Advancing Question-Answering in Ophthalmology with Retrieval-Augmented Generation (RAG): Benchmarking Open-source and Proprietary Large Language Models

Quang Nguyen*[1,2,3], Duy-Anh Nguyen*[4], Khang Dang[7], Siyin Liu[1,3], Khai Nguyen[4], Sophia Y. Wang[5], William Woof [1], Peter Thomas[3], Praveen J. Patel[1,3,8], Konstantinos Balaskas[1,3], Johan H. Thygesen[2], Honghan Wu[6], Nikolas Pontikos[1,3]

[1]UCL Institute of Ophthalmology, 11-43 Bath Street, London EC1V 9EL, UK
[2]UCL Institute of Health Informatics, 222 Euston Rd, London NW1 2DA, UK
[3]Moorfields Eye Hospital NHS Foundation Trust, 162 City Road, London EC1V 2PD, UK
[4]CT MedTech Ltd, Ho Chi Minh city, Vietnam
[5]Stanford Byers Eye Institute, 2452 Watson Ct, Palo Alto CA 94303, USA
[6]School of Health and Wellbeing, University of Glasgow, 90 Byres Road, Glasgow G12 8TB, UK
[7]Department of Informatics, New Jersey Institute of Technology, 218 Central Ave, Newark NJ 07102, USA
[8]NIHR Biomedical Research Centre for Ophthalmology

*equal contribution
Corresponding: {quang.nguyen.21, n.pontikos}@ucl.ac.uk

Abstract

Purpose To evaluate the application of Retrieval-Augmented Generation (RAG), a technique that combines information retrieval with text generation, to benchmark the performance of open-source and proprietary generative large language models (LLMs) in medical question-answering tasks within the ophthalmology domain.

Methods Our dataset comprised 260 multiple-choice questions sourced from two question-answer banks designed to assess ophthalmic knowledge: the American Academy of Ophthalmology's Basic and Clinical Science Course (BCSC) Self-Assessment program and OphthoQuestions. Our RAG pipeline involved initial retrieval of documents in the BCSC companion textbook using ChromaDB, followed by reranking with Cohere to refine the context provided to the LLMs. We benchmarked four models, including GPT-4 and three open-source models (Llama-3-70B, Gemma-2-27B, and Mixtral-8x7B, all under 4-bit quantization), under three settings: zero-shot, zero-shot with Chain-of-Thought and RAG. Model performance was evaluated using accuracy on the two datasets. Quantization was applied to improve the efficiency of the open-source models. Effects of quantization level was also measured.

Results Using RAG, GPT-4-turbo' s accuracy increased from 80.38% to 91.92% on BCSC and from 77.69% to 88.65 % on OphthoQuestions. Importantly, the RAG pipeline greatly enhanced overall performance of Llama-3 from 57.50% to 81.35% (23.85% increase), Gemma-2 62.12% to 79.23% (17.11% increase), and Mixtral-8x7B 52.89% to 75% (22.11% increase). Zero-shot-CoT had overall no significant improvement on the models' performance. Quantization using 4 bit was shown to be as effective as using 8 bits while requiring half the resources.

Conclusion Our work demonstrates that integrating RAG significantly enhances LLM accuracy, especially for privacy-preserving smaller open-source LLMs that can be run in sensitive and resource-constrained environments such within hospitals, offering a viable alternative to cloud-based LLMs like GPT-4-turbo.

Introduction

Generative Large Language Models (LLMs) have been shown to demonstrate a remarkable amount of clinical knowledge, both in general medicine [1] and within specialist domains such as ophthalmology [2], prompting them to be considered as effective tools for assisting healthcare activities. Antaki et al demonstrated that GPT4, the latest iteration of OpenAI's Generative Pretrained Transformer model, performs comparably to humans in answering ophthalmic exam questions [2]. Nonetheless, GPT-4 performance was suboptimal, with 76.5% accuracy on the American Academy of Ophthalmology's Basic and Clinical Science Course (BCSC) Self-Assessment program and 70% on OphthoQuestions question sets, indicating substantial room for improvement to attain the levels of accuracy needed in healthcare settings. In addition, as noted by Antaki et al [2], there is an inherent tendency of LLMs to produce erroneous, nonfactual information, commonly referred to as hallucination [3,4], which is not acceptable in critical domains such as healthcare.

Several techniques have been proposed to improve LLM performance and reduce hallucinations including, enhancing prompting via prompt engineering such as "chain of thought" [5], fine-tuning [6,7] or reinforcement with human feedback (RLHF) [8]. While prompt engineering is a simple and cost-effective way to improve LLM, it cannot enable the model to "know" or correctly infer information beyond its training data. Moreover, prompting can inadvertently amplify bias within a model [9]. For example, Tamkin et al. [9] showed that when a prompt explicitly mentions a candidate's age, race, and gender, the model (Claude 2.0) tended to favour younger individuals over older in hiring decision scenario. On the other hand, fine-tuning and RLHF, although potentially more effective, are extremely computationally expensive [6,10], making them impractical in resource-constrained settings such as hospital settings.

We proposed an alternative, potentially promising approach: the Retrieval-Augmented Generation (RAG) pipeline. RAG, introduced by Lewis et al in 2020, is a hybrid method that combines the strengths of information retrieval with generative models to enhance the quality and relevance of generated text, and was commonly used prior to the advent of LLMs [11]. RAG integrates a privacy-protecting retrieval component that identifies and extracts relevant information from a large user-defined corpus, which can be stored locally, with a generative model that synthesizes this information into coherent and contextually appropriate responses. Using this approach, RAG can effectively leverage additional

knowledge to improve content accuracy and informativeness. This method may be particularly valuable where the generative model's internal knowledge is insufficient or outdated, ensuring the output is both up-to-date and contextually enriched.

Our study presents three key contributions. First, we explore an effective method to enhance the capabilities of LLMs in understanding knowledge in ophthalmology via answering medical exam question without relying on resource-intensive fine-tuning techniques by leveraging external knowledge sources through a RAG pipeline. Secondly, we evaluate the potential of a small, open-source LLM in accurately capturing ophthalmic knowledge when RAG augmented. Finally, we quantify the improvement achieved by implementing the RAG pipeline, offering insights into its effectiveness in enhancing LLM performance within healthcare contexts.

Methods

Data Acquisition and Preparation

We sourced our dataset from the American Academy of Ophthalmology's (AAO) Basic and Clinical Science Course (BCSC) Self-Assessment, and OphthoQuestions question bank, which have recently emerged as the standard for question-answering tests for LLMs [2,12]. Both question banks focus on testing clinical knowledge in ophthalmology, particularly in the diagnosis and management of ophthalmic conditions, in addition to fundamental anatomical and physiological knowledge of the eye. Both datasets are not publicly available thus making them unlikely to be included in ChatGPT training dataset. Data permission was given by AAO for BCSC and a personal subscription for OphthoQuestions by SL.

Following the protocol from Antaki et al [13], a subset of 260 multiple-choice questions across 13 sections (20 questions per section) was selected from each question set. The 13 sections are: *Clinical Optics*, *Cornea*, *Fundamentals*, *General Medicine*, *Glaucoma*, *Lens and Cataract*, *Neuro-ophthalmology*, *Oculoplastics*, *Pathology and Tumors*, *Pediatrics*, *Refractive Surgery*, *Retina and Vitreous* and *Uveitis*.

For each question, we collected the question, choices (4 for each question), correct answer, difficulty level, and cognitive level (**Supplementary Material 1**). Finally, 120 questions were selected, none of which were accompanied by graphical information, as the LLM models in this study are not capable of processing image and text simultaneously.

Although this dataset was obtained independently from Antaki et al, the same acquisition method was followed to ensure similar distribution of Difficulty Level and Cognitive Level for comparability across studies [13] (**Supplementary Figure 1).**

Prompting strategy

Zero-shot prompting: As in previous studies [2] a zero-shot approach was used with the prompt *"please select the best answer from the options below and provide an explanation."* followed by the question and the options. However, we refined the prompt following the design introduced by Taori et al 2023 [14] in which we used markdown format to clearly split the prompt to concrete sections (question, task, instruction) (**Figure B**) . This more explicit and organized prompt structure help smaller, weaker LLMs return result with a more consistent and correct format. It's worth noting that we omit *few-short learning* (otherwise known as *in-context learning*), a common prompt engineering technique in LLM research where one or few exemplars (pair of inputs and desired outputs) are appended to the prompt which often guide the model to generate a more accurate response [6]. However, few-shot prompting has been shown to be effective only on large scale models such as Gopher (280B) [15] or GPT-3 (175B) [6]. On the other hand, choosing examples that represent dataset well are a non-trivial tasks and often require some dynamic sampling process to achieve consistent boost in performance [16,17] . For these reasons, we excluded few-shot strategy from this study.

Chain-of-Thought (CoT) is a technique used to improve the reasoning abilities of models by breaking down a problem into a series of intermediate steps or thoughts [5]. By generating a coherent sequence of thoughts, LLMs can provide more accurate and transparent answers to challenging problems, such as mathematical reasoning and logical inference. Additionally, this approach not only improves the performance of AI systems on a variety of cognitive tasks but also enhances the interpretability of their outputs. The CoT technique is particularly effective in scenarios where traditional one-step responses may lead to errors or ambiguities. Since the original CoT technique requires providing step-by-step reasoning on few exemplars in the prompt which we omitted, we instead opted for a more basic form of CoT called *Zero-shot-CoT (hereby ZRS-CoT)* [18] by simply adding *"Let's think step by step"* to the original prompt to guide the language model's reasoning process. This approach has been shown to be a much stronger baseline than simple zero-shot prompt while preserving a minimal and easily constructed prompt that does not require hand-crafted examples.

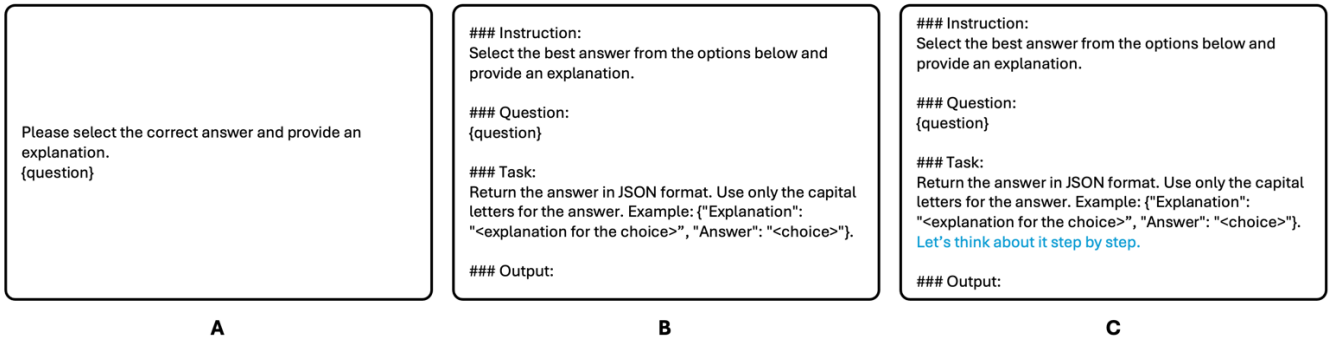

***Figure 1**: **(A)** the original zero-shot prompt from [2]. **(B)** Enhanced prompt following template from Alpaca instruction finetuning strategy [14]. **(C)** Zero-shot Chain of Thought (Zero-shot-CoT).*

Retrieval Augmentation Generation (RAG)

We employed the RAG pipeline to enhance the performance of generative language models by integrating relevant external knowledge into model's prompt [11]. **Figure 2** illustrates our RAG pipeline which has three main components:

Knowledge database stores the external knowledge as vector representations (also often referred to as embeddings) which are numerical encodings that capture the semantic meaning of words, sentences, or documents in a multi-dimensional space. We first split the external source, AAO's 2023-2024 BCSC textbook [19], into smaller units (pages) and convert them into vector representations using OpenAI's *text-embedding-ada-2* model. These embeddings are stored in ChromaDB [20], an open-source vector database designed for efficient storage and retrieval of high-dimensional embeddings. ChromaDB supports fast similarity searches and integrates with various embedding generation models to handle and query large datasets effectively.

Retriever finds the most relevant documents relating to the input query by comparing the similarity between their numerical representations. In our system, the input query is a question and its 4 choices, all of which was converted to an embedding using *text-embedding-ada-2*. ChromaDB (v.0.4.20), with its vector ranking functionality, also serves as the retriever in our RAG.

Re-ranker refines the selection of documents to ensure the highest relevance after the Retriever identifies the top k=30 most relevant candidates. We used Cohere's reranker [21] which offers advanced ranking results, to re-assesses the similarity scores of the retrieved documents and the input query, ensuring the retrieval of the most pertinent data for subsequent generation tasks. This step narrows down the list to the top n=5 most relevant documents, which are then used as context for the final stage of the pipeline. We set n significantly smaller than k to test whether the model can improve

performance with a much smaller amount of context. Smaller context also helps reduce the running cost during deployment.

Generator produces the final answer by taking the top-ranked documents, along with the query, formatting them into a prompt template, and passing the text to an LLM. The Generator leverages the provided context to generate an accurate and contextually relevant response.

Finally, we employed LangChain [22] (v0.0.351), a versatile open-source framework that allows developers to seamlessly connect LLM with external data sources. LangChain provides the tools to integrate all components Knowledge Database, Retriever, Re-ranker, and Generator.

Models

We included three LLMs in our evaluation (**Table 1**):

GPT4-Turbo is the Generative Pretrained Transformer (GPT) language model underlying ChatGPT, a cloud-based chatbot application developed by OpenAI, renowned for its advanced natural language processing capabilities. GPT-4, the latest iteration in this series, follows the release of GPT-3.5 by OpenAI in June 2020. GPT-4 has demonstrated remarkable performance on challenging medical question-answering benchmarks, surpassing its predecessor and even achieving human-level accuracy in many areas [23] . As GPTs are continuously updated, we utilized the most recent version, GPT-4 Turbo, released in November 2023, for this study. Access to GPT-4 is available via an online API, which allows for the automated analysis of large datasets.

Llama-3-70B is a state-of-the-art large language model developed by Meta in April 2024 [24], designed for complex language tasks such as multilingual dialogue, text summarization, and contextual reasoning. With a parameter count of 70 billion, it is part of the Llama 3.1 series, which includes models such as Llama-3.1-8B, and Llama-3.1-405B. Llama-3.1-70B outperforms its predecessor, Llama 2, in various major domains. Additionally, the model maintains competitive speed and cost-efficiency, making it suitable for a range of applications despite its higher hardware requirements (140 GB of GPU VRAM in FP16).

Gemma-2-27B was introduced by Google in June 2024. While Meta's latest model, LLaMA-3, scales to over 70 billion parameters and prioritizes raw performance at large scales, Gemma-2 is designed for efficiency, offering both smaller (9B) and larger (27B) configurations. The 9B model employs knowledge distillation to retain key capabilities of larger models while maintaining computational efficiency. Moreover, Gemma-2 introduces a sliding window attention mechanism, alternating between local (4096

tokens) and global (8192 tokens) attention layers, thus achieving a balance between computational efficiency and the ability to model long-range dependencies.

Mixtral-8x7B is a high-performance Sparse Mixture of Experts (SMoE) language model developed by Mistral AI [25], designed to optimize both computational efficiency and predictive accuracy. The model employs a unique architecture where each layer is divided into 8 experts, with a routing mechanism selecting the top two experts for each token, thereby reducing the effective parameter usage per token to 13 billion from a total of 47 billion parameters. This allows Mixtral to achieve high performance while maintaining low inference costs. The model surpasses Llama-2-70B and matches or exceeds GPT-3.5 on a variety of benchmarks, demonstrating superior capabilities in mathematical reasoning, code generation, and multilingual tasks [25].

***Table 1**: Summary of recently released large language models based on release time, parameter count, and model variants.*

	Time release	By	Number of parameters	Other models in the series
GPT-4-turbo	Nov 2023	OpenAI	N/A	N/A
Llama-3-70B	April 2024	Meta AI	70B	8B, 405B
Gemma-2-27B	June 2024	Google	27B	2B, 9B
Mixtral-8x7B	Dec 2023	Mistral AI	46.7B	Mistral-7B

Quantization

Quantization is a process in machine learning aimed at reducing the precision of model parameters, such as weights and activations, from high-precision floating-point representations (commonly 32-bit, known as FP32) to lower precision formats, such as 16-bit, 8-bit, or even fewer bits [26]. This technique enables efficient computation and storage, as many operations can be executed with lower precision without significantly compromising the model's performance. Empirical evidence indicates that 8-bit quantization maintains performance comparable to full-precision models [27]. Further reduction to 4-bit quantization has been shown to offer additional benefits in terms of model size and inference speed, with minimal impact on performance [28].

In this study, we applied quantization to the weights of the open-source models, converting them to a 4-bit integer (INT4) representation (Q4) from full-precision (FP) float16 or float32. This transformation reduces the number of bits per parameter to 4, leading to a substantial reduction in both disk storage and GPU memory usage, approximately by 75% (e.g. Llama-3-70B’s memory on disk can be reduced

from 140GB (FP) to 40GB (Q4). This reduction is critical for deploying large-scale models in resource-constrained environments.

Statistical analysis

Accuracy was determined by comparing whether LLM's answer was matched with question key (correct or incorrect). For each LLM, accuracy was ascertained in a single run as previous study have shown a great degree of repeatability of GPT-3.5 responses. To evaluate the accuracy of responses across various models, we used generalized estimating equations with an exchangeable correlation structure and a binomial distribution with a logit link, given that the models were tested on identical questions. When significant effects were identified, we conducted post hoc analyses and applied Tukey corrections to adjust the p-values. Python's *statsmodels* (v0.14.2) was used to undertake statistical analysis.

Effect of quantization level and model size

We evaluated model performance and resource efficiency across different quantization levels and sizes. For larger models, including LLaMA-3-70B, Gemma-2-27B, and Mixtral-8x7B, quantization levels from Q8 to Q4 were assessed to gauge the impact of reduced precision on accuracy. Additionally, to capture a broader range of scales, smaller models—Phi-3-mini (3.8B), Phi-3-medium (14B), and Mistral-7B—were also included. This comparison helps provide insights into the most efficient configurations for varying resource constraints.

Compute Infrastructure

The experiments were conducted on an AMD EPYC 9124 CPU (32 cores, 3.0 GHz) with 384GB RAM, and two NVIDIA RTX A6000 GPUs (48GB VRAM each). GPU was distributed when needed. The environment utilized Python 3.10, PyTorch 2.2, transformers 4.44, and llama-cpp-python 0.2 for model loading and evaluation.

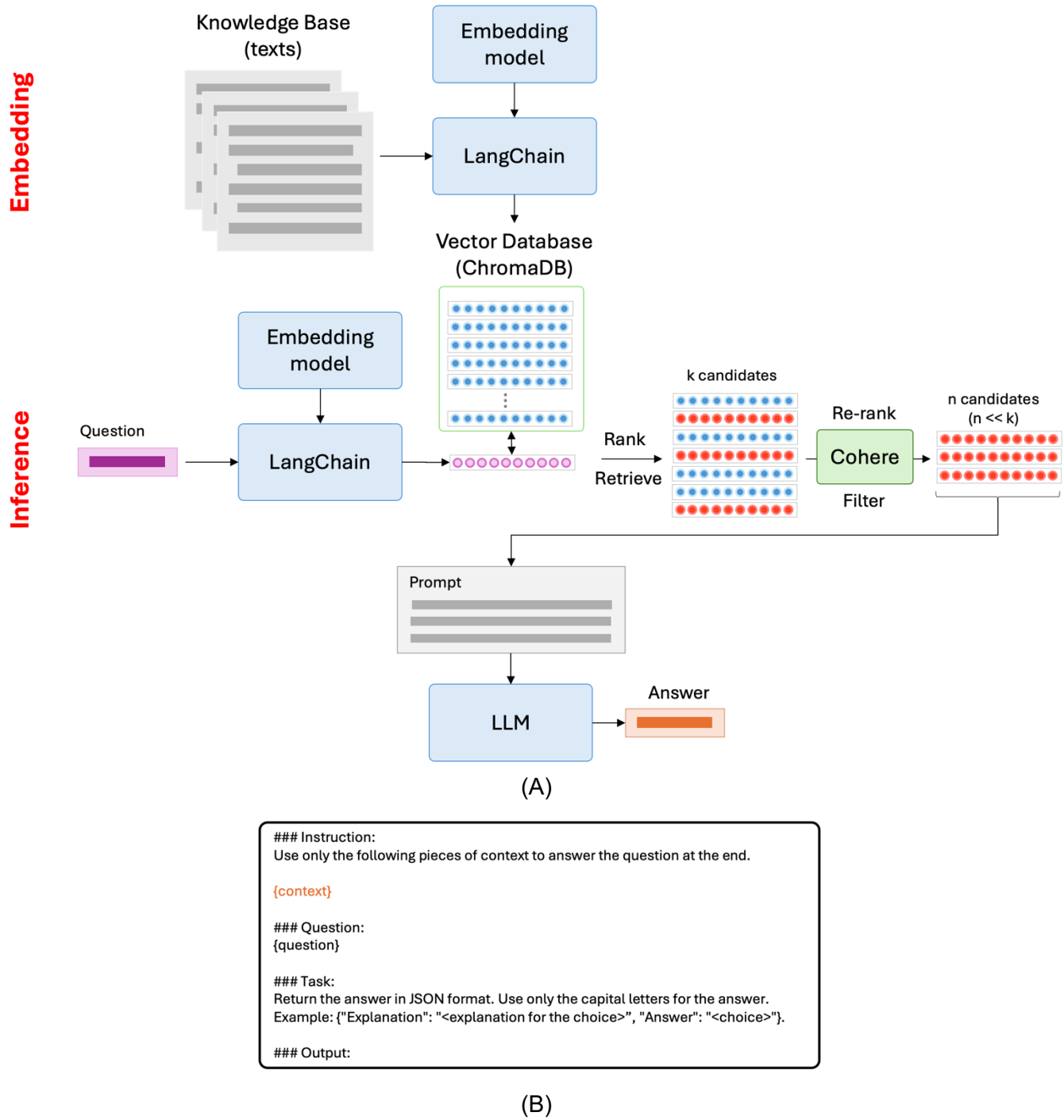

***Figure 2**: **(A)** RAG pipeline. We first split the external source (BCSC textbook) into smaller unit (pages) and convert them to vector using OpenAI's text-embedding-ada-2 model and store the embeddings in a vector database ChromaDB. (Bottom) During inference, we first embedded the question and 4 choices using the same embedding model. ChromaDB was then used to retrieve k=20 most relevant vectors based on similarity score, the list of candidate vectors were further filtered by Cohere's reranker. We only selected the top 5 relevant documents after this step as context. Finally, we put the retrieved documents and the question to a prompt template and send the text to the LLM to get the answer. **(B)** Prompt template which includes task instructions, the concatenation of the top k relevant retrieved documents as {context}, the question and the answer choices as {question} and the output format instruction.*

Results

RAG significantly enhanced the performance of LLMs

On BCSC questions, RAG boosted GPT-4's accuracy from 80.38% to 91.92% (11.54% increase, $p=0.0013$), while the smaller Mixtral-8x7B saw a larger gain from 56.92% to 78.46% (21.54% increase, $p<0.001$). Gemma-2-27B-Q4 improved from 64.23% to 83.46% (19.23% increase, $p<0.001$), and Llama-3-70B-Q4 from 64.62% to 84.62% (20% increase, $p<0.001$). A similar pattern was observed with the OphthoQuestions dataset: GPT-4's accuracy increased by 7.69% (from 77.69% to 85.38%), Llama-3-70B saw a 27.7% gain (50.38% to 78.08%), Gemma-2-24B improved by 15% (60% to 75%), and Mixtral-8x7B by 23.85% (47.69% to 71.54%). Additionally, GPT-4 turbo zero-shot surpassed the previously published GPT-4 results by Antaki et al. [2] for both BCSC (79.03%) and OphthoQuestions (71.7%) .

Llama-3 and Gemma-2, when aided by RAG, surpassed the performance of GPT-4 without RAG. And while Mixtral-8x7B with RAG exhibited a 4% lower performance compared to GPT-4-turbo in the zero-shot setting, this difference was not statistically significant ($p=0.0284$). This suggests that small, open-source model with the additional retrieval component may perform comparably to the baseline GPT-4. The full result breakdown is shown in **Table 1**.

Gemma-2-27B showed exceptional performance, matching Llama-3-70B in all three settings ($p > 0.01$) despite Llama-3-70B having nearly three times more parameters . It also outperformed Mixtral-8x7B, which is twice its size. Surprisingly, despite a nearly 10% difference in mean accuracy in the ZRS, the performance gap between Gemma-2-27B and Mixtral-8x7B was not statistically significant.

***Table 2**: Comparison of GPT-4 and Open-source LLMs' performances in zero-shot (ZRS) zero-shot-CoT (ZRS-CoT) and RAG-enhanced settings (% Accuracy), with performance reference from* [2]*. Performance gain of ZRS-CoT and RAG compared to the zero-shot baseline is provided in the parentheses. Green color indicates improvement, orange reduction.*

Model	mode	BCSC	Ophtho Questions	Mean
GPT-4-turbo	ZRS	80.38	77.69	79.03
	ZRS-CoT	81.54 (1.16↑)	79.62 (1.93↑)	80.58 (1.55↑)
	RAG	91.92 (11.54↑)	85.38 (7.69↑)	88.65 (9.62↑)
Llama-3-70B-Q4	ZRS	64.62	50.38	57.50
	ZRS-CoT	70.77 (6.15↑)	65.77 (15.39↑)	68.27 (10.77↑)
	RAG	84.62 (20.0↑)	78.08 (27.7↑)	81.35 (23.85↑)
Gemma-2-27B-Q4	ZRS	64.23	60.0	62.12
	ZRS-CoT	61.54 (-2.69↓)	56.92 (-3.08↓)	59.23 (-2.89↓)
	RAG	83.46 (19.23↑)	75.0 (15.0↑)	79.23 (17.11↑)
Mixtral-8x7B-Q4	ZRS	57.69	48.08	52.89
	ZRS-CoT	53.85 (-3.84↓)	52.69 (4.61↑)	53.27 (0.385↑)
	RAG	78.46 (20.77↑)	71.54 (23.46↑)	75.00 (22.11↑)
Antanki et al. 2023 [2] (GPT-4, temparature=0.3)	ZSR	75.8	70.8	71.7

Table 3 shows example explanation of GPT-4 where zero-shot setting made a mistake which was corrected by RAG. The first case was clearly a case of hallucination where the internal knowledge of the LLM was incorrect. RAG corrected it by providing guidelines for ROP screening. In the second case, zero-shot answer reflects illustrates an error that could be made by a physician without specialized training in modern cataract extraction techniques e.g. phacoemulsification. Although it may appear logical to withdraw the phacoemulsification probe from the eye when a complication arises, any experienced cataract surgeon would know that this action exacerbates the situation. The correct approach, as outlined in the RAG-aided response, maintain the probe in position while simultaneously injecting ophthalmic viscosurgical devices (OVDs) to stabilize the anterior chamber. In the third example question, the Zero-shot prompt focused on hypotony, which, while a possible complication, is

not the most common. RAG gave a broader perspective by citing specific study data, showing that cataracts affect a significant proportion of trabeculectomy patients. It provided a more evidence-based explanation of the risks, whereas the Zero-shot model focuses on an isolated complication without this context.

Table 3: *Example responses from GPT-4 with zero-shot prompt and RAG where the external sources provided by RAG rectified error from the answer, largely due to hallucination. The source documents retrieved by RAG are provided in the last column.*

	Question	**Zero-shot Answer**	**RAG Answer**	**Source used by RAG**
1	<u>Retina and Vitreous section:</u> A premature infant has immature vascularization in zone I but no evidence of retinopathy of prematurity (ROP) on initial screening examination. What is the appropriate interval for the next screening examination? A. 4 weeks B. 3 weeks C. 2 weeks **D. 1 week** **Correct Answer: D**	For a premature infant with immature vascularization in zone I but no evidence of retinopathy of prematurity (ROP) on initial screening, the recommended interval for the next screening examination is 1-2 weeks. This is because zone I ROP has the highest risk for developing severe ROP, and close monitoring is necessary to detect any progression at an early stage. Therefore, the most appropriate interval for the next screening examination is 2 weeks. **Answer: C**	According to the guidelines, a premature infant with immature vascularization in zone I, even in the absence of ROP, should be reevaluated in 1 week or less. This is because the condition presents a higher risk and requires closer monitoring. **Answer: D**	BCSC: Pediatric Ophthalmology and Strabismus, page 352 "…Recommended Intervals of Follow-up Eye Examinations for ROP Without Plus Disease: 1 Week or Less…"

2	Lens and Cataract section: An ophthalmologist is performing a routine phacoemulsification when she notes that the remaining half of the nucleus has suddenly dropped through an open posterior capsule and appears to be suspended on the anterior vitreous face. What is the best next step? A. lollipopping the nucleus with the phaco tip in order to pull it into the anterior chamber B. vigorous irrigation with a balanced salt solution **C. injection of viscoelastic to stabilize the anterior chamber** D. immediate withdrawal of the phaco instrument and any secondary instrument from the eye **Correct Answer: C**	Immediate withdrawal of the phaco instrument and any secondary instrument from the eye is crucial when a nucleus drops through an open posterior capsule during phacoemulsification. This step helps prevent further manipulation that could exacerbate the situation by pushing the nucleus further into the vitreous or causing additional vitreous prolapse. Once the instruments are removed, the situation can be assessed, and a plan can be made to address the dropped nucleus, possibly involving a vitreoretinal surgeon. Options such as lollipopping the nucleus or vigorous irrigation could worsen the situation by increasing vitreous traction or displacement of the nucleus. **Answer: D**	In the scenario described, where the nucleus has dropped through an open posterior capsule and is suspended on the anterior vitreous face, the best next step is to stabilize the anterior chamber. This can be achieved by injecting a viscoelastic (OVD) to compartmentalize the vitreous and prevent further displacement of the nucleus into the vitreous cavity. This approach helps in managing the anterior segment stability and minimizes the risk of vitreous prolapse and other complications. Options A and B might exacerbate the situation by applying pressure that could push the nucleus further into the vitreous cavity or cause more vitreous to prolapse. Option D, immediate withdrawal of instruments, could create an outward pressure gradient, potentially worsening the situation by bringing more vitreous forward into the anterior chamber. **Answer: C**	BCSC: Lens and Cataract, page 283 "…If capsule rupture occurs during phacoemulsification, lens fragments may enter the posterior segment. At the time of posterior capsule rupture, the surgeon should stabilize the anterior chamber by reducing the high fluid flow and vacuum levels and by compartmentalizing the vitreous with OVD before removing the phaco instrument. The surgeon should avoid immediate withdrawal of the phaco instrument from the eye, because that would result in an outward pressure gradient, bringing more vitreous forward into the anterior chamber and outward toward the incisions….."
3	Glaucoma section: You are offering a trabeculectomy procedure to a 42-year-old man with poorly controlled open-angle glaucoma in the right eye. During the informed consent discussion, the patient asks you "what is the most common complication after trabeculectomy?" What is the most correct answer to this patient's question? A. blebitis B. hypotony	Hypotony, which is an abnormally low intraocular pressure (IOP), is the most common complication following a trabeculectomy. This occurs when the filtration from the trabeculectomy is too effective, leading to a significant drop in eye pressure. Hypotony can lead to various issues such as choroidal effusion, shallow anterior chamber, and potentially vision-threatening complications if not managed appropriately.	The most common complication after trabeculectomy is cataract progression. This is supported by the information that over 40% of patients in the study lost 2 or more lines of Snellen visual acuity over 5 years, most commonly caused by cataract progression and persistent corneal edema.	BCSC: Glaucoma, page 293 "… At 36 months, the average IOP reduction in patients who had cataract surgery was 16.5%. Moreover, 40% of eyes that underwent cataract surgery had a ≥20% reduction in IOP at 36 months. Conversely, the mean IOP in eyes that did not undergo cataract surgery

	C. cataract D. encapsulated bleb **Correct Answer: C**	**Answer: B**	**Answer: C**	was unchanged at 36 months. Mansberger SL, Gordon MO, Jampel H, et al;…”

Zero-shot Chain of Thought did not improve LLM performance

We found no statistically significant difference in performance between the zero-shot and zero-shot Chain of Thought prompting methods across most models, including GPT-4 (p=0.995), Gemma-2 (p=0.9758), and Mixtral-8x7B (p=1.0). The exception was Llama-3, where CoT led to a notable performance improvement of 10.77% (p = 0.0053). Other than that, the prompt showed mixed result with even a slight accuracy reduction with Gemma-2.

Performance by ophthalmic subspeciality section

Figure 3 summarizes models' performance across the 13 ophthalmic subspeciality sections contained in BCSC and OphthoQuestions. In the BCSC dataset (top panel), RAG allowed GPT-4 to consistently performed at or near the top across most sections, particularly excelling in *Fundamentals*, *Clinical Optics*, *Pathology and Tumors*, *Neuro-ophthalmology*, *Uveitis*, and *Refractive Surgery* (95-100% accuracy in RAG setting). Other small, open-source, quantized models Llama-3-70B and Gemma-2-27B lagged behind GPT-4, particularly in challenging sections such as *Periatrics*, *Oculoplastics* and *Glaucoma* where GPT-4 outperformed them by 15-20%. However, they performed relatively well in all other sections with the accuracy only behind GPT-4 by 5%. Notably, RAG had a significant effect on Gemma-2 performance in *Fundamentals* and *Uveitis* by boosting the accuracy by 40% (55% to 95%), and 35% (65% to 100%), respectively. Mixtral-8x7B had lower accuracy compared to the Llama-3-70B and Gemma-2-27B, often lagged behind other two models by 10%. See detailed breakdown in Error! Reference source not found.**a and 2b**.

In OphthoQuestions, *Clinical Optics* and *Pathology and Tumors* appeared to be particularly challenging when all models' performance dipped significantly (more than 15%). GPT-4-turbo had more modest improvement in accuracy using RAG. Most of the improvements were around 10% with the highest being, 20% in *Refractive Surgery* and 15% in *Pediatrics*. In contrast, A higher improvement gain was seen in Mixtral8x7B across all sections of both question banks, with up to 40% in BCSC's Fundamentals and 45% in OphthoQuestions' *Uveitis* section.

Interestingly, RAG had a small negative effect on *Cornea* section for GPT-4 in BCSC (5%). In OphthoQuestions, RAG reduced performance on *Fundamentals*, *Retina* and *Vitreous* sections.

However, RAG did not have negative effect on open-source models with only one exception of Llama-3-70B on BCSC's *Glaucoma* section where ZRS was better than RAG by 10%.

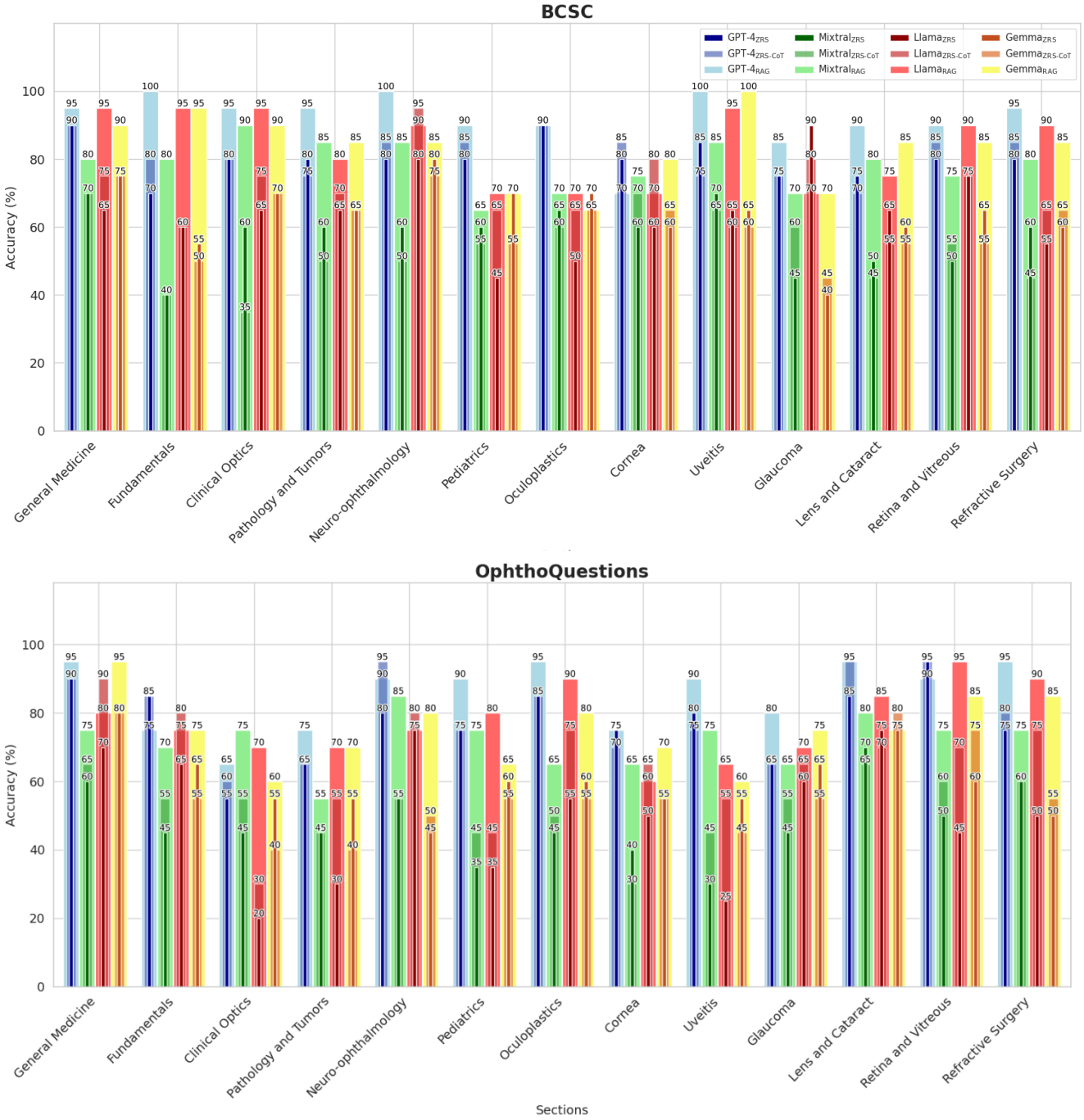

***Figure 3**: Accuracy (%) of GPT4, Mixtral-8x7B, Llama 3-70B-Q4, and Gemma-2-27B-Q4 in three configurations: Zero-shot (ZRS), Zero-shot-Chain of Thought (ZRS-CoT), and Retrieval Augmented Generation (RAG) across 12 sections of BCSC assessment. (Top) BCSC. (Bottom) OphthoQuestions. RAG was run with n=5 retrieved document.*

Four-bit quantization is as effective as eight-bit

Despite eight bit quantization (Q8) showing a slightly higher absolute mean accuracy (1-3%), our experiments revealed no statistically significant difference in performance between four bit (Q4) and eight bit (Q8) quantization levels across all models and settings (Error! Reference source not found.). The only exception was the Llama-3-70B zero-shot setting, where Q8 significantly outperformed Q4 by 11.92% ($p<0.001$).

Effects of model size

Unsurprisingly, as number of model parameters increases, the performance of the models tends to improve. There was a clear positive correlation between model size and performance, indicated by the regressive trend line ($p=0.035$) in *Figure 4* (See detailed breakdown of two datasets in **Supplementary Figure 4**).

The mid-range model Gemma-2 demonstrated outstanding performance compared to smaller (2-6 GB) and medium-sized (7-13 GB) models which occupies a relative same amount of disk memory (Gemma-2-27B-Q8 versus Mixtral-7B and phi-3-medium, or Gemma-2-27B-Q4 versus Llama-3-8B). This indicates that even with similar memory usage, models with more parameters tend to perform better, even when those parameters are of lower precision.

We also observed that small and medium models struggled to surpass 60% accuracy line, further corroborating the finding that models with scale lower than a certain point could not perform complex tasks well regardless of architectural innovations [6].

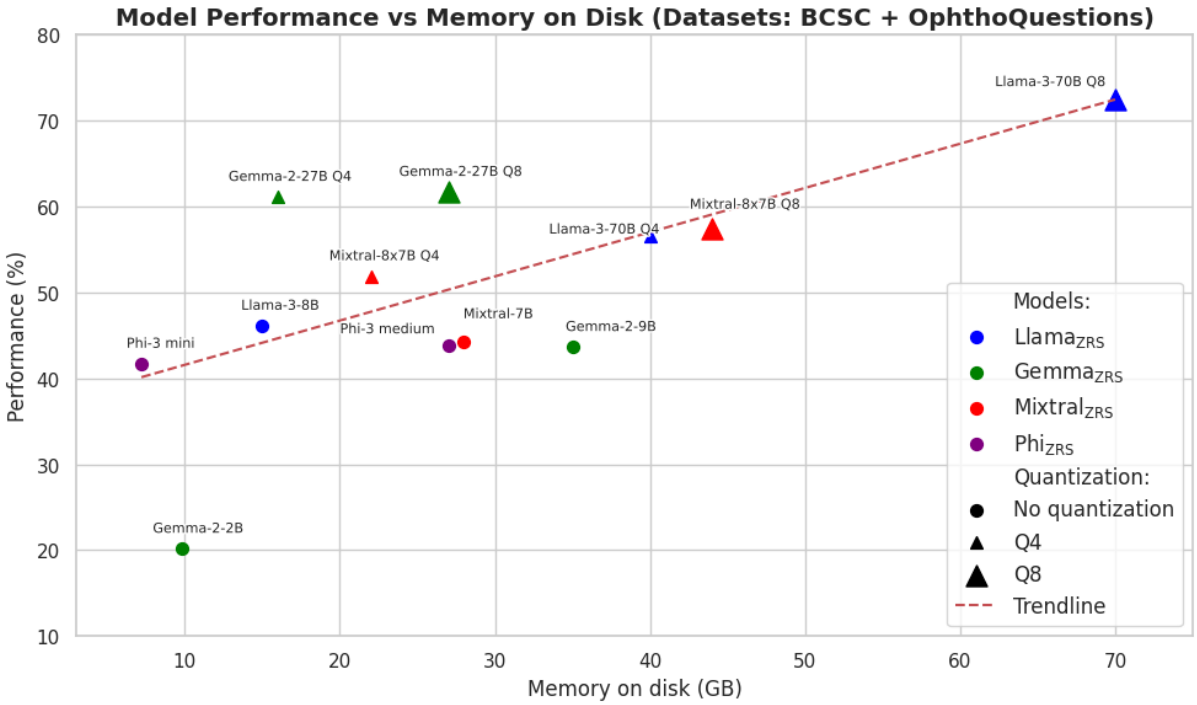

Figure 4: *Relationship between model performance and memory usage for various LLM series (Llama-3, Gemma-2, Mistral and Phi-3) different quantization and prompting strategies. The linear regression trendline (red, dashed) was estimated using numpy.polyfit() (R^2=0.45, p=0.017).*

Discussion

Our findings demonstrate several key insights into the performance of both open-source and proprietary language models on the BCSC and OphthoQuestion ophthalmic question datasets and how their performance can be improved.

Firstly, the zero-shot results achieved in this study surpass those reported previously [2]. This improvement is likely attributable to the more advanced version of GPT-4 employed in our experiments, and more explicit and a well-structured prompt template.

The performance of the surveyed open-source, quantized language modes with RAG was found to be comparable to that of GPT-4 on both datasets and even surpasses the human reference accuracy of 71.91% [2]. This finding is significant as it highlights the potential of open-source models, when enhanced with RAG, to rival much larger proprietary models. These findings offer a promising alternative in resource-constrained and/or privacy-preserving scenarios, where the use of a third-party proprietary model may not be feasible.

The impact of RAG on model performance was substantially more pronounced in this study compared to previous research. For instance, Xiong et al [29] reported only a 1-2% improvement on publicly available medical Question-Answering datasets such as PubMedQA, MedMCQA, MMLU-Med and BioASQ. In contrast, we observed an approximate 10% improvement with RAG on the BCSC dataset. A plausible explanation for this difference is that the BCSC and OphthoQuestions datasets consist of exclusive, paywalled materials that were likely not part of the training data of the language models. Therefore, the additional knowledge provided through RAG played a crucial role in boosting the model's performance. This finding underscores the importance of external knowledge sources when dealing with specialized subject matter and less commonly encountered datasets.

Interestingly, smaller models like Mixtral-8x7B benefited significantly more from RAG compared to GPT-4, with an average performance gain of 22.69%, more than double the 9.62% improvement observed in GPT-4. This suggests that RAG might be particularly advantageous for smaller models, which may lack the extensive pre-existing knowledge of larger models and therefore depend relying more heavily on external data to improve accuracy.

Despite significant improvement by RAG, the OphthoQuestions dataset was challenging with relatively lower improvements observed with RAG [2]. This is particularly pronounced in sections such as *Glaucoma* and *Clinical Optics*, a trend consistent with other studies. The limited performance gain with OphthoQuestions might be attributed to the fact that the external resources utilised by RAG were more aligned with the BCSC content, leading to less significant gains for the OphthoQuestions set.

The results of quantization analysis confirmed that 4-bit precision (Q4) is the best balance between performance and required resourced as previously shown in the literature [28]. Q4 can be used as a viable alternative to Q8, with the added benefit of requiring half the computational resource. In the case of Llama-3-70B, the performance gap between Q4 and Q8 can be bridged effectively by incorporating a simple Chain of Thought prompt "Let's think step by step".

Lastly, zero-shot-CoT was not found to be significantly beneficial to most models including GPT-4. This suggests that in medical question-answering tasks, model performance could be more dependent on the volume of encoded knowledge than on reasoning capabilities introduced through CoT prompting.

A limitation of our study is that we did not include an analysis of the explanations generated by the language models due to the large volume of responses (5,520 in total). Evaluating all these responses manually was beyond the study's scope. Future work could address this limitation by employing smart sampling methods to select a smaller subset of responses for analysis without introducing significant bias.

Additionally, we only utilized the BCSC textbooks as the external knowledge source in this study. It remains unclear whether incorporating additional or more specialized textbooks could boost performance, particularly in weaker areas like Glaucoma. Future studies could explore the potential benefits of expanding the range of resources used for retrieval.

Further improvements in methodology could involve evaluating the relevance of the documents retrieved through RAG, helping to decouple the model's built-in knowledge from external information. A sensitivity analysis on document ranking methods could be valuable. Understanding the impact of different ranking strategies on performance may lead to better optimization of retrieval systems, further improving model accuracy in specialized medical domains. Lastly, conducting a detailed error analysis to identify challenging question types could also provide guidance for refining retrieval algorithms and model training.

In summary, our study is the first to demonstrate the impact of RAG on ophthalmic question sets. Additionally, our study is also the first to evaluate the performance of open-source language models such as Llama-3-70B, Gemma-2-27B, and Mixtral-8x7B on ophthalmic question sets, shedding light on the potential of such models in this specialized medical fields. The findings presented here underscore the potential of RAG for enhancing the capabilities of LLMs in the domain of ophthalmology. By integrating external, dynamically retrieved documents, RAG enables smaller, open-sourced models to efficiently comprehend and process specialized medical information in ophthalmology. Further refinement in document retrieval and ranking strategies could further bridge the performance gap between open-source in-house models and proprietary cloud-based systems, such as GPT-4.

Acknowledgements

QN acknowledges support from a UCL UKRI Centre for Doctoral Training in AI-enabled Healthcare studentship (EP/S021612/1). NP acknowledges support from NIHR AI Award AI_AWARD02488. SL acknowledges support from Medical Research Council Clinical Research Training Fellowship (MR/X006271/1).

References

1. Singhal K, Azizi S, Tu T, Sara Mahdavi S, Wei J, Chung HW, et al. Large Language Models Encode Clinical Knowledge. arXiv [cs.CL]. 2022. Available: http://arxiv.org/abs/2212.13138

2. Antaki F, Milad D, Chia MA, Giguère C-É, Touma S, El-Khoury J, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. Br J Ophthalmol. 2023. doi:10.1136/bjo-2023-324438

3. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural Language Generation. ACM Comput Surv. 2022. doi:10.1145/3571730

4. Qi S, He Y, Yuan Z. Can we catch the elephant? A survey of the evolvement of hallucination evaluation on Natural Language Generation. arXiv [cs.CL]. 2024. Available: http://arxiv.org/abs/2404.12041

5. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv [cs.CL]. 2022. Available: http://arxiv.org/abs/2201.11903

6. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2020. pp. 1877–1901.

7. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. OpenAI; 2019.

8. Stiennon N, Ouyang L, Wu J, Ziegler DM, Lowe R, Voss C, et al. Learning to summarize from human feedback. arXiv [cs.CL]. 2020. Available: http://arxiv.org/abs/2009.01325

9. Tamkin A, Askell A, Lovitt L, Durmus E, Joseph N, Kravec S, et al. Evaluating and mitigating discrimination in language model decisions. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2312.03689

10. Xia Y, Kim J, Chen Y, Ye H, Kundu S, Hao C, et al. Understanding the performance and estimating the cost of LLM fine-tuning. arXiv [cs.CL]. 2024. Available: http://arxiv.org/abs/2408.04693

11. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2020. pp. 9459–9474.

12. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of its Successes and Shortcomings. medRxiv. 2023. p. 2023.01.22.23284882. doi:10.1101/2023.01.22.23284882

13. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings. Ophthalmol Sci. 2023;3: 100324.

14. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, et al. Stanford Alpaca: An Instruction-following LLaMA model. GitHub repository. GitHub; 2023. Available: https://github.com/tatsu-lab/stanford_alpaca

15. Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, et al. Scaling language models: Methods, analysis & insights from training Gopher. arXiv [cs.CL]. 2021. Available: http://arxiv.org/abs/2112.11446

16. Diao S, Wang P, Lin Y, Pan R, Liu X, Zhang T. Active prompting with chain-of-thought for large language models. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2302.12246

17. Qian K, Sang Y, Bayat† F, Belyi A, Chu X, Govind Y, et al. APE: Active learning-based tooling for finding informative few-shot examples for LLM-based entity matching. Proceedings of the Fifth

Workshop on Data Science with Human-in-the-Loop (DaSH 2024). Stroudsburg, PA, USA: Association for Computational Linguistics; 2024. doi:10.18653/v1/2024.dash-1.1

18. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. arXiv [cs.CL]. 2022. Available: http://arxiv.org/abs/2205.11916

19. AMERICAN ACADEMY OF OPHTHALMOLOGY. BASIC AND CLINICAL SCIENCE COURSE COMPLETE SET 2023,2024 (BCSC). Amer Academy Of Ophthalmo; 2021.

20. ChromaDB. [cited 2 Apr 2024]. Available: https://docs.trychroma.com/

21. Reranking. In: Cohere AI [Internet]. [cited 2 Apr 2024]. Available: https://docs.cohere.com/docs/reranking

22. LangChain. [cited 29 Aug 2024]. Available: https://www.langchain.com/

23. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2303.13375

24. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, et al. The Llama 3 herd of models. arXiv [cs.AI]. 2024. Available: http://arxiv.org/abs/2407.21783

25. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, et al. Mixtral of Experts. arXiv [cs.LG]. 2024. Available: http://arxiv.org/abs/2401.04088

26. Nagel M, Fournarakis M, Amjad RA, Bondarenko Y, van Baalen M, Blankevoort T. A white paper on neural network quantization. arXiv [cs.LG]. 2021. Available: http://arxiv.org/abs/2106.08295

27. Dettmers T. 8-bit approximations for parallelism in deep learning. arXiv [cs.NE]. 2015. Available: http://arxiv.org/abs/1511.04561

28. Dettmers T, Zettlemoyer L. The case for 4-bit precision: k-bit inference scaling laws. Proceedings of the 40th International Conference on Machine Learning. JMLR.org; 2023. pp. 7750–7774.

29. Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking retrieval-augmented generation for medicine. arXiv [cs.CL]. 2024. Available: http://arxiv.org/abs/2402.13178

30. Taib F, Yusoff MSB. Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. J Taibah Univ Med Sci. 2014;9: 110–114.

31. American Academy of Ophthalmology (AAO). OKAP User's Guide. San Francisco, CA: American Academy of Ophthalmology (AAO); 2024. Available: https://www.aao.org/Assets/d2fea240-4856-4025-92bb-52162866f5c3/637278171985530000/user-guide-2020-pdf