. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ global health*, 3(4): e000798.

Wang, F.; Zhang, Z.; Zhang, X.; Wu, Z.; Mo, T.; Lu, Q.; Wang, W.; Li, R.; Xu, J.; Tang, X.; et al. 2024. A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness. *arXiv preprint arXiv:2411.03350*.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 214–229.

Wong, M.; Lim, Z. W.; Pushpanathan, K.; Cheung, C. Y.; Wang, Y. X.; Chen, D.; and Tham, Y. C. 2024. Review of emerging trends and projection of future developments in large language models research in ophthalmology. *British Journal of Ophthalmology*, 108(10): 1362–1370.

Wu, C.; Lin, W.; Zhang, X.; Zhang, Y.; Xie, W.; and Wang, Y. 2024. PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, ocae045.

Wu, X.; Duan, R.; and Ni, J. 2024. Unveiling security, privacy, and ethical concerns of ChatGPT. *Journal of Information and Intelligence*, 2(2): 102–115.

Yan, B.; Li, K.; Xu, M.; Dong, Y.; Zhang, Y.; Ren, Z.; and Cheng, X. 2024. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*.

Yang, X.; Chen, A.; PourNejatian, N.; Shin, H. C.; Smith, K. E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A. B.; Flores, M. G.; et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1): 194.

Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; and Zhang, Y. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.

Zhang, P.; Zeng, G.; Wang, T.; and Lu, W. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Zhao, H.; Ling, Q.; Pan, Y.; Zhong, T.; Hu, J.-Y.; Yao, J.; Xiao, F.; Xiao, Z.; Zhang, Y.; Xu, S.-H.; et al. 2023. Ophtha-llama2: A large language model for ophthalmology. *arXiv preprint arXiv:2312.04906*.

A Fine-tuning Dataset Curation Pipeline

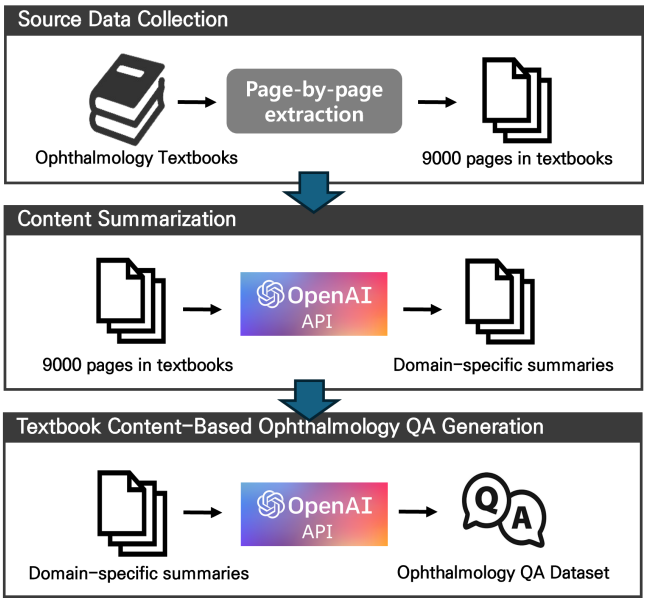

Figure 4: Fine-tuning Dataset Building Process Pipeline

B Additional Results

B.1 Comparison of Results by Topic

Table 6 reports the performance of each model on the various topics. See Sec. 4.5 for a detailed discussion of the results.

B.2 Comparison of performance after fine-tuning on additional Optics dataset

Table 7 shows the performance after fine-tuning on additional questions generated specifically for optics. The improvement in the optics category can be observed. Interestingly, 10 out of the 19 topics also demonstrated an improvement from this additional fine-tuning. However, as shown in Table 5, overall performance decreases. This can be attributed to the large number of general ophthalmology QA pairs in the evaluation dataset.

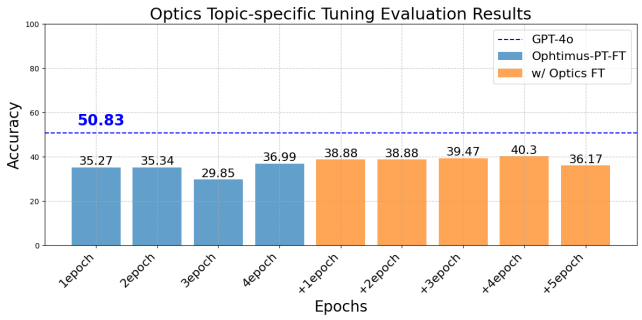

Figure 5: Evaluation results when additional trained on optics-specific QA samples.

Fig. 5 depicts the accuracy on the optics topic with additional epochs of training. It can be observed that the maximum of 40.4% is achieved after 4 epochs. This demonstrates the need for an early-stopping criterion.

C Detailed evaluation by training epoch

Table 8 presents the results of training for additional epochs. The optimal model was selected using a validation set of 1% of the training data. Generally, the best performance was achieved after 4 epochs of training. Future work will need to determine a better early stopping criterion, or if multiple iterations of training are needed.

D Description of Evaluation Datasets

MedMCQA (Pal, Umapathi, and Sankarasubbu 2022). MedMCQA is a four-choice MCQ dataset developed for evaluating question-answering capabilities in the medical domain. The questions span various areas—including pathology, pharmacology, and clinical scenarios—and are designed to mirror standardized exam formats. Consequently, the dataset allows for an extensive assessment of medical knowledge, as well as foundational reasoning skills essential for medical practice.

PubMedQA (Jin et al. 2019). PubMedQA is composed of questions derived from biomedical paper abstracts on PubMed. Each question is answered with a "yes" or "no," testing the model's ability to read, understand, and analyze scientific abstracts. Covering a broad range of biomedical and medical research topics, PubMedQA serves as an effective benchmark for evaluating both reading comprehension and specialized domain knowledge.

Topic	Model				
	ChatGPT-4o	**Ophtimus-1B-FT**	**Ophtimus-3B-FT**	**Ophtimus-8B-FT**	**Ophtimus-8B-PT-FT**
ANTERIOR SEGMENT	**63.33%**	43.33%	53.33%	53.33%	50.00%
Cataract	**75.00%**	37.50%	56.25%	62.50%	62.50%
Conjunctiva	**85.48%**	23.81%	42.86%	57.14%	52.38%
Cornea	**72.73%**	31.82%	45.45%	58.82%	53.41%
Error of Refraction	**78.95%**	36.84%	47.37%	57.89%	63.16%
General Ophthalmology	**74.37%**	36.92%	48.88%	60.95%	67.19%
Glaucoma	**76.46%**	38.65%	53.37%	61.35%	55.00%
Neuro Ophthalmology	**78.18%**	29.70%	40.61%	57.58%	52.12%
Ocular Trauma	**75.00%**	46.43%	53.57%	64.29%	53.57%
Oculoplastic	**63.64%**	36.36%	27.27%	45.45%	45.45%
Optics	**50.83%**	23.14%	31.82%	35.54%	36.99%
ORBIT_LIDS_ADNEXA	**76.07%**	37.61%	41.03%	57.26%	55.56%
Pathology	**73.13%**	44.78%	58.21%	68.66%	67.32%
PEDIATRICS_STRABISMUS	**65.45%**	23.64%	40.00%	61.82%	61.82%
PHARMACOLOGY	**87.50%**	42.50%	62.50%	67.50%	75.00%
POSTERIOR SEGMENT	**71.11%**	44.44%	42.22%	62.22%	57.78%
Retina & vitreous	**61.54%**	23.08%	30.77%	**61.54%**	30.77%
Systemic diseases	**87.50%**	50.00%	50.00%	62.50%	62.50%
UVEITIS	**70.00%**	27.50%	50.00%	**70.00%**	60.00%

Table 6: Comparison of Results by Topic (Highest in bold, 2nd highest underlined)

Topic	Model	
	Ophtimus-PT-FT	**w/ Optics FT**
ANTERIOR SEGMENT	50.0%	63.3% (+13.3%)
Cataract	50.0%	62.5% (+12.5%)
Conjunctiva	52.4%	57.1% (+4.7%)
Cornea	53.4%	58.0% (+4.6%)
Error of Refraction	63.2%	57.9% (-5.3%)
General Ophthalmology	67.2%	61.0% (-6.2%)
Glaucoma	55.0%	47.2% (-7.8%)
Neuro Ophthalmology	52.1%	58.8% (+6.7%)
Ocular Trauma	53.6%	67.9% (+14.3%)
Oculoplastic	45.5%	36.4% (-9.1%)
Optics	**37.0%**	**40.3% (+3.3%)**
ORBIT_LIDS _ADNEXA	55.6%	60.7% (+5.1%)
Pathology	67.3%	67.2% (-0.1%)
PEDIATRICS_STRABISMUS	61.8%	47.3% (-14.5%)
PHARMACOLOGY	75.0%	67.5% (-7.5%)
POSTERIOR SEGMENT	57.8%	44.4% (-13.4%)
Retina & vitreous	30.8%	38.5% (+8.5%)
Systemic diseases	62.5%	50.0% (-12.5%)
UVEITIS	60.0%	57.5% (-2.5%)

Table 7: Comparison of performance after fine-tuning on additional Optics dataset.

MedQuaAD (Ben Abacha and Demner-Fushman 2019). MedQuAD (Medical Question Answering Dataset) consists of question–answer pairs extracted from credible medical websites, including those associated with the U.S. National Institutes of Health (NIH). Each question is paired with a concise, evidence-based answer that underscores the importance of factual verification. The diverse clinical coverage in MedQuAD makes it an excellent resource for assessing how accurately a model can retrieve, interpret, and convey medically relevant information. **Medical Flashcards/Wikidoc (Han et al. 2023).** Medical Flashcards/Wikidoc is a dataset used at Stanford University for training the Alpaca model. In contrast to the multiple-choice format of the other datasets, these questions are open-ended (essay-style), sourced from medical flashcards and wiki-based materials. The questions range from basic medical knowledge to more complex clinical scenarios, requiring succinct yet precise responses. This dataset thus provides a robust environment for testing the model's ability to produce accurate, fact-based answers in an unstructured format.

E Description of Descriptive/Essay Question (EQA) Metrics

Rouge-L (Lin 2004). Rouge-L is a recall-oriented metric that looks for the longest common subsequence between the reference and the candidate.

BLEU (Papineni et al. 2002). The BLEU (BiLingual Evaluation Understudy) is a metric that was originally developed for the automatic quality evaluation of machine-translated texts. The BLEU metric is a corpus-level metric based on the modified n-gram precision measure with a length penalization for the candidate sentences that are shorter than the reference ones.

METEOR (Lavie and Agarwal 2007). The METEOR score evaluates text generation by comparing it to references, considering synonyms, stemming, and word order. It combines precision, recall (with more weight on recall), and penalizes word order errors for better alignment with human judgment.

SemScore (Aynetdinov and Akbik 2024). SemScore is an evaluation metric for assessing LLM outputs by measuring semantic similarity to reference responses, offering a closer alignment with human judgment compared to traditional metrics like BLEU or ROUGE.

	Multi-Choice Question			Essay Question			
Model	**Ophtimus-Eval**	**MedMCQA(Ophthal)**	**PubmedQA(Ophthal)**	**RougeL**	**BLEU**	**METEOR**	**SemScore**
ChatGPT-4o	71.95%	81.95%	89.90%	0.193	0.082	0.341	0.761
Llama-3-8B-Instrct	48.60%	74.02%	63.97%	0.193	0.064	0.244	0.684
Llama-3.1-8B-Instrct	39.78%	57.96%	83.84%	0.177	0.054	0.215	0.641
Eye-Llama	32.56%	59.43%	66.11%	0.183	0.062	0.211	0.686
PMC-Llama-13B	48.28%	63.45%	72.48%	0.223	0.082	0.288	0.714
Ophtimus-1B-FT (Llama-3.2-1B+FT)							
Iteration 1	28.46%	37.00%	68.79%	0.228	0.080	0.232	0.723
Iteration 2	31.10%	38.47%	68.46%	0.225	0.078	0.223	0.720
Iteration 3	32.87%	39.28%	69.13%	0.221	0.078	0.220	0.716
Iteration 4	34.77%	38.44%	68.46%	0.219	0.076	0.217	0.711
Iteration 5	34.12%	38.14%	68.13%	0.219	0.076	0.220	0.713
Ophtimus-3B-FT (Llama-3.2-3B+FT)							
Iteration 1	42.90%	51.77%	69.46%	0.233	0.085	0.239	0.732
Iteration 2	44.99%	51.86%	68.79%	0.228	0.082	0.229	0.724
Iteration 3	46.56%	51.44%	68.46%	0.226	0.081	0.225	0.724
Iteration 4	46.01%	51.01%	69.80%	0.224	0.077	0.225	0.726
Iteration 5	45.26%	51.21%	69.13%	0.224	0.081	0.225	0.724
Ophtimus-8B-FT (Llama-3.1-8B+FT)							
Iteration 1	52.74%	61.64%	73.49%	0.238	0.087	0.243	0.741
Iteration 2	56.04%	60.27%	72.48%	0.233	0.088	0.233	0.733
Iteration 3	55.48%	60.05%	72.15%	0.233	0.089	0.236	0.737
Iteration 4	57.78%	59.49%	72.48%	0.226	0.083	0.230	0.733
Iteration 5	56.71%	58.81%	73.15%	0.220	0.078	0.226	0.730
Ophtimus-8B-PT-FT (Optimus-8B-PT+FT)							
Iteration 1	54.28%	60.80%	71.14%	0.237	0.089	0.243	0.740
Iteration 2	56.78%	61.21%	72.15%	0.227	0.081	0.235	0.734
Iteration 3	56.35%	60.16%	71.81%	0.227	0.084	0.233	0.734
Iteration 4	59.13%	58.82%	71.14%	0.222	0.079	0.224	0.735
Iteration 5	58.52%	59.46%	71.48%	0.222	0.080	0.230	0.730

Table 8: Detailed performance results by training epoch.