LEME: Open Large Language Models for Ophthalmology with Clinical Validation

Hyunjae Kim[1], Xuguang Ai[1], Sahana Srinivasan[2], LEME Research Group, Lucila Ohno-Machado[1], Hua Xu[1], Ron A. Adelman[1], Luciano V. Del Priore[1], Yih-Chung Tham[2,*], Qingyu Chen[1,*]

[1]*Yale School of Medicine, CT, USA*
[2]*Yong Loo Lin School of Medicine, National University of Singapore, Singapore*
[*]Correspondence: thamyc@nus.edu.sg, qingyu.chen@yale.edu

Yale SCHOOL OF MEDICINE

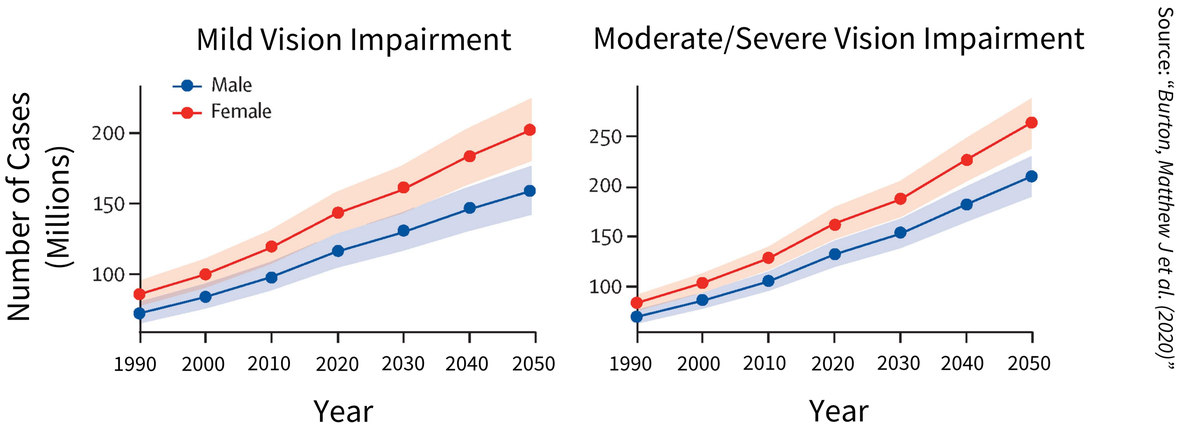
NUS National University of Singapore — Yong Loo Lin School of Medicine

Paper

GitHub

1. Problem

The Global Burden of Eye Diseases is Both Substantial and Increasing.



Source: "*Burton, Matthew J et al. (2020)*"

Few Models Have Been Tailored for Ophthalmology, and None Have Yet Been Rigorously Validated.

Model (Release)	Availability	Base LLM	Largest	Preference Tuning	Clinical Validation
Ophtha-LLaMA2 (Dec 2023)	✗	LLaMA-2	7B	✗	✗
Tan et al. (Feb 2024)	✗	LLaMA-2	13B	✗	✗
EyeGPT (Feb 2024)	✗	LLaMA-2	7B	✗	✗
Eye-LLaMA (Apr 2024)	License-restricted	LLaMA-2	7B	✗	✗
Ophtimus-LLM (Mar 2025)	✗	LLaMA-3 Series	8B	✗	✗
LEME	✓	LLaMA-3 Series	70B	✓	✓

Limitations (Threefold)

- **Transparency**: Only *Eye-LLaMA* is public, but trained on copyrighted data.
- **Development**: Models use small or outdated architectures with minimal reasoning-focused tuning.
- **Evaluation**: Mostly limited to knowledge-based QA without real-world validation.

2. Aims

1 Model Development

Introducing **LEME**: Reasoning-aligned, open-weight LLMs tailored for ophthalmology.

2 Clinical Validation

Clinically grounded evaluation of LLMs in ophthalmology.

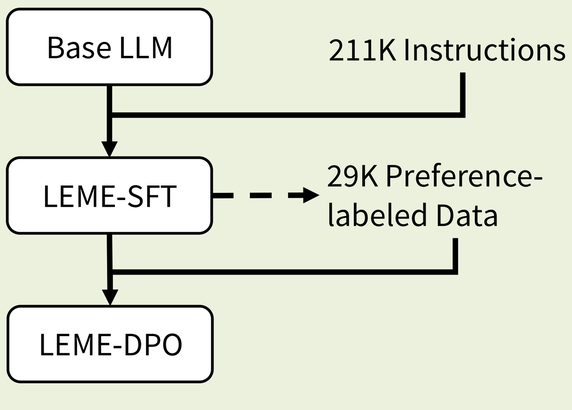
3 Open Research

Models, datasets, and codes openly released for the community.

3. Model Training

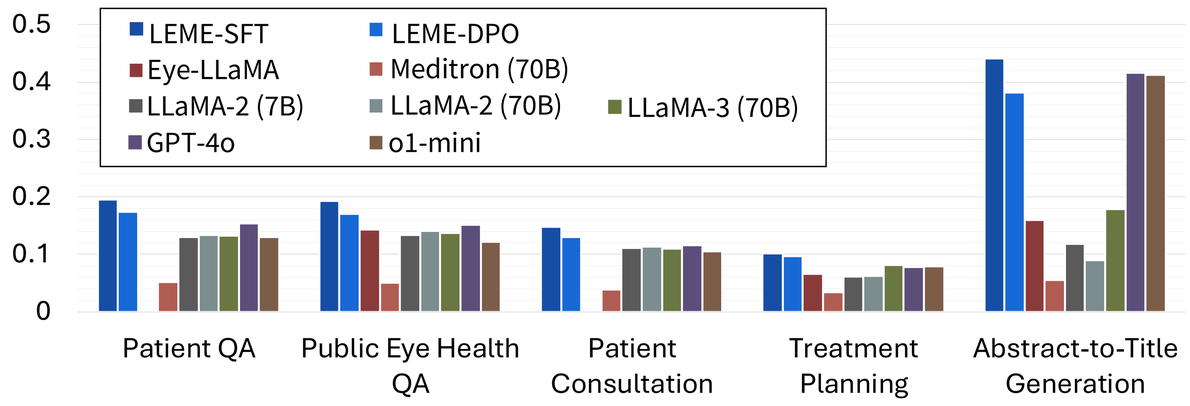
Phase I: Supervised Fine-tuning (SFT)

Phase II: Direct Preference Optimization (DPO)



4. Benchmark Evaluation

LEME Outperformed All Baselines ($p < 0.004$).



5. Clinical Validation

(i) In Answering Patient Queries, LEME Achieved the Highest Clinician Ratings Across Three Metrics.

Model	Factuality	Specificity	Complete.	Safety
Expert	**4.73**	4.44	4.33	4.67
LLaMA-3	4.12	4.01	4.59	4.62
LEME-DPO	4.67	**4.77**	**4.79**	**4.88**

(ii) In Visual Acuity Extraction, LEME Scored the Highest F1.

Input: Progress Note

…
Right Left
Dist cc 20/30 -2 20/40
Dist ph cc 20/25 +2 20/25 +2
Near cc J1+ J1+

Output: Distance VA Values

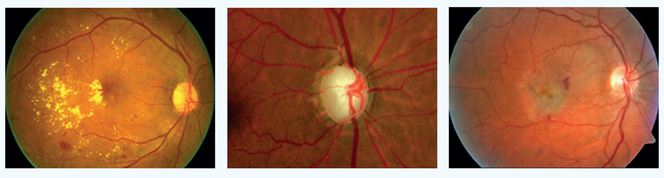
OD, corrected: 20/30– 2
OD, corrected, pinhole: 20/40
OS, corrected: 20/25 +2
OS, corrected, pinhole: 20/25 +2
…

- F1 Scores: Eye-LLaMA (3.0%); LLaMA-3-70B (47.9%); **LEME (62.0%)**

(iii) In Assessment and Treatment Planning, LEME Approaches Attending-level Clinicians.

Input: Progress Note + Imaging Texts

Patients with the following conditions:



Diabetic Retinopathy | Glaucoma | AMD

Output:

Assessment: … Examination findings include 2+ nuclear sclerosis, posterior vitreous detachment, inferior disc sloping, … The patient also has dermatochalasis, pinguecula, …
Treatment Plan: 1. Continue current glaucoma medications: … 2. Refill prescriptions for Brimonidine and Latanoprost as requested. …

Model	Factuality	Specificity	Complete.	Safety
Expert	**4.85**	**4.83**	**4.60**	**4.71**
LEME-DPO	4.36	4.55	4.42	4.36

6. Conclusion

- LEME shows promising potential to ease the burden in ophthalmology.
- In pilot tests on assessment and treatment planning, LEME exhibited emerging capability in challenging clinical routine tasks.
- Open release of our resources will accelerate clinical translation.