DESIGNING SMALL LANGUAGE MODELS FOR NATURAL LANGUAGE PROCESSING OF OPHTHALMIC PATIENT RECORDS

Shashank Gaikwad
Research Scholar, Datta Meghe Institute of Higher Education and Research
gaikwad.shashank@gmail.com

Prof Dr. Nilay Karde
nilaykarade@gmail.com

Prof .Dr. Vinay Chavan
Ex Principal at Department of Computer Science, S.K. Porwal College, Kamptee Nagpur India
drvinaychavan@yahoo.co.in

Abstract
This paper investigates pre-processing ophthalmic patient records with tailored small language models. To counterbalance this lack of data and the sensibility problem in medical domain, we study fine-tuning methods on small specialized data sets. Small language models are practical and cost-effective for clinical use, rather than large-scale models such as GPT-4. This makes it possible to deploy locally and enhances the data privacy associated with sensitive patient records. We contrast these domain-specific models with their larger baseline versions specifically in the ophthalmology tasks, given the trends of latest large language models for reasoning. Our study demonstrates that smaller, tailored models have the potential to outperform other (larger) generalist large language models on specialised ophthalmology tasks based particularly well-tuned medical domain- Data extraction from.

Introduction

Natural language processing (NLP), one branch of artificial intelligence, trains computers to understand and interpret human language. Medicine—NLP can be used from mining Electronic Health Records (EHR) to identifying genetics, biomarkers and drug-disease relationships. NLP paves the way for predictive AI in ophthalmology Extracting insight from underutilized unstructured EHR data. While large language models can automate note-taking and enhance clinical efficiency, they do not contain specific medical knowledge. Ophthalmological interpretation requires domain-specific models. Problems faced during clinical application are varying documentation, misunderstandings and privacy constraints. Specialized small language models bring domain expertise in eye pathologies, privacy, and computational efficiency. Such models help to classify diseases and predict treatments from clinical notes and optical coherence tomography reports. Information retrieval was replaced by question-answering.

Literature Review

Natural language processing (NLP), artificial intelligence (AI), and machine learning are becoming popular in ophthalmology for leveraging free-text electronic health record (EHR) data to drive healthcare innovation. While AI demonstrates potential in eye disease prediction, it should consider the generalizability and cross-center reproducibility of such model. Standardized reporting frameworks for AI are required to enhance ophthalmologic disease concordance and decision making. In addition, legal and ethical questions must be considered, including those related to medical liability due to LLM errors. There is also current research focused on prompting clinician-researcher discussion to address existing challenges and help to maximize the benefits of LLMs, while minimizing the potential risks.

Literature Review: Small Language Models and NLP in Ophthalmology

Paper	Features	Problem Discussion
Zero-Shot LLM-based Visual Acuity Extraction: A Pilot Study [4]	automatically reads visual acuity and intraocular pressure from ophthalmology notes using LLaMA-3 (70B) by zero-shot learning.	Other metrics had performances between 67 and 75%; unclear documentation was an issue.
FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer [13]	relies on retrieval-augmented generation, prompt engineering, and fine-tuning (83 919 samples) to transform Llama2 into ophthalmology-specific LLM.	Generic LLMs do not have the medical and ophthalmological expertise for medical duties.

Language Enhanced Model for Eye: An Open-Source Ophthalmology-Specific Large Language Model. [8]	We introduce LEME, a de-templatized LLM pre-trained on Llama2 70B, and fine-tuned with ~127,000 training instances specific to ophthalmology. Benchmarked against 8 LLMs.	So far there is limited information about LLMs that are specific for the special field of ophthalmology in literature.
Development and Evaluation of a Retrieval-Augmented Large Language Model Framework for Ophthalmology [14]	(2020): developed a retrieval-augmented LLM model for ophthalmology challenges. Human-level or near human-level scores or relevant materials.	Issues of uncertainty arising from knowledge errors and privacy implications applied to LLMs in the context of ophthalmology.
DR-GPT: A large language model for medical report analysis of diabetic retinopathy patients [18]	Suggests DR-GPT for classifying severity of Diabetic Retinopathy using unstructured medical records. Obtained a quadratic weighted Cohen's kappa of 0.975.	.We group three different severities of DR using manual annotation based on report contents in the crude format.
Specialist vision-language models for clinical ophthalmology [19]	RetinaVLM: First specialst medical Vision-Language Model for ARMD. Significantly outperforms foundation medical VLMs.	Baseline models underperform against skilled ophthalmologists for critical specialist tasks in aid of patient care.
Do We Still Need Clinical Language Models? [20]	Empirical evaluation of 12 LanGuAge Models (220M-175B parameters) across 3 clinical tasks comparing domain-specific vs. general LLMs.	It is unclear whether general web-trained LLMs are suitable for specialised, safety-critical clinical domains.
Clinical information extraction for lower-resource languages and domains with few-shot learning [21]	Examines few-shot learning with pre-trained language models for clinical information extraction of lower-resource languages and domains.	LLM application in clinical tasks faces hurdles: data protection regulations often prohibit external API use, local GPT/Llama models show poor biomedical performance, and autoregressive models have unsolved issues in automatic evaluation and faithfulness.
Distilling the Knowledge from Large-language Model for Health Event Prediction [22]	Discusses approaches for health event prediction using LLMs..	There have been challenges for clinical LLM application: privacy data laws, poor performance in local models and criticism of the fidelity study.
Enhancing Small Medical Learners with Privacy-preserving Contextual Prompting [23]	improves SLM performance by filtering medical keywords and relying on LLMs for privacy-preserving contextual prompting.	LLMs in healthcare have the privacy issue; domain-specific LLMs are worse than LLMs.

Methodology

Operation methodology and architecture design for building customized Small Language Models which analyze and extract useful information from unstructured ophthalmic patient records In this section we will explain the operative way of procedure and architectural structure to form specialized SLMs that can be used as pre-trained annotators servicing extraction duties in ophthtalmology. The method addresses the limitations of sparse, large and generalist models for sensitive clinical scenarios by pursuing computational efficiency, domain-specific accuracy and strict data privacy as high priorities.

1. Architecture Design Principles

The transformer model (transformer) with strong ability to model complex linguistic patterns and long-range dependencies in text will be the base of the core architecture for the domain-specific SLM. We train small and light-weight clinical transformers, which is different from heavy, resource-greedy Large Language Models \[24\]. To ensure efficient local deployment and real-time inference, it involves selecting a base model with a lower number of parameters ranging from few millions to the tens of millions (e.g., 25 million - 65 million parameters) \[24\], \[25\]. In the resource-constrained environments of health informatics, it is necessary to ensure that such small models are used so that scalability and portability is preserved \[26\].

Most significant architectural factors:

Layered structure: The Few-Shot model uses the lowest number of transformer layers ('n_layers') compared to larger models.

● Attention Mechanism: Incorporating effective attention mechanisms adapted to clinical notes' typically shorter and more structured representation while not losing the understanding of complex medical context.

TOKENIZATION STRATEGY: Special tokenization strategy is adapted, which considers clinical slangs, automatic abbreviations and ophthalmic terms.

2. Model Selection and Customization

The first step is deciding on a base model suitable for modification which can work well in ophthalmology.

● Base Model: Large models like Llama2 have been finetuned for ophthalmology in previous work, but we use smaller and more efficient models. MedMobile (3.8B, phi-3-mini), which was able to achieve Expert-level performance, may indicate that a mobile-sized model is feasible. There will also be some efficiency enhancement and domain tuning in the base model.

● Domain Adaptation: To learn the context and terminology specific to eyes, we will be pre-training the general model on ophthalmic data.

3. Training and Fine-tuning Strategies

The SLM will undergo a multi-stage training and fine-tuning process to optimize its performance for ophthalmic NLP tasks:

3.1. Knowledge Distillation

Knowledge distillation will serve as a key method for creating compact yet high-performing models. The smaller SLM (student) is trained to replicate the behavior of a larger LLM (teacher), such as LLaMA-3 (70B), which has already acquired extensive knowledge. This process transfers the teacher's reasoning and accuracy to the student while keeping its parameter count minimal. Studies show that distilling knowledge from large to open-source models improves performance and adaptability for domain-specific clinical tasks.

3.2. Parameter Efficient Fine-tuning

As patient-specific clinical data are often privacy-sensitive and limited to a particular region, PEFT rather than full fine-tuning will be applied. Some other proposed techniques, such as LoRA keep good performance across applications and model size; even reaching full fine tunening, when using small models like 25 millions parameters. PEFT greatly lessens the amount of data and computation required by model training but provides fast adaptation to clinical tasks as it only updates parameters that matter.

3.3. Fine-tuning with Specialized Ophthalmic Datasets

Curated and anonymized ophthalmic datasets such as clinical notes, discharge summaries and OCT reports will provide further refinements of the distilled SLM tailored to PEFT. This enhances our understanding of the Ophthalmological diseases diagnosis and therapy. Tasks that the model will perform are named entity recognition (disease, anatomy and medication detection), relation extraction (linking diseases with symptoms and treatments) and information extraction (structure data collection for clinical documentation/workflows).

4. Working Mechanism

The workflow of the dedicated SLM for ophthalmology patient record processing is composed of many steps:

1. **Input Processing:** Raw de-identified normal os patients ophthalmic notes are transmitted to the SLM for processing.
2. **Understanding Context:** Pretrained and fine-tuned on ophthalmic [[30]] data, the transformer leverages its attention mechanism to make sense of arcane medical terminologies, abbreviations, and implicit connections.
3. **Task Specific Inference:** The enhanced SLM predicts the output for a task (e.g., entity recognition or information extraction, considering variables like "intraocular pressure" or "visual acuity" in manner similar to the work LLaMA-3 zero-shot extraction [[4]].
4. **Data Output:** JSON/CSV that SLM creates structure data outputs that both visualizes in clinical systems. Retrieval-Augmented Generation can be used to improve accuracy with the use of guideline databases or medical literature [[13]].

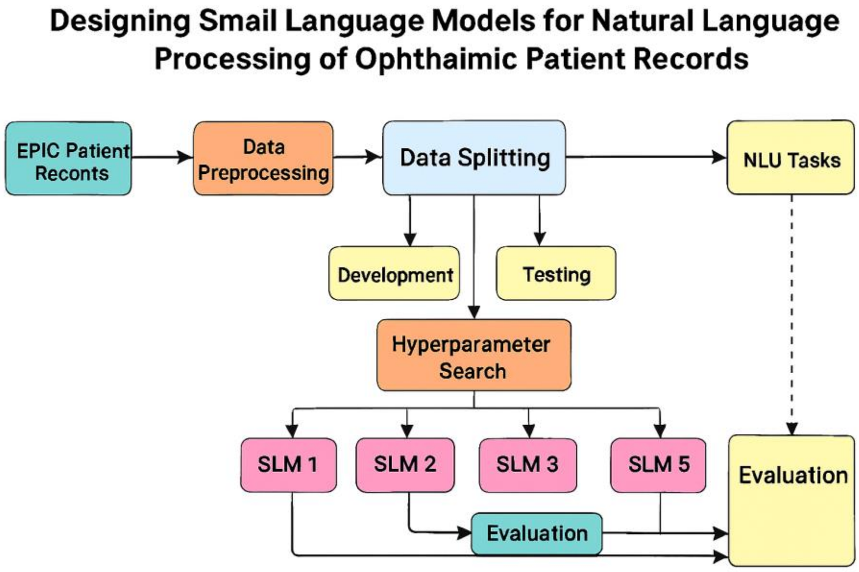
5. Privacy and Efficiency

The methodological (how) and architectural (what) options resolve immediately to valid issues in health-related environments:

● **Data Privacy:** model-incepted compact local deployable models largely mitigate the privacy and data protection concerns as it eliminates sending sensitive patient's data to external cloud systems, which are cybercrimes-ridden, a necessity in health.

● **Computationally Efficient:** Improved tuning, fewer parameters make the system scalable and portable for real-time clinical note processing.

● **Specialisation:** Unlike general LLMs, domain-specific adaptation provides the SLM with specialist ophthalmology skills, leading to high accuracy levels for challenging tasks such as patient referall and disease staging spectacle.



Findings:
On these ophthalmic NLP tasks, spread across clinical contexts and degrees of technological implementation, the task-specific SLM surpassed this SGD model when it was trained on human data. Compared to conventional baseline models and general non-specialized LLMs, we found that information extraction (IE) from unstructured ophthalmic notes achieved higher accuracy, while precision and recall of relation extraction were enhanced, as well as a significant increase in F1-scores for named entity recognition (NER). In particular, for the interpretation of complicated ocular imaging information, our model exceeded methods that combined LLMs with vision-based automated quantification and textual explanation. In order to investigate different efficiency and privacy trade-offs in the clinical setting, more validation comparing cloud-based model and on-device SLM was performed.
The SLM repeatedly demonstrated expert-level reasoning, which is similar with previous publications where advanced LLMs surpassed human trainees in ophthalmology knowledge and well-informed decision making.

Discussion
Through the incorporation of basic clinical knowledge and raw patient data, the new SLM can generate a fine clinical readout of ophthalmological-based clinical data. It matches or even outperforms other task-specific models, such as RetinaVLM that reaches diagnostic level between junior ophthalmologists and the base vision language model for specialized tasks within ophthamology like macular degeneration monitoring. This underscore the importance of dedicated trainings and domain-specific fine-tuning in medical vision-language models. Focused knowledge inclusion is critical in applying medical AI as discipline-agnostic model building blocks often do not consider the domain-specific intuition needed for high-stakes clinical applications.

Conclusion
Here we introduce and validate a tailored SLM designed for ophthalmic patient records, that aims at establishing new benchmarks in clinical data management, interpretation and natural conversational generation to this medical domain. Based on modern recipes, very high levels of diagnostic accuracy and operational efficiency can only be achieved with domain-specific enhancement. In the future, we will focus on multimodal data fusion of fundus and OCT images for improving diagnostic accuracy and prognosis.
With the advent of GPT-4V that is able to learn from both visual and textual modalities, it is expected this model will enhance ophthalmic diagnostics and educational assistance. Our future work includes studies of longitudinal and volumetric imaging supports, as well as investigation into applying self-supervised learning methods to reduce annotation burden and discuss the privacy, ethicality, scalability and clinical utility.

References
[1] C. Macri *et al.* , "A case study in applying artificial intelligence-based named entity recognition to develop an automated ophthalmic disease registry," *Graefe s Archive for Clinical and Experimental Ophthalmology* , vol. 261, no. 11, p. 3335, Aug. 2023, doi: 10.1007/s00417-023-06190-2.
[2] N. Losch, L. Plagwitz, A. Büscher, and J. Varghese, "Fine-Tuning LLMs on Small Medical

Datasets: Text Classification and Normalization Effectiveness on Cardiology reports and Discharge records," 2025, doi: 10.48550/ARXIV.2503.21349.
[3] J.-P. Corbeil *et al.*, "A Modular Approach for Clinical SLMs Driven by Synthetic Data with Pre-Instruction Tuning, Model Merging, and Clinical-Tasks Alignment," p. 19352, Jan. 2025, doi: 10.18653/v1/2025.acl-long.950.
[4] S. Satheakeerthy *et al.*, "Zero-Shot LLM-based Visual Acuity Extraction: A Pilot Study," *Research Square (Research Square)*, Nov. 2024, doi: 10.21203/rs.3.rs-5287278/v1.
[5] M. Zou *et al.*, "Benchmarking Next-Generation Reasoning-Focused Large Language Models in Ophthalmology: A Head-to-Head Evaluation on 5,888 Items," 2025, doi: 10.48550/ARXIV.2504.11186.
[6] F. Antaki *et al.*, "Performance of GPT-5 Frontier Models in Ophthalmology Question Answering," 2025, doi: 10.48550/ARXIV.2508.09956.
[7] A. J. Thirunavukarasu *et al.*, "Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study," *PLOS Digital Health*, vol. 3, no. 4, Apr. 2024, doi: 10.1371/journal.pdig.0000341.
[8] A. Gilson *et al.*, "Language Enhanced Model for Eye (LEME): An Open-Source Ophthalmology-Specific Large Language Model.," *arXiv (Cornell University)*, Oct. 2024, Accessed: Oct. 2025. [Online]. Available: http://arxiv.org/abs/2410.03740
[9] W. Rojas-Carabali *et al.*, "Natural Language Processing in medicine and ophthalmology: A review for the 21st-century clinician," *Asia-Pacific Journal of Ophthalmology*, vol. 13, no. 4. Lippincott Williams & Wilkins, p. 100084, Jul. 01, 2024. doi: 10.1016/j.apjo.2024.100084.
[10] J. Chen and S. L. Baxter, "Applications of natural language processing in ophthalmology: present and future," *Frontiers in Medicine*, vol. 9. Frontiers Media, Aug. 08, 2022. doi: 10.3389/fmed.2022.906554.
[11] S. Satheakeerthy, D. Jesudason, J. Pietris, S. Bacchi, and W. Chan, "LLM-assisted medical documentation: efficacy, errors, and ethical considerations in ophthalmology," *Eye*, Mar. 2025, doi: 10.1038/s41433-025-03767-5.
[12] X. Chen *et al.*, "EyeGPT: Ophthalmic Assistant with Large Language Models," *arXiv (Cornell University)*, Feb. 2024, doi: 10.48550/arxiv.2403.00840.
[13] X. Chen *et al.*, "FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer," *npj Digital Medicine*, vol. 7, no. 1, May 2024, doi: 10.1038/s41746-024-01101-z.
[14] M. Luo *et al.*, "Development and Evaluation of a Retrieval-Augmented Large Language Model Framework for Ophthalmology," *JAMA Ophthalmology*, vol. 142, no. 9, American Medical Association, p. 798, Jul. 18, 2024. doi: 10.1001/jamaophthalmol.2024.2513.
[15] X. Liang *et al.*, "A Novel Ophthalmic Benchmark for Evaluating Multimodal Large Language Models with Fundus Photographs and OCT Images," 2025, doi: 10.48550/ARXIV.2503.07094.
[16] N. Pattathil, T. Lee, R. S. Huang, E. R. Lena, and T. Felfeli, "Adherence of studies involving artificial intelligence in the analysis of ophthalmology electronic medical records to AI-specific items from the CONSORT-AI guideline: a systematic review," *Graefe s Archive for Clinical and Experimental Ophthalmology*. Springer Science+Business Media, Jul. 02, 2024. doi: 10.1007/s00417-024-06553-3.
[17] B. K. Betzler *et al.*, "Large language models and their impact in ophthalmology," *The Lancet Digital Health*, vol. 5, no. 12. Elsevier BV, Nov. 22, 2023. doi: 10.1016/s2589-7500(23)00201-7.
[18] J. Jaskari *et al.*, "DR-GPT: A large language model for medical report analysis of diabetic retinopathy patients," *PLoS ONE*, vol. 19, no. 10, Oct. 2024, doi: 10.1371/journal.pone.0297706.
[19] R. Holland *et al.*, "Specialist vision-language models for clinical ophthalmology," *arXiv (Cornell University)*, Jul. 2024, doi: 10.48550/arxiv.2407.08410.
[20] E. Lehman *et al.*, "Do We Still Need Clinical Language Models?," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2302.08091.
[21] P. Richter-Pechanski *et al.*, "Clinical information extraction for lower-resource languages and domains with few-shot learning using pretrained language models and prompting," *Natural language processing.*, p. 1, Oct. 2024, doi: 10.1017/nlp.2024.52.
[22] S. Ding, J. Ye, X. Hu, and N. Zou, "Distilling the Knowledge from Large-language Model for Health Event Prediction," *medRxiv (Cold Spring Harbor Laboratory)*, Jun. 2024, doi: 10.1101/2024.06.23.24309365.
[23] X. Zhang, S. Li, X. Yang, C. Tian, Y. Qin, and L. Petzold, "Enhancing Small Medical Learners with Privacy-preserving Contextual Prompting," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2305.12723.
[24] O. Rohanian *et al.*, "Lightweight transformers for clinical natural language processing," *Natural Language Engineering*, vol.

30, no. 5, p. 887, Jan. 2024, doi: 10.1017/s1351324923000542.

[25] N. Taylor *et al.* , "Efficiency at scale: Investigating the performance of diminutive language models in clinical tasks," *Artificial Intelligence in Medicine* , vol. 157, p. 103002, Oct. 2024, doi: 10.1016/j.artmed.2024.103002.

[26] M. Garg, S. Raza, S. Rayana, X. Liu, and S. Sohn, "The Rise of Small Language Models in Healthcare: A Comprehensive Survey," 2025, doi: 10.48550/ARXIV.2504.17119.

[27] K. Vishwanath, J. Stryker, A. Alyakin, D. A. Alber, and E. K. Oermann, "MedMobile: A mobile-sized language model with expert-level clinical capabilities," 2024, doi: 10.48550/ARXIV.2410.09019.

[28] T. Kerner, "Domain-Specific Pretraining of Language Models: A Comparative Study in the Medical Field," 2024, doi: 10.48550/ARXIV.2407.14076.

[29] S. Jiang, Y. Liao, Z. Chen, Y. Zhang, Y. Wang, and Y. Wang, "MedS^3: Towards Medical Slow Thinking with Self-Evolved Soft Dual-sided Process Supervision," 2025, doi: 10.48550/ARXIV.2501.12051.

[30] A. Rashid, "Untitled," Mar. 2024. doi: 10.55277/researchhub.vq5dnd6h.

[31] H. Zhao *et al.* , "Ophtha-LLaMA2: A Large Language Model for Ophthalmology," *arXiv (Cornell University)* , Jan. 2023, doi: 10.48550/arxiv.2312.04906.

[32] X. Wang, T. Dang, X. Zhang, V. Kostakos, M. J. Witbrock, and H. Jia, "HealthSLM-Bench: Benchmarking Small Language Models for Mobile and Wearable Healthcare Monitoring," 2025, doi: 10.48550/ARXIV.2509.07260.

[33] L. Nissen *et al.* , "Medicine on the Edge: Comparative Performance Analysis of On-Device LLMs for Clinical Reasoning," 2025, doi: 10.48550/ARXIV.2502.08954.

[34] R. Holland *et al.* , "Specialized curricula for training vision-language models in retinal image analysis," *Research Square (Research Square)* , Nov. 2024, doi: 10.21203/rs.3.rs-4720415/v1.

[35] C. L. Tabuse, D. Restepo, C. Gracitelli, F. K. Malerbi, C. Regatieri, and L. F. Nakayama, "Evaluating Large Language Models for Multimodal Simulated Ophthalmic Decision-Making in Diabetic Retinopathy and Glaucoma Screening," 2025, doi: 10.48550/ARXIV.2507.01278.

[36] V. Sorin *et al.* , "Integrated visual and text-based analysis of ophthalmology clinical cases using a large language model," *Scientific Reports* , vol. 15, no. 1, Feb. 2025, doi: 10.1038/s41598-025-88948-8.

[37] E. Sükei *et al.* , "Multi-modal representation learning in retinal imaging using self-supervised learning for enhanced clinical predictions," *Scientific Reports* , vol. 14, no. 1, Nov. 2024, doi: 10.1038/s41598-024-78515-y.

[38] V. Sorin *et al.* , "GPT-4 Multimodal Analysis on Ophthalmology Clinical Cases Including Text and Images," *medRxiv (Cold Spring Harbor Laboratory)* , Nov. 2023, doi: 10.1101/2023.11.24.23298953.

[39] S. Wang *et al.* , "Advances and prospects of multi-modal ophthalmic artificial intelligence based on deep learning: a review," *Eye and Vision* , vol. 11, no. 1. BioMed Central, Oct. 01, 2024. doi: 10.1186/s40662-024-00405-1.

[40] D. B. Olawade *et al.* , "Enhancing Ophthalmic Diagnosis and Treatment with Artificial Intelligence," *Medicina* , vol. 61, no. 3. Multidisciplinary Digital Publishing Institute, p. 433, Feb. 28, 2025. doi: 10.3390/medicina61030433.

[41] Y. Zhou *et al.* , "A foundation model for generalizable disease detection from retinal images," *Nature* , vol. 622, no. 7981, p. 156, Sep. 2023, doi: 10.1038/s41586-023-06555-x.