RetiRAG: A Retrieval-Augmented Generation Framework for Specialized Ophthalmology Applications

Ting Xu[1,2‡], Siyani Chen[3‡], Meng Wang[1,2], Wenbin Liao[4], Jie Zhang[5], Tian Lin[3],

Huazhu Fu[6], Dianbo Liu[1,2,✉], Haoyu Chen[3,✉], Ching-Yu Cheng[1,2,7,✉]

[1] *Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119228, Singapore.*

[2] *Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119228, Singapore.*

[3] *Joint Shantou International Eye Center, Shantou University and the Chinese University of Hong Kong, 515041, Shantou, Guangdong, China.*

[4] *School of Computer Science and Technology, University of Chinese Academy of Sciences, 100190, Beijing, China.*

[5] *Center for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore.*

[6] *Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore.*

[7] *Singapore Eye Research Institute, Singapore National Eye Centre, Republic of Singapore.*

‡ T. Xu, and S. Chen are the co-first authors.

✉C.Y. Cheng, H. Chen, and D. Liu are the co-corresponding authors and contributed equally.

1. Purpose

Large language models (LLMs) have seen rapid advancements in recent years, but their performance in specialized fields like ophthalmology remains constrained [1]. A critical challenge lies in integrating domain-specific knowledge into these models to enhance their expertise. To address this, we developed an ophthalmology knowledge base and implemented a retrieval-augmented generation (RAG) approach to create RetiRAG, a specialized ophthalmology-focused language model. Our framework enables local deployment to safeguard patient privacy and eliminate the risk of data leakage. By leveraging the knowledge base, RetiRAG empowers LLMs with precise and context-aware understanding, enhancing their applications in ophthalmology.

2. Methods

We built a comprehensive knowledge base comprising hundreds of ophthalmology textbooks, research papers, and clinical guidelines to develop RetiRAG, a specialized language model tailored for ophthalmology. To optimize information retrieval, the text was divided into smaller chunks of five sentences each. A Retrieval-Augmented Generation (RAG) framework [2] was employed, utilizing the e5-base-v2 model as the retriever and Llama-3-8B-Instruct as the generator. The retriever encoded both user queries and knowledge chunks into 768-dimensional vectors, facilitating efficient retrieval based on cosine similarity. The top five most relevant chunks were then retrieved and provided as input to the generator, which produced a response using a predefined prompt template. The pipeline of our framework is illustrated in Figure 1. We evaluated the model on two tasks: answering questions related to common ophthalmic conditions and generating keywords for rare disease.

3. Results

For common ophthalmic conditions, we evaluated the model's performance using 19 clinical questions, with examples shown in the left section of Figure 2. RetiRAG achieved an accuracy of 63.2%, compared to 52.6% for the non-RAG model. Additionally, three junior ophthalmology residents completed the same questions, with accuracies of 63.2%, 42.1%, and 47.4%, respectively. These results showed that the RetiRAG model outperformed the non-RAG one and performed on par with junior-level ophthalmology residents, showcasing its potential as a valuable tool for clinical decision-making in common ophthalmic scenarios. For rare diseases, we evaluated the model's ability to generate keywords for 10 ophthalmology-specific rare conditions (the right section of Figure 2). An ophthalmologist assessed the quality of the outputs, and we found that RetiRAG produced keywords more closely aligned with expert descriptions. In contrast, the non-RAG model generated partially relevant keywords—approximately half were directly related to the diseases, with the remainder being generic or less meaningful. The comparison of generated keywords in Figure 2 highlights RetiRAG's ability to produce more focused, contextually relevant, and clinically meaningful outputs for rare diseases.

4.Conclusion

This study developed RetiRAG, a retrieval-augmented language model for ophthalmology, demonstrating improved accuracy on clinical questions and more relevant keyword generation for rare diseases compared to a non-RAG model. Operating efficiently on a single GPU with an ophthalmology-specific knowledge base, RetiRAG highlights the potential for compact, scalable LLMs to achieve high performance in specialized medical domains while ensuring resource efficiency and safety.

References

[1] Betzler, B. K., Chen, H., Cheng, C. Y., Lee, C. S., Ning, G., Song, S. J., ... & Wong, T. Y. (2023). Large language models and their impact in ophthalmology. The Lancet Digital Health, 5(12), e917-e924.

[2] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2.

RetiRAG: A Retrieval-Augmented Generation Framework for Specialized Ophthalmology Applications

Ting Xu[1,2‡], Siyani Chen[3‡], Meng Wang[1,2], Wenbin Liao[4], Jie Zhang[5], Tian Lin[3],

Huazhu Fu[6], Dianbo Liu[1,2,✉], Haoyu Chen[3,✉], Ching-Yu Cheng[1,2,7,✉]

[1] *Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119228, Singapore.*

[2] *Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119228, Singapore.*

[3] *Joint Shantou International Eye Center, Shantou University and the Chinese University of Hong Kong, 515041, Shantou, Guangdong, China.*

[4] *School of Computer Science and Technology, University of Chinese Academy of Sciences, 100190, Beijing, China.*

[5] *Center for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore.*

[6] *Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore.*

[7] *Singapore Eye Research Institute, Singapore National Eye Centre, Republic of Singapore.*

‡ T. Xu, and S. Chen are the co-first authors.

✉C.Y. Cheng, H. Chen, and D. Liu are the co-corresponding authors and contributed equally.

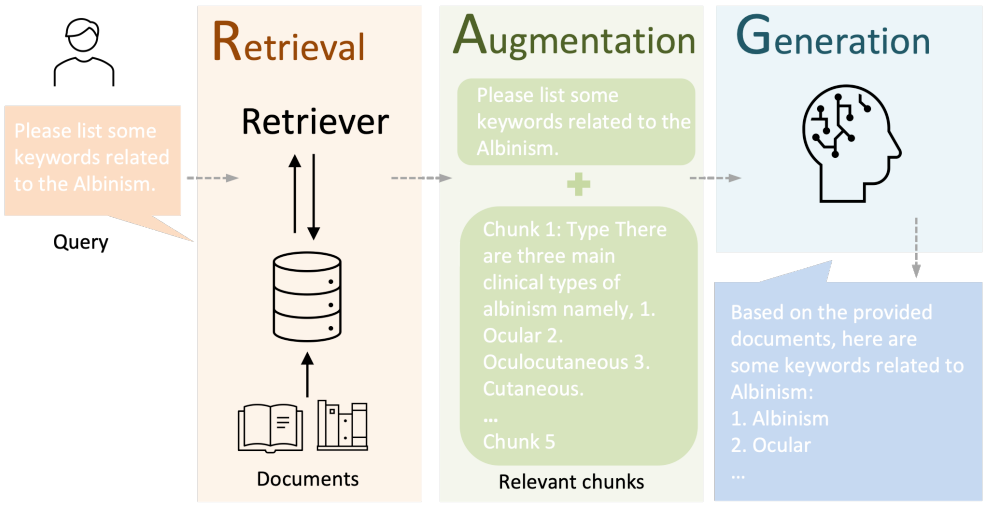

Figure 1: Overview Framework of RetiRAG: A Retrieval-Augmented Generation Pipeline.

Prompt: Please select a correct answer from the following options:		Prompt: Please list some keywords related to the Albinism.	
Question 1	Question 2	RetiRAG	Non-RAG
A 60-year-old male presents with a decrease in the vision and distortion of images in the right eye for 3 months. He gave a history of cataract surgery a year ago. Examination reveals a 20/60 vision on Snellen's chat and a vertical metamorphopsia on the Amsler grid. The optical coherence tomography scan shows a hyperreflective layer above the internal limiting membrane in the macula with an increase in the thickness of macula with fine cystic changes. Which of the following is the most appropriate diagnosis? A. Recurrence of cystoid macular edema B. Macular edema secondary to macular traction C. Persistent epithelial detachment D. Clinically significant macular edema *Standard Answer: B RetiRAG: B Non-RAG: D* *Doctor1: B Doctor2: D Doctor3: B*	A 65-year-old male with a 30-year history of uncontrolled hypertension presents complaining of blurry vision for two months that has become more noticeable during the past week. Confrontational visual fields show inferonasal visual field defects confirmed by visual field testing. What is the most likely pathophysiologic mechanism responsible for this patient's blurry vision? A. Macular edema B. Peripheral retinal edema C. Arteriovenous knicking D. Cotton wool spots *Standard Answer: A RetiRAG: A Non-RAG: B* *Doctor1: B Doctor2: D Doctor3: C*	Based on the provided documents, here are some keywords related to Albinism: 1. Albinism 2. Ocular 3. Oculocutaneous 4. Cutaneous 5. Melanin 6. Hypopigmentation 7. Nystagmus 8. Ophthalmology 9. Genetic 10. Congenital 11. Autosomal recessive 12. X-linked recessive 13. Pigmentation 14. Iris 15. Metabolic disease	Here are some keywords related to Albinism: 1. Albinism 2. Congenital disorder 3. Lack of melanin 4. Vision impairment 5. Skin condition 6. Hair condition 7. Eye problems 8. Nystagmus 9. Photophobia 10. Hypopigmentation 11. Oculocutaneous albinism 12. Mucocutaneous albinism 13. Osteogenesis imperfecta 14. Genetic disorder 15. Rare condition 16. Inheritance pattern 17. Autosomal recessive 18. Albinism and vision 19. Albinism and skin 20. Albinism and hair

Figure 2: Comparison of RetiRAG and Non-RAG Model Performance on Common Ophthalmic Questions (Left) and Rare Disease Keyword Generation (Right).