Advancing Question-Answering in Ophthalmology with Retrieval-Augmented Generations (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], and **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 study demonstrates that integrating Retrieval-Augmented Generation (RAG) significantly enhances the accuracy of large language models (LLMs) in medical question-answering tasks within the ophthalmology domain. This improvement is especially pronounced for smaller, privacy-preserving open-source LLMs, which can be effectively deployed in sensitive and resource-constrained environments such as hospitals. The results highlight that open-source models, when combined with RAG, can provide

© .

a viable alternative to proprietary cloud-based LLMs like GPT-4-turbo. Furthermore, the effectiveness of 4-bit quantization in maintaining model performance while reducing computational resource requirements underscores its potential for optimizing the deployment of LLMs in real-world applications.

1. Introduction

Generative Large Language Models (LLMs) have been shown to demonstrate a remarkable amount of clinical knowledge, both in general medicine Singhal et al. (2023) and within specialist domains such as ophthalmology Antaki et al. (2023a), prompting them to be considered as effective tools for assisting healthcare activities. Antaki et al. demonstrated that GPT-4, the latest iteration of OpenAI's Generative Pretrained Transformer model, performs comparably to humans in answering ophthalmic exam questions Antaki et al. (2023a). 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. Antaki et al. (2023a), there is an inherent tendency of LLMs to produce erroneous, nonfactual information, commonly referred to as hallucination Ji et al. (2022); Qi et al. (2024), 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"* Wei et al. (2022), fine-tuning Brown et al. (2020); Radford et al. (2019) or reinforcement with human feedback (RLHF) Stiennon et al. (2020). 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 Tamkin et al. (2023). For example, Tamkin et al. Tamkin et al. (2023) 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 Brown et al. (2020); Xia et al. (2024), 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 Lewis et al. (2020). 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.

2. Method

2.1. 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 Antaki et al. (2023a,b). 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. Antaki et al. (2023b), 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 S1). 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. (2023b), the same acquisition method was followed to ensure similar distribution of *Difficulty Level* and *Cognitive Level* for comparability across studies (Supplementary Figure S1).

2.2. Formulation

A standard Medical QA task involves of a question q and n candidate answers $a_1, a_2, \ldots, a_n$ where n is typically small (4 or 5). Given an input $x = concat(q, a)$, a Language Model $\mathcal{L}$ attempts to identify the correct option a^* from the candidates given the question q and a question prompt template t_q:

$$a^* = \mathcal{L}(x, t_x)$$

This straightforward vanilla approach is often insufficient for highly specialized, knowledge-intensive fields like medicine, frequently leading to hallucinations or incorrect answers Ji et al. (2022); Farquhar et al. (2024); Pal et al. (2023). RAG aims to provided $\mathcal{L}$ with extra relevant information through a retriever on external knowledge base.

Retrieval: a knowledge base can be considered as a corpus $\mathcal{C} = \{d_1, d_2, \ldots, d_D\}$ which is a collection of D documents. The retriever is a function $\mathcal{R} : (x, C) \rightarrow \mathcal{C}_\mathcal{R}$ that takes an input x and return a subset $\mathcal{C}_\mathcal{R} \in \mathcal{C}$ ($|\mathcal{C}_\mathcal{R}| \ll |\mathcal{C}|$) containing documents most relevant to x. Let $E_R(.)$ be an encoder that maps a text input to a d-dimensional vector, the retriever first calculates the relevancy between a document d_i and x by a similarity score

$$s_r(x, d_i) = E_R(x)^T E_R(d_i)$$

where E_R can be any contextual embedding model like BERT Devlin et al. (2018), or other neural networks trained to capture semantic meaning from text effectively. Then the desired set $\mathcal{C}_\mathcal{R}$ can be selected simply by taking the top k documents that have the highest s_r score:

$$\mathcal{C}_\mathcal{R} = \underset{d \in \mathcal{C}}{\text{Top-K}}\ s_r\left(x, d\right)$$

Re-ranking: While the retriever efficiently filters irrelevant documents, its bi-encoder design, where the query and document are independently encoded, may lack the granularity to rank documents by true relevance. A re-ranker, typically a cross-encoder model, evaluates query-document pairs jointly, capturing fine-grained contextual relationships Dai and Callan (2019); Nogueira and Cho (2019); MacAvaney et al. (2019). By concatenating the query and document (e.g., *[CLS] query [SEP] document [SEP])* and processing them simultaneously with transformer-based models like BERT, the cross-encoder produces a relevance score reflecting nuanced interactions.

Formally, the re-ranker is a function $\mathcal{F} : (x, C_R) \rightarrow \mathcal{C}_\mathcal{F}$ which further refines the subset $\mathcal{C}_\mathcal{R}$ by reordering its documents based on a relevance score

$$s_f(x, d_i) = g(E_F(concat(x, d_i)) \quad \forall d_i \in \mathcal{C}_\mathcal{R}$$

where E_F is the cross-encoder that processes the concatenated query x and document d_i to produce a contextualized representation, and g maps this representation to a relevance score (e.g g could be a linear layer applied to the [CLS] token representation output by a BERT model). The top k documents with the highest $s_f(x, d)$ scores form the final ranked set:

$$\mathcal{C}_\mathcal{F} = \underset{d \in \mathcal{C}_\mathcal{R}}{\text{Top-K}}\ s_f(x, d)$$

This approach ensures precise rankings while remaining efficient by focusing only on the reduced candidate set $\mathcal{C}_\mathcal{F}$ ($|\mathcal{C}_\mathcal{F}| \ll |\mathcal{C}_\mathcal{R}|$).

2.3. Prompting Strategy

Zero-shot prompting: As in previous studies Antaki et al. (2023a) 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. Taori et al. (2023) in which we used markdown format to clearly split the prompt to concrete sections (question, task, instruction) (Figure 1**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 Brown et al. (2020). However, few-shot prompting has been shown to be effective only on large scale models such as Gopher (280B) Rae et al. (2021) or GPT-3 (175B) Brown et al. (2020). 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 Diao et al. (2023); Qian et al. (2024) . 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 Wei et al. (2022). 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) Kojima et al. (2022) 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.

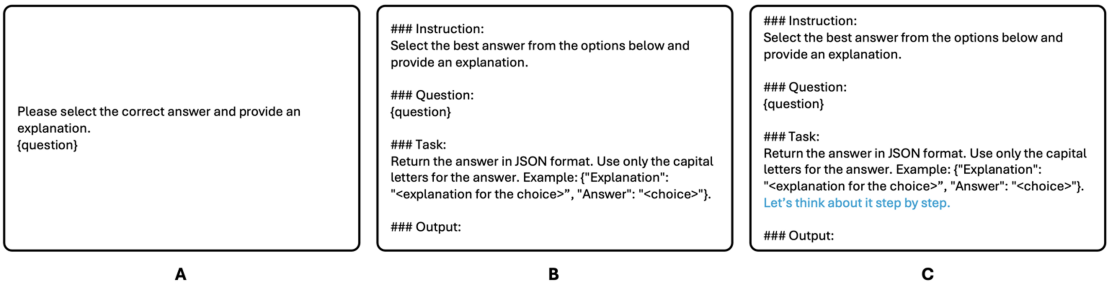

Figure 1: **(A)** Original zero-shot prompt from Antaki et al. (2023a). **(B)** Enhanced prompt following template from Alpaca instruction finetuning strategy Taori et al. (2023) **(C)** Zero-shot Chain of Thought (Zero-shot-CoT).

2.4. 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 Lewis et al. (2020). 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 AMERICAN ACADEMY OF OPHTHALMOLOGY. (2021), into smaller units (pages) and convert them into vector representations using OpenAI's *text-embedding-ada-2* model. These embeddings are stored in ChromaDB[1], 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[2] (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[3] 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.

1. https://www.trychroma.com
2. https://github.com/chroma-core/chroma
3. https://docs.cohere.com/docs/reranking

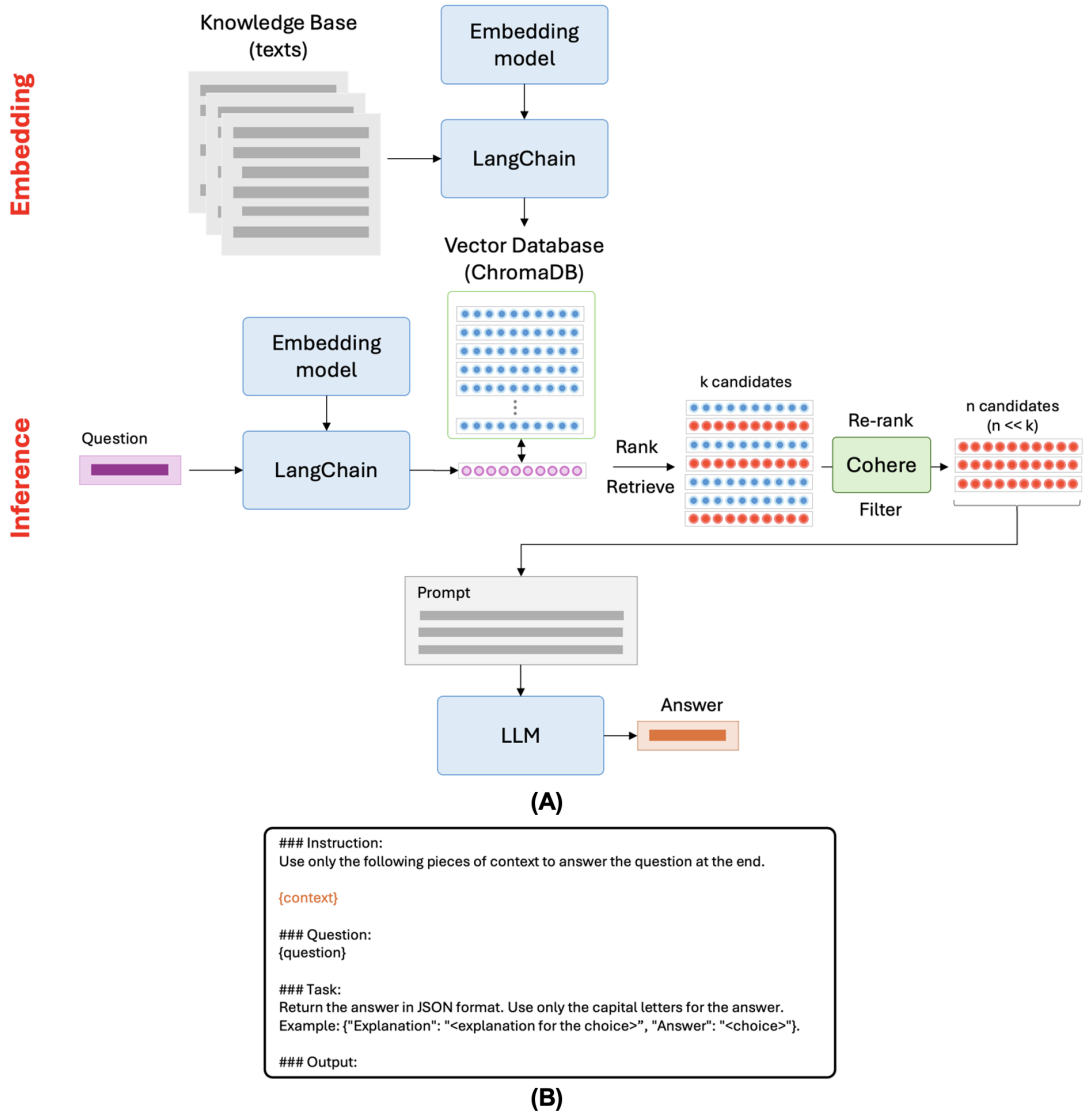

Figure 2: **(A)** RAG pipeline. We first split the external source (BCSC textbook) into smaller unit (pages) and convert them to vector using OpenAI's text-embedding-ada-2 model and store the embeddings in a vector database ChromaDB. (Bottom) During inference, we first embedded the question and 4 choices 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.

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[4] (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.

2.5. Large Language Model

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 Nagel et al. (2021). 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 Dubey et al. (2024), 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 Gemma Team et al. (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 Jiang et al. (2024), 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 Jiang et al. (2024).

2.6. 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 Nagel et al. (2021). 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

4. https://www.langchain.com

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

maintains performance comparable to full-precision models Dettmers (2015). Further reduction to 4-bit quantization has been shown to offer additional reduction in model size and inference speed, with minimal impact on performance Dettmers and Zettlemoyer (2023).

In this study, we applied quantization to the weights of the open-source models, converting them to a 4-bit integer (INT4) and 8-bit (INT8) representations (or Q4 and Q8, respectively) 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.

2.7. 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.

2.8. 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.

2.9. 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.

3. Results

3.1. 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. Antaki et al. (2023b) 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 2.

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 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 focused 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.

Example 1. Retina and Vitreous Section:
Question: 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? **Options:**

A. 4 weeks
B. 3 weeks
C. 2 weeks
D. 1 week
Correct Answer: D

Zero-shot Answer:
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

RAG Answer:
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

Source used by RAG:
BCSC: Pediatric Ophthalmology and Strabismus, page 352
"...Recommended Intervals of Follow-up Eye Examinations for ROP Without Plus Disease: 1 Week or Less..."

Example 2. Lens and Cataract Section:

Question:
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?
Options:
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

Zero-shot Answer:
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

RAG Answer:

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

Source used by RAG:

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..."

Example 3. Glaucoma Section:

Question:

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?

Options:

A. blebitis

B. hypotony

C. cataract

D. encapsulated bleb

Correct Answer: C

Zero-shot Answer:

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.

Answer: B

RAG Answer:

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.

Answer: C

Source used by RAG:

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 was unchanged at 36 months. Mansberger SL, Gordon MO, Jampel H, et al;..."

3.2. 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.

3.3. 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 Supplementary Table S2.

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%.

3.4. 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 (Supplementary Table S1). The only exception was the Llama-3-70B zero-shot setting, where Q8 significantly outperformed Q4 by 11.92% ($p < 0.001$).

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↑)
Antaki et al.(GPT-4)	ZSR	75.8	70.8	71.7

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] (GPT-4 with temperature=0.3). Performance gain of ZRS-CoT and RAG compared to the zero-shot baseline is provided in the parentheses. Green color indicates improvement, orange reduction.

3.5. 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 S3).

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 Brown et al. (2020).

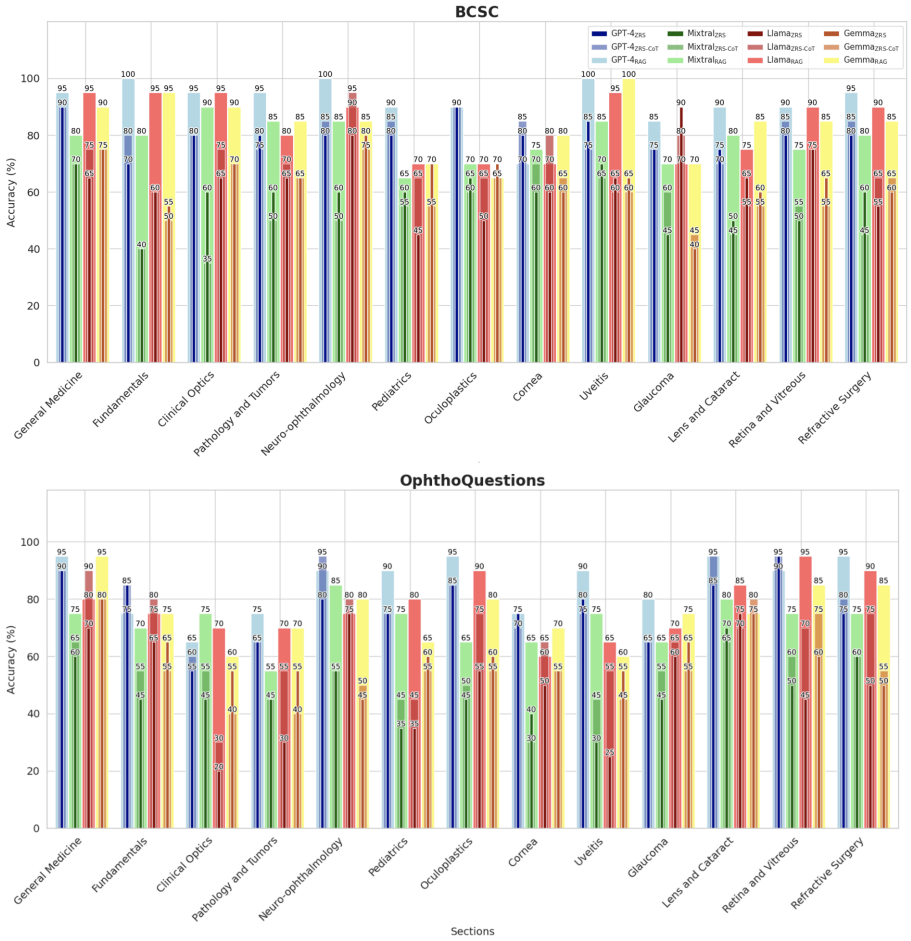

Figure 3: Accuracy (%) of GPT4, Mixtral-8x7B, Llama 3-70B-Q4, and Gemma-2-27B-Q4 in three configurations: Zero-shot (ZRS), Zero-shot-Chain of Thought (ZRS-CoT), and Retrieval Augmented Generation (RAG) across 12 sections of BCSC assessment. (Top) BCSC. (Bottom) OphthoQuestions. RAG was run with $n =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

4. 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 Antaki et al. (2023a). 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% Antaki et al. (2023a). 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

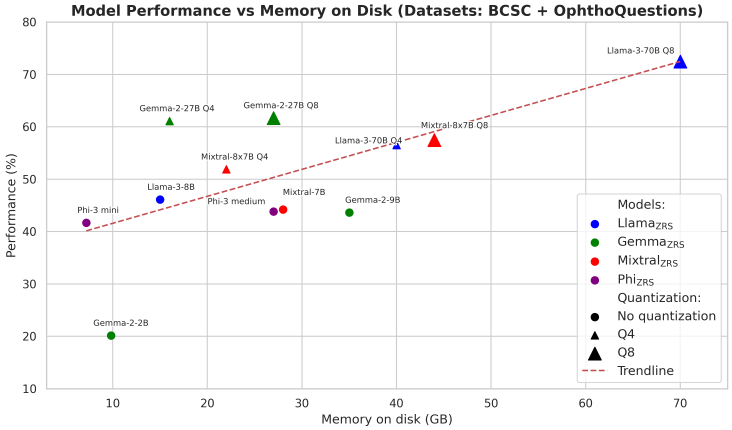

Figure 4: Relationship between model performance and memory usage for various LLM series (Llama-3, Gemma-2, Mistral and Phi-3) different quantization and prompting strategies. The linear regression trendline (red, dashed) was estimated using *numpy.polyfit()* ($R^2 = 0.45$, $p = 0.017$)

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. Xiong et al. (2024) 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 Antaki et al. (2023a). 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 Dettmers and Zettlemoyer (2023). 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.

Acknowledgments

QN is grateful for the support of 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

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

American Academy of Ophthalmology (AAO). OKAP user's guide, 2024.

Fares Antaki, Daniel Milad, Mark A Chia, Charles-Édouard Giguère, Samir Touma, Jonathan El-Khoury, Pearse A Keane, and Renaud Duval. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. *Br. J. Ophthalmol.*, November 2023a.

Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval. Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmol. Sci.*, 3 (4):100324, December 2023b.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

Zhuyun Dai and Jamie Callan. Deeper text understanding for IR with contextual neural language modeling. *arXiv [cs.IR]*, May 2019.

Tim Dettmers. 8-bit approximations for parallelism in deep learning. *arXiv [cs.NE]*, November 2015.

Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 7750–7774. JMLR.org, July 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 1:arXiv:1810.04805, 2018.

Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. Active prompting with chain-of-thought for large language models. *arXiv [cs.CL]*, February 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi,

Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman

Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models. *arXiv [cs.AI]*, July 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-Yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M R Arnold,

Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size. *arXiv [cs.CL]*, July 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, November 2022.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv [cs.LG]*, January 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv [cs.CL]*, May 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-Tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, number Article 793 in NIPS '20, pages 9459–9474, Red Hook, NY, USA, December 2020. Curran Associates Inc.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. CEDR: Contextualized embeddings for document ranking. *arXiv [cs.IR]*, April 2019.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv [cs.LG]*, June 2021.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv [cs.IR]*, January 2019.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Stroudsburg, PA, USA, December 2023. Association for Computational Linguistics.

Siya Qi, Yulan He, and Zheng Yuan. Can we catch the elephant? a survey of the evolvement of hallucination evaluation on natural language generation. *arXiv [cs.CL]*, April 2024.

Kun Qian, Yisi Sang, Farima Bayat†, Anton Belyi, Xianqi Chu, Yash Govind, Samira Khorshidi, Rahul Khot, Katherine Luna, Azadeh Nikfarjam, Xiaoguang Qi, Fei Wu, Xianhan Zhang, and Yunyao Li. APE: Active learning-based tooling for finding informative few-shot examples for LLM-based entity matching. In

Proceedings of the Fifth Workshop on Data Science with Human-in-the-Loop (DaSH 2024), Stroudsburg, PA, USA, 2024. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *arXiv [cs.CL]*, December 2021.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180, August 2023.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. *arXiv [cs.CL]*, September 2020.

Fahisham Taib and Muhamad Saiful Bahri Yusoff. 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.*, 9(2):110–114, June 2014.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions. *arXiv [cs.CL]*, December 2023.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following LLaMA model, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv [cs.CL]*, January 2022.

Yuchen Xia, Jiho Kim, Yuhan Chen, Haojie Ye, Souvik Kundu, Cong Hao, and Nishil Talati. Understanding the performance and estimating the cost of LLM fine-tuning. *arXiv [cs.CL]*, August 2024.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. *arXiv [cs.CL]*, February 2024.