Boosting Healthcare LLMs Through Retrieved Context

Jordi Bayarri-Planas
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
jordi.bayarri@bsc.es

Ashwin Kumar Gururajan
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
ashwin.gururajan@bsc.es

Dario Garcia-Gasulla
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
dario.garcia@bsc.es

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language processing, and yet, their factual inaccuracies and hallucinations limits their application, particularly in critical domains like healthcare. Context retrieval methods, by introducing relevant information as input, have emerged as a crucial approach for enhancing LLM factuality and reliability. This study explores the boundaries of context retrieval methods within the healthcare domain, optimizing their components and benchmarking their performance against open and closed alternatives. Our findings reveal how open LLMs, when augmented with an optimized retrieval system, can achieve performance comparable to the biggest private solutions on established healthcare benchmarks (multiple-choice question answering). Recognizing the lack of realism of including the possible answers within the question (a setup only found in medical exams), and after assessing a strong LLM performance degradation in the absence of those options, we extend the context retrieval system in that direction. In particular, we propose OpenMedPrompt a pipeline that improves the generation of more reliable open-ended answers, moving this technology closer to practical application.

1 Introduction

Large Language Models (LLMs) are the default solution for most text-related tasks. However, a critical concern remains: their factual accuracy[1], a limitation inherent to their generative nature. LLMs are not designed to retrieve precise information, but rather to generate plausible text based on learned patterns. A popular approach to enhance the factuality of LLMs is to contextualize them by biasing their output through relevant input tokens. This ranges from simple prompting techniques, such as "Let's think step by step," to more sophisticated Retrieval Augmented Generation (RAG) systems. Indeed, integrating context retrieval systems can significantly boost the performance and reliability of LLMs.

In the domain of medical multi-choice question-answering (MCQA), the current state-of-the-art is dominated by private models like GPT-4 and MedPalm-2. According to popular evaluation methods, open models significantly lag behind, limiting their practical use. However, a comprehensive assessment must include context retrieval systems and consider more realistic evaluation methods. This study addresses two key research questions: First, how competitive can open LLMs be in healthcare when enhanced with optimized context retrieval systems? Second, how can these systems be extended beyond the limited domain of multi-choice QA?

[1]We consider reasoning to be beyond a text task

To address the first question, this work explores the limits of model boosting through context retrieval by optimizing its components. Open models of various sizes are then enhanced with this optimized system and compared with private solutions. For the second question, we propose an extension of the context retrieval system to generate open-ended (OE) answers, increasing the quality of responses not biased by possible answers.

- C1: Guidelines for an optimized context retrieval setup
- C2: Updated benchmark on open/private LLMs for healthcare MCQA
- C4: Novel context retrieval design for OE answer generation
- C4: Automated library for model boosting

2 Related Work

Addressing the challenge of LLM factuality has spurred extensive research. Initial attempts to improve it centered on harnessing the inherent In-Context Learning (ICL) abilities of LLMs [5], allowing them to adapt to new tasks with minimal examples and without specific training. This paved the way for the development of sophisticated prompting techniques designed to elicit more accurate and reasoned responses. Chain of Thought (CoT) prompting [32] guides LLMs through intermediate reasoning steps, enhancing their performance on complex tasks. In contrast, Self-Consistency (SC) [31] exploits the stochastic nature of LLMs, producing and comparing several outcomes for the same input, before producing a unified answer. Both are combined in Self-Consistency Chain of Thought (SC-CoT). In a similar fashion, tree of Thought (ToT) [33] builds a search space across generated steps towards the answer. At each step several follow-up options are generated, evaluated, and selected, building a pruned tree in the process, from which the final answer is extracted. Reflection methods [23, 25] employ the same generation model as a critique model to iteratively critique the intermediate output and then generate an improved output based on this feedback.

Recognizing the limitations of relying solely on internal knowledge, researchers turned to external knowledge integration through prompting techniques, ultimately leading to the emergence of Retrieval Augmented Generation (RAG) [13]. RAG systems retrieve and integrate relevant information from external knowledge bases, significantly enhancing LLM performance by biasing responses with factual data. In the healthcare domain, few-shot, CoT and SC are recurrently used to improve factuality [15, 21, 24, 29], and a combination of those was proposed in Medprompt [19], a context retrieval system designed for medical MCQA that achieves state-of-the-art results with GPT-4. While Medprompt has been adapted for open-source models like [16], a thorough investigation into the optimal configuration of its components (*e.g.,* DBs, embeddings) remains an open area of research.

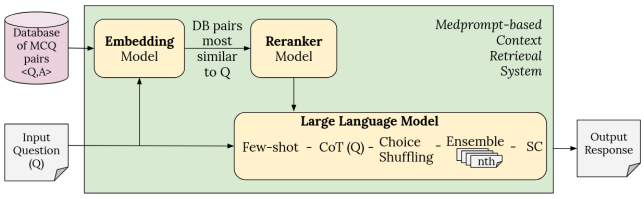

Figure 1: Components of a question-answering system based on context retrieval for LLMs.

Despite recent advances, a significant performance gap persists between large private models and their open-weight counterparts [8, 15]. However, the rapid evolution of accessible LLMs [3, 16], coupled with the potential of optimized RAG systems, suggests that this gap may be narrowing. The ability to leverage medium-sized open models with optimized RAG, potentially achieving performance comparable to larger closed-source alternatives, holds significant promise for reducing adoption costs and increasing accessibility.

At this point, it is worth noting the limitations of MCQA benchmarks. While providing a valuable assessment tool (given its reliable ground truth), it represents a simplified and often unrealistic setup [8] only found in professional medicine exams. Indeed, when considering real-world clinical scenarios, we should not expect the question to include a list of possible answers. That is, models ought to generate open-ended answers. While CoT mechanisms have also been applied to the task of open-ended answers in QAs, contributions in that regard are scarce and non-exhaustive [15, 24].

3 Methods

This section details the methodology employed to optimize a context retrieval system and evaluate its performance. We first outline the core components of the context retrieval system under investigation (§3.1), followed by a description of the benchmark datasets used for evaluation (§3.2), and conclude with the specific models and computational resources employed in the study (§3.3).

Let us start with the retrieval system architecture, which is based on the Medprompt design. Figure 1 illustrates this question-answering system, highlighting the key components that will be explored and optimized in this work.

3.1 Retrieval Components

In this work, we consider the role and impact of the following components:

- **Choice shuffling**. This technique involves randomly shuffling the order of the answer choices presented in multiple-choice questions to mitigate position bias [15, 19]).
- **Number of ensembles**. Refers to the number of independently generated responses produced by the LLM for a given question. These responses are then aggregated, to arrive at the final answer (through techniques such as majority voting). The effect of this component is experimented with the SC-CoT technique.
- **Database.** This component represents the external knowledge source used to contextualize and bias the model. We experiment with two distinct database approaches: utilizing the validation set of the datasets to generate examples in execution time as seen in the original Medprompt methodology, and constructing custom databases derived from the training sets.
- **Embedding model**. The embedding model is crucial for retrieving relevant information from the database. It transforms both the input question and the database entries into numerical vector representations, allowing for similarity comparisons. We evaluate various embedding models, considering factors like dimensionality (e.g., 768 vs. 4096) and domain specificity (e.g., general purpose vs. healthcare-specific).
- **Reranking model**. This optional component aims to refine the initial retrieval results by re-ranking the candidate QA's retrieved from the database based on their relevance to the input question. We employ the MedCPT-Cross-Encoder [12], a specialized medical reranking model trained on a large corpus of biomedical literature.

Beyond these core components, we also explore the impact of hyperparameters such as temperature and the number of few-shot examples, though their impact is less pronounced.

3.2 Datasets

To evaluate the performance of our optimized system, we employ four widely recognized medical Multiple-Choice Question Answering (MCQA) datasets:

- **MedQA** [11]: Consists of 1,273 multiple-choice questions in the format of the US Medical License Exam (USMLE).
- **MedMCQA** [22]: Large-scale multiple-choice question answering dataset with questions from Indian medical school entrance exams. We use the validation set which is composed of 4,183 questions, as the answers of the test set are private.
- **CareQA** [6]: Multiple-choice question answering dataset based on the access exam for Spanish Specialised Healthcare Training. It consists of 5,621 questions.
- **MMLU** [9]: MMLU is a multitask benchmark suite of 57 different datasets spanning domains across STEM. From all these tasks we use the medical related: *anatomy, clinical knowledge, college biology, college medicine, medical genetics* and *professional medicine*, which account for a total of 1,089 questions.

These datasets collectively provide a comprehensive and diverse evaluation platform for assessing the performance of our system across different multi-choice medical question styles and sources.

3.3 Models and Compute

The main model used in our experiments is Llama3-Aloe-8B-Alpha [7], a state-of-the-art open-source LLM specifically fine-tuned for the healthcare domain. With 8 billion parameters, this model builds upon the Meta Llama-3 architecture, leveraging a curated combination of high-quality medical data, synthetic data, and targeted alignment via DPO and Red-teaming datasets. It offers a compelling balance between performance and computational cost.

All experiments are conducted on a single compute node equipped with 4x NVIDIA H100 (64GB) GPUs. We utilize a single GPU for smaller models (<10B parameters) and employ tensor parallelism

across multiple GPUs for larger models (>20B parameters). To assess the environmental impact of our experiments, we monitor the node's power consumption and extrapolate these measurements to estimate the carbon footprint of each experiment, utilizing the latest CO_2 emissions ratio provided by the European Union.

Furthermore, we developed a custom software repository [2] to facilitate this research, which we release open-source to benefit the broader research community. This repository provides a streamlined framework for evaluating various prompt strategies and optimizing context retrieval systems for diverse LLMs. Detailed documentation and an online tutorial are provided to facilitate its adoption.

4 Retrieval Experiments

This section details the experiments conducted to evaluate the impact of different Context Retrieval (CR) components on the performance of Large Language Models (LLMs) in answering healthcare questions. We begin by establishing a baseline performance without any context retrieval (§4.1), followed by a systematic investigation of individual components within the Self-Consistency with Chain-of-Thought (SC-CoT) framework. We then study the impact of adding external knowledge sources and the various components in the Medprompt architecture (§4.2). Based on these findings, we propose an optimized CR configuration (§4.3) and compare its performance against state-of-the-art models (§4.4).

4.1 SC-CoT Experiments

Table 1 presents the baseline performance of Llama3-Aloe-8B-Alpha, our primary evaluation model, on four benchmark datasets using zero-shot next token prediction, CoT, and SC-CoT. This serves as a reference point for evaluating the subsequent improvements achieved through the integration of various components.

	CareQA	MedMCQA	MedQA	MMLU
Zero-shot	67.57	58.91	62.45	72.76
CoT	65.11	55.10	64.26	72.93
SC-CoT	67.64	56.78	64.81	73.68

Table 1: Baseline results of *Llama3-Aloe-8B-Alpha* using zero-shot next token prediction, CoT, and SC-CoT with 5 ensembles.

We investigate the impact of choice shuffling and a variable number of ensembles. The impact of *choice shuffling*, is among the most well-documented phenomenon [15], as LLMs answering MCQs seem to be biased towards the first response (*e.g.,* "A"). Table 2 demonstrates the consistent benefits of incorporating choice shuffling within the SC-CoT framework, yielding improvements across all datasets and ensemble configurations. Based on these results, all subsequent experiments utilize choice shuffling to ensure unbiased evaluation.

Next, we consider the second SC-CoT factor, the *number of ensembles* used in the self-consistency process. This choice has a significant effect, not only on task performance but also on computational cost and footprint. Such trade-off can be seen in Table 3, where performance gains as footprint consistently grow up to 5-6% and 1.76Kg of CO_2 respectively. The first five ensembles provide around 3.5% accuracy gains, while another five adds half of that. Gains beyond that are marginal, requiring up to twenty-five ensembles (fifteen more) to gain another 1%. Figure 2 visually depicts the trend of diminishing returns.

The recommended number of ensembles depends on the criticality of the task, and the available resources. In our experimentation, and unless otherwise specified, all SC experiments will be conducted using 5 ensembles.

N	CareQA	MedMCQA	MedQA	MMLU
5	67.64	56.78	64.81	73.68
5 + CS	+0.85	+2.13	+0.24	+1.88
20	68.89	56.78	64.10	73.79
20 + CS	+1.08	+2.53	+3.53	+3.75

Table 2: Accuracy change when adding choice shuffling (CS) to the baseline with SC-CoT. N represents the number of ensembles of the SC component.

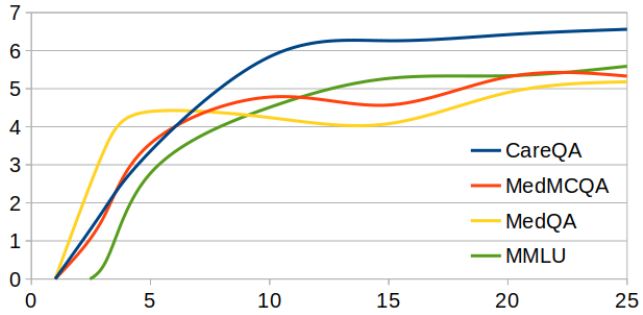

Figure 2: Trends on the accuracy gain obtained by using an increasing number of ensembles (horizontal axis) in a SC-CoT setup.

N	CareQA	MedMCQA	MedQA	MMLU	CO_2
1	65.11	55.10	64.26	72.93	0.08 Kg
3	+1.78	+1.58	+3.30	+0.32	0.20 Kg
5	+3.36	+3.56	+4.40	+2.78	0.33 Kg
10	+5.84	+4.78	+4.24	+4.51	0.69 Kg
15	+6.26	+4.57	+4.08	+5.27	1.02 Kg
20	+6.42	+5.31	+4.90	+5.34	1.38 Kg
25	+6.56	+5.33	+5.18	+5.59	1.76 Kg

Table 3: Accuracy change of SC-CoT with a variable number of ensembles, when compared to the baseline without SC (N=1). CO_2 indicates the associated footprint.

[2] https://github.com/HPAI-BSC/prompt_engine

4.2 Medprompt Experiments

At this point, we extend the SC-CoT scheme with retrieval components, exploiting external data sources. This will include the main elements included in Medprompt: An *embedding* model, a *database*, and *reranker* model. The *embedding* model encodes both input and database items before computing their similarity scores. For this component, we consider four different models, ranging in embedding size and in number of parameters. We also consider whether they have been specialized in the healthcare domain or not. These properties are listed in Table 4, and their performance in Table 5. Results show all models achieve comparable performance in most datasets, with no embedding model clearly outperforming the rest. As a result, we select the healthcare-specific, cheaper model PubMedBERT to be the embedding model for our future experiments.

Model	Domain	Emb. size	Params.
PubMedBERT [18]	Medical	768	109M
MedCPT [12]	Medical	768	109M
UAE-Large-V1 [14]	General	1024	335M
SFR-Mistral [17]	General	4096	7B

Table 4: Properties of the models used to embed the questions.

Embedding	CareQA	MedMCQA	MedQA	MMLU
PubMedBERT	68.65	59.55	69.60	75.55
MedCPT	68.81	59.29	67.16	75.44
UAE-Large-V1	68.08	59.53	69.05	76.70
SFR-Mistral	68.61	60.60	70.15	73.33

Table 5: Accuracy for each embedding model on each dataset. We set 5 as the number of ensembles and few-shot examples, choice shuffling is activated. CoT database in use.

The second retrieval component is the *database*. That is the documental source from which pieces of text are extracted and introduced into the model prompt for contextualization. Both the quality and the diversity of those databases have a large impact on performance. To test this hypothesis we test two setups. First, a smaller database, composed by the validation set of each dataset. For CareQA (which lacks a validation split) we use the MedMCQA validation split as database. The second setup includes a larger and augmented database. Instead of the validation split, we use the training splits of MedMCQA (180K QA pairs, also used for CareQA and MMLU) and MedQA (10K QA pairs).

In Medprompt, these samples are enhanced using CoT. In addition, we consider ToT as well. In both cases, samples are enhanced by prompting an LLM chosen for its instruction-following capability. For CoT, we prompted the Mixtral-8x7B [2] model with the question, the possible answers, and the correct choice, and then asked to analyse each option individually, to explain the answer through detailed reasoning, and to end with a re-identification of the selected option (which is tested for validity). For ToT, we followed the same approach but used Llama-3.1-70B-Instruct to generate the answers. We adapted the original ToT prompt to simulate three logical experts collaborating to answer the question. The size of both databases is exactly the same.

While the MedMCQA and MedQA experiments study the impact of size and quality in databases, the experiments on CareQA and MMLU add a factor of generalization (by using a DB from a different source). Results are shown in Table 6. In three out of four cases, the synthetically enhanced data improves the performance of the retrieval system. Even when the database comes from a different source, the extended database contributes to increase accuracy. Between CoT and ToT, the first outperforms the second in three out of four datasets. All further experiments will make use of the CoT extended databases.

Database	MedMCQA	MedQA	CareQA	MMLU
Validation	68.65	59.55	69.60	75.55
Train+CoT	+7.15	-1.26	+0.78	+3.23
Train+ToT	-1.83	+4.40	+0.63	+1.88

Table 6: Accuracy change when extending the database through CoT and ToT (MedMCQA and MedQA train splits). Also when using a database extended with CoT/ToT coming from a different source (CareQA and MMLU use MedMCQA train split as database).

The final component we study is the use of a *reranker* model. The role of the reranker is to sort the most similar items retrieved from the database, which yields performance boosts in certain domains [30]. To that end, we use the MedCPT-Cross-Encoder model. That is a contrastively pre-trained transformer, tuned on PubMed information retrieval. Table 7 shows the inconsistent performance gains achieved with the reranker, leading us to exclude it from further experimentation due to its added computational cost.

CareQA	MedMCQA	MedQA	MMLU
-0.18	+0.10	+1.02	-1.35

Table 7: Accuracy change when adding the reranker, sorting the samples retrieved from the database.

4.3 Proposed Scheme

Based on our empirical findings, we propose an optimized context retrieval configuration that follows the Medprompt (Figure 1) scheme. The main component removed is the reranker model. The proposed setup utilizes the following:

- **Choice Shuffling**: Enabled to mitigate position bias.
- **Number of Ensembles**: 5 as the default, as this provides the best trade-off between performance gains and computational cost associated. 20 for benchmarking.
- **Database**: CoT-augmented training sets, providing the most comprehensive and effective knowledge source.
- **Embedding Model**: PubMedBERT, a small and efficient healthcare-specific embedding model, is chosen.
- **Reranking Model**: Excluded due to its inconsistent performance gains and added computational overhead.

This setup is the one provided by default in the software library released in association with this work.

4.4 State-of-the-art Comparison

In this section, we benchmark the performance of our optimized CR system against a diverse set of open-source state-of-the-art general-purpose and healthcare-specific LLMs, including models of varying sizes and architectures.

- **Aloe-8B**: A fine-tuned version of Llama3 8B for the healthcare domain, tuned with multi-choice QA data. The retrieval system was optimized for this model, its results may be biased.
- **Llama-3.1-8B/70B** [3]: Latest models released by Meta in 2024. Instruct versions are tuned using supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF).
- **Qwen-2-7B/72B** [4]: Instructed versions of the new series of Qwen large language models.
- **Mistral-7B-Instruct-v0.3** [10]: Third version of the 7B model developed by Mistral AI.
- **Gemma-2-27B-it** [27]: Instruct tuned version of the larger model from the Gemma family, which is a family of state-of-the-art open models from Google.
- **Yi-1.5B-34B-Chat-16K** [1]: The Yi series of models are a family of LLMs trained from scratch by 01.AI. Yi-1.5 is an upgraded version of Yi, continuously pre-trained on top of it. The chat version is the instruct-tuned version.

For each model in this list, we conduct the same evaluation, with and without the context retrieval setup. Results, shown in Table 8, indicate a unanimous boost in performance in all datasets and models. The gains are generalized but non-homogeneous. These are higher on less-performing models, clearly influenced by both the smaller model size and the larger room for improvement.

Performance improvement seems to be highly dependent on the model family. Llama-based models (Llama3-Aloe 8B, Llama 3.1 8B/70B) seem to benefit particularly from context retrieval systems, while other models enjoy lower gains (*e.g.,* Qwen2 72B). LLM performance on a retrieval system is in fact being consistently affected by training policies. There are also differences in gains across benchmarks, with some obtaining consistently higher benefits from context retrieval than others (*e.g.,* MedMCQA). That is a remarkable difference taking into account the huge task similarities shared by all four datasets (*i.e.,* multiple-choice medical question answering). This points towards the importance of data sources, and the challenges of generalization.

Overall, the performance gain provided by the context retrieval scheme is highly valuable, as it reduces the costs of having highly reliable healthcare systems. It shows a well-tuned system based on small LLMs reaching the accuracy levels of much bigger models.

4.5 Private Comparison

To contextualize our results further, we include performance data reported for prominent private models, not been reproduced by this work. The open models under study include:

- **GPT-4** [19, 20]: Developed by OpenAI, with an undisclosed number of parameters but estimated to be at least in the hundreds of billions. We report GPT-4 results on these benchmarks with Medprompt.
- **MedPalm-2** [26]: Developed by Google. We report MedPalm-2, likely based on the Palm-2 model, with Ensemble refinement as a prompting technique.

Model	CareQA	MedMCQA	MedQA	MMLU
Mistral-7B	60.72	48.22	52.32	66.05
with CR	+0.85	+12.24	+9.98	+8.55
Qwen2-7B	68.07	55.06	57.03	73.41
with CR	+1.28	+8.27	+4.71	+3.57
Aloe-8B	67.57	58.91	62.45	72.76
with CR	+4.11	+8.86	+9.98	+6.81
Llama-3.1-8B	70.25	59.31	63.71	75.73
with CR	+3.36	+8.85	+8.88	+5.45
Gemma-2-27B	78.12	61.61	66.93	81.70
with CR	+0.05	+8.77	+2.75	+2.19
Yi-1.5-34B	73.23	57.54	61.43	78.69
with CR	+2.33	+8.68	+10.37	+4.12
Qwen2-72B	83.30	69.33	76.83	86.51
with CR	+2.12	+4.57	+2.44	+2.41
Llama-3.1-70B	83.06	72.17	79.81	87.21
with CR	+4.07	+4.40	+4.79	+3.46
Avg. with CR	+2.27	+8.08	+6.74	+4.87

Table 8: Accuracy of LLMs with and without the context retrieval components proposed, when evaluated on multi-choice medical QA.

Table 9 presents a consolidated leaderboard, sorted by average performance across available datasets, by unifying the results shown in Table 8 with those reported for private models. This comparison reveals that our optimized CR configuration not only boosts the accuracy of open-source models but also enables them to achieve performance levels comparable to much larger private models. In particular, when augmented with our CR system, the Llama-3.1-70B and Qwen-2-72B models demonstrate competitive performance with Google's MedPalm-2 or OpenAI's GPT4.

5 Open-Ended Answer Generation

While multiple-choice question answering (MCQA) benchmarks have been valuable in evaluating Large Language Models (LLMs) for medical applications, they fail to fully capture the complexities of real-world clinical scenarios. In practice, healthcare professionals often need to formulate comprehensive answers without pre-defined options. This necessitates a shift towards open-ended question-answering capabilities in medical AI systems.

Our preliminary analysis revealed a significant performance gap when transitioning from multiple-choice to open-ended formats. Table 10 illustrates this disparity, showing a substantial decrease in accuracy of 10% for the Llama-3.1-8B-Instruct model on the MedQA dataset when transitioning from multiple choice questions (MCQs) to open-ended (OE) questions. Surprisingly, the incorporation of

Model	RAG	MedMCQA	MedQA	MMLU	Avg.
GPT-4	MP	79.1	90.2	94.2	87.83
Llama-3.1-70B	CR	76.57	84.60	90.67	83.94
MedPalm-2	ER	72.3	85.4	89.4	82.36
Qwen2-72B	CR	73.89	79.26	88.92	80.69
GPT-4	5S	72.40	81.40	87.40	80.40
MedPalm-2	5S	71.3	79.7	87.8	79.60
Gemma-2-27B	CR	70.38	69.68	83.89	74.65

Table 9: Accuracy of top performing models in medical MCQA, sorted by average performance. MP: Medprompt. CR: Context Retrieval. ER: Ensemble Refinement (Google's custom prompt technique). 5S: Five-shot Underlined values are reported by others [19, 26].

Chain-of-Thought (CoT) and Tree-of-Thought (ToT) reasoning did not improve performance for open-ended questions, contrary to their effectiveness in other domains.

K	Type	Accuracy
0	Multiple-choice	63.71
0	Open-Ended	53.34
5	CoT Open-Ended	52.40
5	ToT Open-Ended	51.93

Table 10: Baseline results of Llama-3.1-8B-Instruct for the MedQA dataset

To address this gap, we propose OpenMedprompt, an extension of our optimized retrieval system specifically designed for open-ended medical question answering. We introduce the methodology in §5.1, our modifications for datasets and databases for OE generation in §5.2, the evaluation procedure in §5.3 and the results and discussion in §5.4.

5.1 OpenMedprompt

OpenMedprompt adapts the Medprompt architecture for open-ended question answering. We replace the MCQA database with an OE-QA database and remove components specific to multiple-choice formats. We propose two strategies for consensus-building and answer refinement:

OpenMedprompt with Ensemble Refining (OM-ER): This strategy leverages the diversity of multiple generated answers to produce a refined and more accurate final response. It involves generating N initial answers with randomized temperature and top_p parameters, incorporating K relevant examples from the database into the prompt. Then, the LLM synthesizes these N answers into a single, refined response.

OpenMedprompt with Self-reflection(OM-SR) [23, 25]: This strategy employs a feedback loop to improve the generated answer. It begins by generating an initial answer using the K most similar examples from the database. Then, it performs N iterations of self-reflection, where the model generates feedback on its previous response and produces an improved answer based on this feedback. We integrate attribute scores from ArmoRM-Llama3-8B [28], a reward model along with the critique model's reflection as an external feedback to guide answer generation.

The ArmoRM-Llama3-8B reward model provides reward scores across 19 attributes. Out of the nineteen total, we give the generation model only the following scores *ultrafeedback-truthfulness*, *ultrafeedback-honesty*,*ultrafeedback-helpfullness* and *prometheus-score* which have a good correlation with correct responses.

Figure 3 and Figure 4 illustrate the architecture of OM-ER and OM-SR, respectively.

5.2 Dataset and Database Construction

We utilized the MedQA dataset, which is particularly suitable for this task as it contains medical questions that can be answered without providing multiple-choice options, with minimal or no rephrasing of the original question required, making it ideal for testing our proposed framework.

To create a robust context retrieval system, we construct two distinct databases using the MedQA training set. We use LLaMA-3.1-70B-Instruct to generate CoT and ToT answers for each question in the training set. Both databases are carefully curated to ensure the accuracy of the generated responses and their alignment with the ground truth answers.

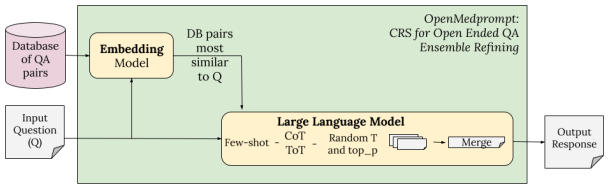

Figure 3: Components of the OpenMedprompt with Ensemble Refining (OM-ER) system.

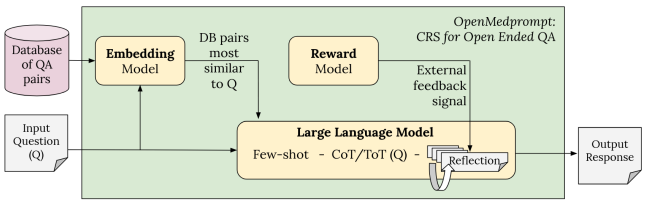

Figure 4: Components of the OpenMedprompt with Self-reflection(OM-SR) system.

5.3 Experiments

We evaluate the performance of OM-ER and OM-SR using an automated LLM-based judge (LLaMA-3.1-70B-Instruct). This judge assesses the correctness of the generated open-ended answers by comparing them to the ground truth answers.

Table 11 presents the results of our experiments of OM-ER using the CoT and ToT databases. We see that both CoT and ToT databases show improvement over the baseline open-ended performance. ToT

N	CoT-Accuracy	ToT-Accuracy
3	56.17	57.82
5	56.40	56.40
7	56.87	57.58
10	56.72	59.54
15	59.31	59.00
20	58.76	60.02

Table 11: OM-ER accuracy results of Llama-3.1-8B-Instruct in MedQA dataset using the CoT and ToT database. *N* represents the number of ensembles.

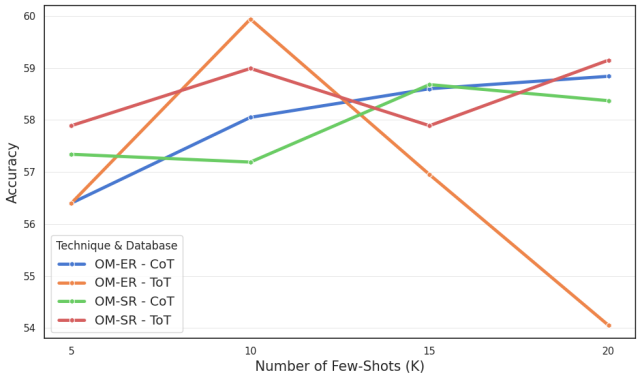

Figure 5: OM-ER and OM-SR accuracy when changing the number of few-shots (K) with a fixed number of ensembles (5).

generally outperforms CoT, with the best result (60.02%) achieved using 20 ensembles.

Table 12 shows the results for OM-SR with the CoT and ToT databases. We see that the OM-SR system shows more consistent improvement over the baseline compared to the OM-ER system. The CoT database performance peaks at N=15 (60.88%), outperforming ToT in this configuration. The ToT database on the other hand shows less variation across different N values, suggesting more stable performance.

N	CoT-Accuracy	ToT-Accuracy
3	55.93	57.50
5	57.34	57.89
7	59.07	58.21
10	60.25	60.02
15	60.88	58.52
20	58.68	58.68

Table 12: OM-SR accuracy results of Llama-3.1-8B-Instruct in MedQA dataset using the CoT and ToT database. *N* represents the number of refinement iterations.

While we expect performance to increase when the number of few shots increases we notice that in Figure 5 this isn't the case for OM-ER ToT. This could be attributed to the effective context length

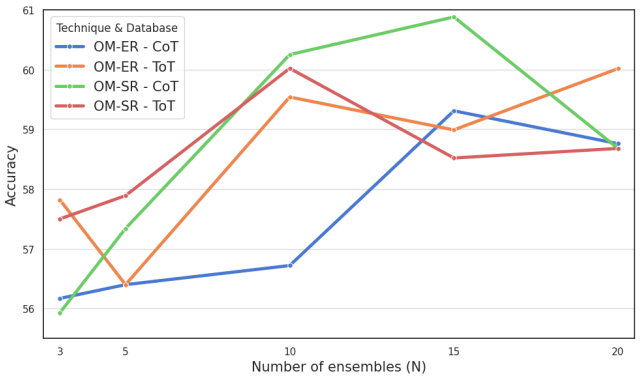

Figure 6: OM-ER and OM-SR accuracy when changing the number of ensembles (N) with a fixed number of few-shots (5).

of the LLM. ToT responses are very verbose and when increasing the number of few-shots the context size significantly grows making it difficult for the LLM to reason over this large context.

We also see that the increase in the number of ensembles doesn't consistently improve the accuracy in Figure 6. OM-ER benefits from ensemble sizes up to 20 while for the OM-SR method ensemble size of 15 provides the best performance for CoT while ensemble size 10 provides the best performance for ToT.

5.4 Results and Discussion

Our results demonstrate the effectiveness of OpenMedprompt in improving open-ended answer generation accuracy in the medical domain. Both OM-ER and OM-SR contribute to performance gains compared to the baseline. The choice of database and the number of retrieved examples also play a significant role in performance.

Specifically, we observe that OM-SR generally outperforms OM-ER across most configurations. This suggests that the iterative feedback loop and incorporation of reward model scores provide a more effective mechanism for refining the generated answers. The choice of database (CoT vs. ToT) also has an impact on the performance differently for each approach. OM-ER benefits more from the ToT database, while OM-SR shows stronger results with the CoT database at higher iteration counts. OM-SR often achieves good performance with fewer iterations compared to the number of ensembles required for OM-ER to reach similar accuracy levels. However, OM-ER might be preferred when speed and simplicity are prioritized, or when exploring diverse perspectives is beneficial.

Choosing the Right Configuration:

- **Accuracy Priority**: Choose OM-SR, particularly with higher iteration counts ($N \geq 10$).
- **Complex Reasoning**: OM-SR's self-reflection mechanism may be more effective for questions requiring intricate logical steps.
- **Speed and Simplicity**: OM-ER with moderate ensemble sizes ($N = 5-10$) offers a good balance of improved accuracy and computational efficiency.

- **Diverse Perspectives**: OM-ER inherently generates multiple initial answers, which can be valuable when diversity of thought is important.
- **Task**: The OM-ER approach is not suitable when large output texts are expected, as models start running out of effective context windows when merging the responses to create a unified answer.

6 Conclusions

This work underscores the significant potential of augmenting Large Language Models (LLMs) with context retrieval systems to enhance their accuracy and reliability in the healthcare domain. Our exploration of Self-Consistency with Chain-of-Thought (SC-CoT) components revealed substantial gains through choice shuffling and an optimal number of ensembles, striking a balance between performance and computational cost. Further investigation into the Medprompt architecture highlighted the effectiveness of small, healthcare-specific embedding models and the value of enriching databases with Chain-of-Thought augmented examples. Conversely, the inclusion of a reranking model was found to be computationally expensive with inconsistent benefits, leading us to recommend against its use.

Our optimized context retrieval configuration, when applied to a diverse set of open-source LLMs, consistently boosted performance across multiple medical question-answering benchmarks. Notably, smaller models experienced the most significant improvements, showcasing the ability of well-tuned retrieval systems to bridge the performance gap between smaller open models and larger private alternatives. This finding has profound implications for democratizing access to high-performing healthcare AI systems, reducing reliance on resource-intensive large models. Moreover, our results demonstrate that open LLMs augmented with our optimized CR system can achieve accuracy comparable to, and in some cases surpassing, state-of-the-art private solutions like MedPalm-2 and GPT-4.

Recognizing the limitations of multiple-choice question answering (MCQA) in mirroring real-world clinical scenarios, we extended our approach to develop OpenMedprompt, a novel framework for open-ended medical question answering. Two distinct strategies, OpenMedprompt with Ensemble Refining (OM-ER) and OpenMedprompt with Self-Reflection (OM-SR), were introduced and evaluated, revealing their effectiveness in improving open-ended answer generation accuracy. OM-SR, with its iterative feedback loop and integration of reward model scores, generally outperformed OM-ER, suggesting the potential of self-reflection mechanisms for complex medical reasoning.

By releasing our custom software repository as an open-source resource, we aim to empower the research community to further explore and refine these techniques, contributing to the development of more accurate, accessible, and impactful LLM applications in healthcare.

6.1 Future Work

Several avenues for future research remain, encompassing both the optimization of retrieval-augmented generation for multiple-choice questions and the further development of OpenMedprompt for open-ended answer generation.

Expanding Retrieval System Capabilities: **Dynamic Retrieval:** Exploring dynamic retrieval techniques that adapt the number of retrieved examples based on the complexity of the question, potentially improving efficiency and accuracy. **Multi-Database Integration:** Investigating the integration of multiple knowledge sources, such as medical ontologies or clinical guidelines, to enrich the retrieval database and enhance the LLM's understanding of complex medical concepts. **Cross-Lingual Retrieval:** Adapting the retrieval system to support multiple languages, facilitating broader access to medical information and enabling cross-lingual medical question answering. Enhancing OpenMedprompt: **Hybrid Approaches:** Exploring combinations of OM-ER and OM-SR to leverage the strengths of both methods, potentially leading to a more robust and adaptable system for open-ended medical question answering. **Advanced Reward Models:** Investigating more sophisticated reward models tailored specifically to medical knowledge evaluation, capturing nuanced aspects of medical reasoning, factual accuracy, and clinical relevance. **Prompt Engineering:** Fine-tuning prompt structures for both initial answer generation and refinement stages, incorporating specific instructions, constraints, or contextual information to guide the LLM towards more accurate and comprehensive responses. **Larger Models:** Evaluating the performance of OpenMedprompt with more powerful language models to assess scalability and explore the potential for further accuracy gains. **Domain Adaptation:** Extending OpenMedprompt to other specialized domains beyond medicine, such as law or engineering, focusing on tailoring the retrieval database and reward models to the specific knowledge requirements of each domain.

6.2 Carbon Footprint

	Time	Power consumption	Footprint
Medprompt	652.90	828.47	169.84
OpenMedprompt	195.40	134.01	27.47
Total	848.30	962.48	197.31

Table 13: Computational cost of the experiments conducted in this study. Time is expressed in hours, power consumption in KWh and carbon footprint in kg/CO2.

6.3 Ethical Considerations

This work seeks to increase the factuality and reliability of LLM-based systems. However, any work that deals with generative models like LLMs should be done in awareness of the several ethical limitations that the technology entails. This includes the computational footprint (which is considered in this work), as well as the factuality (which is the main target of this work). Limitations related to the potential for impersonation, self-diagnose and other non-approved uses, are further discussed in publications related with healthcare model releases (*e.g.,* [7]).

Acknowledgments

This work has been partly funded by the AI4Europe projects from the European Union's Horizon 2020 programme (Grant Agreements N°951911 and N°101070000), and by a SGR-GRE grant from the Generalitat de Catalunya (code 2021 SGR 01187).

References

[1] 01. AI, :, and Alex Young et al. 2024. Yi: Open Foundation Models by 01.AI. arXiv:2403.04652 [cs.CL] https://arxiv.org/abs/2403.04652

[2] Mistral AI. 2024. Mixtral of Experts. https://mistral.ai/news/mixtral-of-experts/

[3] Abhimanyu Dubey et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[4] An Yang et al. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671* (2024).

[5] Tom B. Brown et al. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[6] Lucia Urcelay Ganzabal and Pablo Bernabeu Pérez. 2024. *careqa.* https://huggingface.co/datasets/HPAI-BSC/CareQA

[7] Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Jordi Bayarri-Planas, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Lucia Urcelay-Ganzabal, Marta Gonzalez-Mallo, Sergio Alvarez-Napagao, Eduard Ayguadé-Parra, and Ulises Cortés Dario Garcia-Gasulla. 2024. Aloe: A Family of Fine-tuned Open Healthcare LLMs. arXiv:2405.01886 [cs.CL] https://arxiv.org/abs/2405.01886

[8] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine* (2024), 1–10.

[9] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY] https://arxiv.org/abs/2009.03300

[10] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[11] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11, 14 (2021), 6421.

[12] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics* 39, 11 (2023), btad651.

[13] 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. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/2005.11401

[14] Xianming Li and Jing Li. 2023. AnglE-optimized Text Embeddings. *preprint arXiv:2309.12871* (2023).

[15] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns* 5, 3 (2024).

[16] Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. 2024. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports* 14, 1 (2024), 14156.

[17] Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. SFR-Embedding-Mistral:Enhance Text Retrieval with Transfer Learning. Salesforce AI Research Blog. https://blog.salesforceairesearch.com/sfr-embedded-mistral/

[18] David Mezzetti. 2023. Embeddings for Medical Literature. (2023). https://medium.com/neuml/embeddings-for-medical-literature

[19] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. arXiv:2311.16452 [cs.CL] https://arxiv.org/abs/2311.16452

[20] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[21] Ankit Pal and Malaikannan Sankarasubbu. 2024. Gemini goes to med school: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. *arXiv preprint arXiv:2402.07023* (2024).

[22] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. arXiv:2203.14371 [cs.CL] https://arxiv.org/abs/2203.14371

[23] Matthew Renze and Erhan Guven. 2024. Self-Reflection in LLM Agents: Effects on Problem-Solving Performance. arXiv:2405.06682 [cs.CL] https://arxiv.org/abs/2405.06682

[24] Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H. Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digital Medicine* 7, 1 (2024), 20. https://doi.org/10.1038/s41746-024-01010-1

[25] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366 [cs.AI] https://arxiv.org/abs/2303.11366

[26] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617 [cs.CL] https://arxiv.org/abs/2305.09617

[27] Gemma Team and Morgane Riviere et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118 [cs.CL] https://arxiv.org/abs/2408.00118

[28] Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. [n. d.]. Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. *arXiv preprint arXiv:2406.12845* ([n. d.]).

[29] Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine* 7, 1 (2024), 41. https://doi.org/10.1038/s41746-024-01029-4

[30] Xiaodan Wang, Lei Li, Zhixu Li, Xuwu Wang, Xiangru Zhu, Chengyu Wang, Jun Huang, and Yanghua Xiao. 2023. Agree: Aligning cross-modal entities for image-text retrieval upon vision-language pre-trained models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 456–464.

[31] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171 [cs.CL] https://arxiv.org/abs/2203.11171

[32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] https://arxiv.org/abs/2201.11903

[33] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601 [cs.CL] https://arxiv.org/abs/2305.10601