To Generate or to Retrieve? On the Effectiveness of Artificial Contexts for Medical Open-Domain Question Answering

Giacomo Frisoni♠* Alessio Cocchieri♠* Alex Presepi♠
Gianluca Moro♠* Zaiqiao Meng♢
♠Department of Computer Science and Engineering, University of Bologna
♢School of Computing Science, University of Glasgow
♠{giacomo.frisoni, a.cocchieri, gianluca.moro}@unibo.it
♠alex.presepi@studio.unibo.it ♢zaiqiao.meng@glasgow.ac.uk

Abstract

Medical open-domain question answering demands substantial access to specialized knowledge. Recent efforts have sought to decouple knowledge from model parameters, counteracting architectural scaling and allowing for training on common low-resource hardware. The retrieve-then-read paradigm has become ubiquitous, with model predictions grounded on relevant knowledge pieces from external repositories such as PubMed, textbooks, and UMLS. An alternative path, still under-explored but made possible by the advent of domain-specific large language models, entails constructing artificial contexts through prompting. As a result, "to generate or to retrieve" is the modern equivalent of Hamlet's dilemma. This paper presents MEDGENIE, the first generate-then-read framework for multiple-choice question answering in medicine. We conduct extensive experiments on MedQA-USMLE, MedMCQA, and MMLU, incorporating a practical perspective by assuming a maximum of 24GB VRAM. MEDGENIE sets a new state-of-the-art in the open-book setting of each testbed, allowing a small-scale reader to outcompete zero-shot closed-book 175B baselines while using up to 706× fewer parameters. Our findings reveal that generated passages are more effective than retrieved ones in attaining higher accuracy.[1]

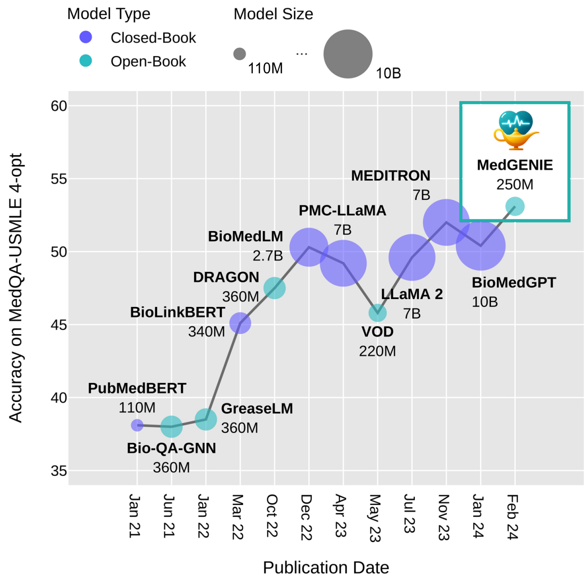


Figure 1: MEDGENIE performance (Flan-T5-base, Fusion-In-Decoder) on USMLE-style questions. Comparison against fine-tuned open-source baselines with a maximum of 10B parameters, using the MedQA (4 options) test set. Model size displayed on a log scale.

1 Introduction

Question answering is a challenging task that requires complex reasoning on explicit constraints described in the question and unstated domain knowledge. Open-domain question answering (ODQA) aims to tackle natural questions across various topics without predefined evidence (Chen et al., 2017). This setting mirrors real-world scenarios where there cannot be a labeled passage for every potential user inquiry. ODQA has particular significance in medicine due to the high-quality standards it demands, including a deep understanding of specialized terminology and background concepts, and an effective recall of expert insight for clinical decision making (Frisoni et al., 2022).

Recent efforts have transitioned from a *closed-book* strategy, where models rely solely on their opaque parametric knowledge, to an *open-book* alternative, allowing them to consult external sources for grounding. In particular, the *retrieve-then-read* framework is a common thread (Zhu et al., 2021; Zhang et al., 2023a), where the input is augmented with relevant knowledge chunks *retrieved* from an external datastore, which can be unstructured (e.g., PubMed, textbooks) or structured (e.g., UMLS). However, performance is highly depen-

* Equal contribution (co-first authorship).
[1]Our code, fine-tuned models, and generated contexts are publicly available at https://github.com/unibo-nlp/medgenie.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9878–9919
August 11-16, 2024 ©2024 Association for Computational Linguistics

dent on the quality of the retriever. Developing custom retrieval modules generally requires extensive question–context pairs (Karpukhin et al., 2020) or intensive computational resources (Lee et al., 2019), particularly when dealing with massive sources (Gan et al., 2023). Furthermore, the retrieved fragments may be incomplete or not specifically tailored to the query, leading to noise (Oh and Thorne, 2023).

In parallel, medical large language models (LLMs) have gained increasing research interests to aid professionals and improve patient care (Zhou et al., 2023). After pre-training on an extreme-scale collection of specialized text corpora, they implicitly encode an impressive amount of domain knowledge that can be evoked through prompting, akin to summoning a genie from a lamp. This facilitates a paradigm shift towards a *generate-then-read* approach, wherein contexts are directly *generated* by an LLM. Despite preliminary work (Yu et al., 2023; Su et al., 2023; Zhang et al., 2023b), there is an ongoing debate on "whether generative augmentation is preferable to retrieval augmentation".

In this paper, we introduce MEDGENIE, the first *generate-then-read* framework for multiple-choice medical ODQA. Specifically, we study the effectiveness of grounding generalist LLMs and small language models (SLMs) on multi-view contexts generated by a medical LLM via in-context learning (ICL) and fine-tuning, respectively. To foster accessibility and match prevalent hardware configurations, we assume a low-resource infrastructure with 24GB VRAM.

We evaluate MEDGENIE on three standard ODQA benchmarks designed to quantify professional medical competencies: MedQA-USMLE (Jin et al., 2020), MedMCQA (Pal et al., 2022), and MMLU-Medical (Hendrycks et al., 2021). MEDGENIE demonstrates significant performance gains, improving the accuracy of few-shot LLM readers on all testbeds by up to ≈16 points. By fine-tuning the reader, MEDGENIE allows Flan-T5-base to outcompete closed-book zero-shot 175B LLMs and supervised 10B baselines on MedQA (Figure 1), using up to 706× fewer parameters. Furthermore, our research demonstrates a clear inclination of cutting-edge rerankers towards favoring generated contexts over retrieved ones. When treated as knowledge sources or incorporated into human-curated ones, generated passages notably enhance the effectiveness of retrieve-then-read workflows (up to ≈6 extra points). RAGAS evaluation confirms the quality of generated contexts, even allowing for more faithful answers from the reader. Finally, we release a comprehensive dataset of ≈1 million artificial contexts in the medical field, adhering to principles of open science and encouraging further research endeavors.

2 Related Work

Medical Language Models Transformer-based language models have become the go-to solution for any task in medical NLP. Early work on domain adaption focused on encoder-only models pretrained on PubMed articles and abstracts, counting 40+ models introduced between 2019 and 2021 (Kalyan et al., 2022). With the recent trend of scaling up pre-training data and model parameters, multiple studies have delved into medical LLMs, transitioning to decoder-only architectures and generative tasks (He et al., 2023; Zhou et al., 2023). Popular open-source milestones include ChatDoctor (Li et al., 2023), PMC-LLaMA (Wu et al., 2023), and MEDITRON (Chen et al., 2023). "Pre-train, prompt, and predict" is today's prevalent methodology for applying LLMs to new problems, circumventing the need for extensive fine-tuning on human-labeled examples, which are rarely accessible in clinical scenarios. We are the first to examine the prowess of medical LLMs in generating background context for ODQA without supervision.

Open-Book Question Answering Closed-book LLMs, such as GPT-IV and Med-PaLM-2 540B, exhibit impressive results in ODQA with fine-tuning (Singhal et al., 2023) or prompting (Nori et al., 2023a,b) techniques, performing non-trivial reasoning steps (Liévin et al., 2022). As LLMs grow to achieve predictable accuracy gains (Kaplan et al., 2020), their computational demands exceed the resources of most product teams. To counteract the scaling trend, reach satisfactory results with commodity hardware, and ensure greater control over knowledge, various open-book contributions have surfaced in medicine, comprising BioReader (Frisoni et al., 2022), DRAGON (Yasunaga et al., 2022), and VOD (Liévin et al., 2023). Nevertheless, to date, generate-then-read pipelines have predominantly undergone scrutiny solely within the general domain. GenRead (Yu et al., 2023) and CGAP (Zhang et al., 2023b) are among the first to demonstrate the efficacy of LLMs as strong context generators, focusing on datasets like Natural Questions, TriviaQA, and WebQuestions. They utilized

LLMs such as InstructGPT (175B) and Megatron (530B) both as generators and readers. GenRead introduced a clustering-based prompting technique involving context generation, filtering, encoding, k-means clustering, and per-cluster sampling. While we acknowledge the efficacy of these techniques, as demonstrated in ad-hoc ablations (Appendix D), we caution against their feasibility in real-world scenarios due to associated high costs. Additionally, investigations into the fusion of retrieved and LLM-generated contexts for ODQA, exemplified by COMBO (Su et al., 2023), have shown promising results in enhancing performance across aforementioned benchmarks. However, their approach entails utilizing two discriminators trained on silver labels to compute compatibility scores, with InstructGPT and ChatGPT as context generators and a FiD Flan-T5-large as the reader, scaling up to 10 contexts using four A40 GPUs, each equipped with 46 GB VRAM.

3 Method

In this section, we discuss our MEDGENIE framework, illustrated in Figure 2. First, it prompts a medical LLM to furnish multi-view background contexts for a given question (§3.1). Then, it diverges into two distinct paths depending on the reader type: ICL for LLMs (§3.2), and Fusion-In-Decoder (FID) fine-tuning for SLMs (§3.3). Both strategies equip readers with custom background passages, allowing them to tackle medical questions effectively even without prior knowledge.

Problem Statement In multiple-choice ODQA, a dataset record consists of a question $\mathbf{q}$ and an answer set $\mathcal{A}=\{\mathbf{a}_1, \ldots, \mathbf{a}_{|\mathcal{A}|}\}$, all expressed in text. Our goal is to identify which answer $\mathbf{a} \in \mathcal{A}$ is correct ($\hat{\mathbf{a}}$). A closed-book solution would estimate $p_\theta(\mathbf{a}|\mathbf{q}, \mathcal{A})$ parameterized by θ, leaving considerable domain knowledge unexploited (Levine et al., 2022). Conversely, we assume access to a set of record-specific artificial contexts and add an auxiliary variable $\mathcal{C}=\{\mathbf{c}_1, \ldots, \mathbf{c}_{|\mathcal{C}|}\}$. We thus move to a marginal likelihood with a reader-generator model: $p_\theta(\mathbf{a}|\mathbf{q}, \mathcal{A})=\sum_i p_\theta(\mathbf{a}|\mathcal{C}_i, \mathbf{q}, \mathcal{A}) p_\theta(\mathcal{C}_i|\mathbf{q}, \mathcal{A})$. Since we cannot sum over all possible contexts, we approximate the maximum a posteriori estimation with a decoding strategy, using a single value for $\mathcal{C}$.

3.1 Multi-view artificial contexts

In the first stage, we prompt a medical LLM to output $\mathcal{C}$. The primary challenge in generating multiple synthetic contexts for the same question is to promote diversity and knowledge coverage while avoiding repetitive information. Addressing this concern, Yu et al. (2023) proposed a three-step approach: (i) building a supporting repository with an artificial document for each $\mathbf{q}$ of the training set, (ii) encoding artificial documents and clustering them, (iii) given a training or inference query, sampling various question–document pairs from each cluster, and using them as ICL demonstrations to craft a series of topic-guided contexts. Although we acknowledge that clustering-based prompting can amplify the effectiveness of MEDGENIE (Appendix §D), it introduces complexity, largely impacting time and memory costs. To streamline the process and reduce resource constraints, we devise a multi-view prompt technique, dividing $\mathcal{C}$ into two segments, $\mathcal{C}_a$ and $\mathcal{C}_b$. We first ask the LLM to generate a set of contexts $\{\mathbf{c}_{a,1}, \ldots, \mathbf{c}_{a,l}\}$ conditioned on $\mathbf{q} \cup \mathbf{a} \in \mathcal{A}$ (*option-focused*), and then a set of contexts $\{\mathbf{c}_{b,1}, \ldots, \mathbf{c}_{b,m}\}$ conditioned solely on $\mathbf{q}$ (*option-free*). The former aims to provide information for each candidate answer, whereas the latter targets a broader context, encompassing pertinent domain concepts that could aid reasoning. Figure 3 clarifies their distinction.

We rely on few-shot prompting (Brown et al., 2020) to guide the LLM on two examples $\mathcal{E}_c$ meticulously curated by a healthcare professional in distinct and representative domains:

$$Pr_a = \left[\bigoplus_{i}^{|\mathcal{E}|} [\mathbf{q}_i \oplus \mathcal{A}_i \oplus \mathbf{c}_i] \right] \oplus \mathbf{q} \oplus \mathcal{A} \tag{1}$$

$$Pr_b = \left[\bigoplus_{i}^{|\mathcal{E}_c|} [\mathbf{q}_i \oplus \mathbf{c}_i] \right] \oplus \mathbf{q} \tag{2}$$

where Pr_a is the option-focused prompt, Pr_b the option-free one, and $\oplus$ denotes concatenation. Full prompt templates are listed in Appendix §K.

To promote intra-context diversity, we (i) use a random sampling decoding strategy with high temperature, ensuring a wider exploration of the model's output space, and (ii) penalize new tokens based on their frequency in the prompt and the inferred text. Details in Appendix §A.

According to in-depth ablation studies (Appendix §B.3), we set l=3 and m=2, corroborating the importance of each view. Importantly, we preprocess each context by removing any answer-guessing phrase that could bias the reader.

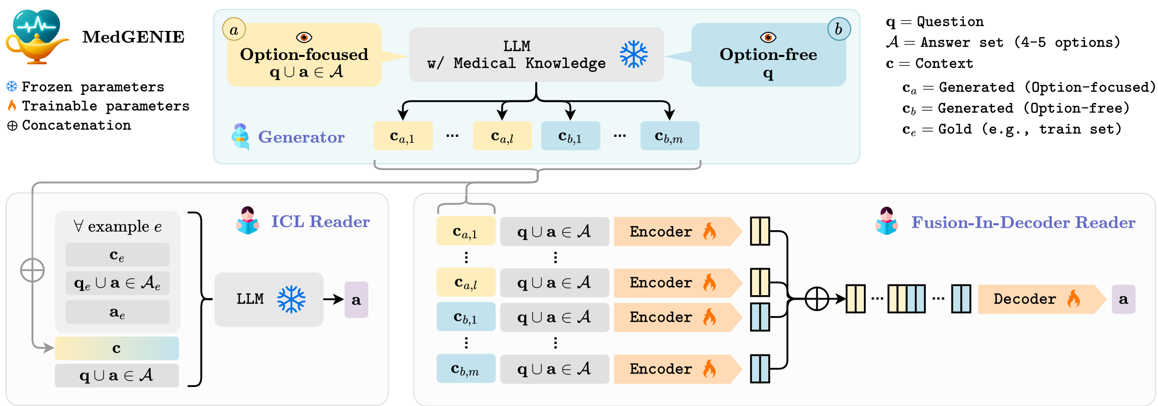


Figure 2: Overview of the MEDGENIE framework. It generates multi-view artificial contexts with a specialized LLM (top), and then uses them to ground a prompted LLM or a fine-tuned SLM (bottom).

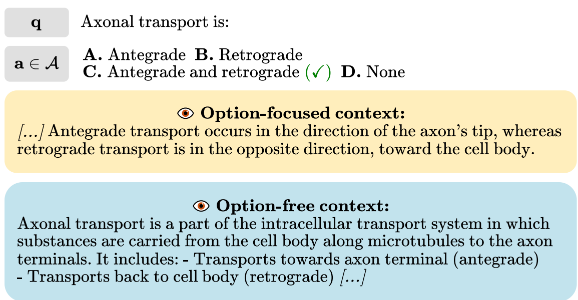


Figure 3: Example of multi-view context generation for a MedMCQA eval instance. The knowledge verbalized by a medical LLM is highly valuable in determining the correct answer (unseen by the generator).

3.2 ICL reader (Unsupervised)

For the unsupervised setting, we use ICL which does not need the complete training data but only a few demonstration examples. In particular, we feed an LLM reader with few-shot ODQA demonstrations and the test query preceded by its artificial context. Inferring the prompt continuation indirectly accomplishes the *unseen* ODQA task. Mechanically, we build the following two-shot prompt:

$$\left[\bigoplus_{i}^{|\mathcal{E}_a|} \mathbf{c}_i \oplus \mathbf{q}_i \oplus \mathcal{A}_i \oplus \hat{\mathbf{a}}_i\right] \oplus \mathbf{c} \oplus \mathbf{q} \oplus \mathcal{A} \quad (3)$$

where $\mathcal{E}_a$ encloses two train instances of the target benchmark, and $\mathbf{c}_i$ is an option-focused context from the medical LLM—proved to be more effective than a human sample (Appendix §C). $\mathbf{c}$ is the concatenation of option-focused and option-free contexts from §3.1. Consistently, we assign the initial positions within $\mathbf{c}$ to option-focused contexts due to their higher probability of containing the correct answer. This conjecture is further validated by the outcomes of our experiments in §5.2. Our design choice wants to prevent $\mathcal{C}_a$ from "getting lost in the middle" (Liu et al., 2023). Indeed, LLMs are prone to difficulty extracting relevant information when it is buried within the central portion of a lengthy context. Contrarily, accuracy improves when pertinent information is situated at either the beginning or end of the input context.

3.3 Fine-tuned reader (Supervised)

While the ICL methodology unlocks the utilization of off-the-shelf LLMs and harnesses their innate language understanding abilities, it mandates loading billions of weights into memory, making inference alone expensive. LLM readers could remarkably benefit from additional fine-tuning steps, but training generally requires up to 20 extra bytes per parameter, easily evolving prohibitive. Therefore, for the supervised regime, we train a lightweight FID reader (Izacard and Grave, 2021). Each $<\mathbf{c}, \mathbf{q}, \mathcal{A}>$ tuple is joined with special separator tokens, and processed independently by the encoder. The representations produced by the last encoder layer are then concatenated and sent to the decoder layers. In this way, the computation cost grows linearly with $|\mathcal{C}|$. We keep the relative order of contexts fixed across all input pairs, always putting $\mathcal{C}_a$ in front of $\mathcal{C}_b$.

4 Experimental Setup

4.1 Benchmarks

We evaluate MEDGENIE on three multiple-choice ODQA benchmarks (Table 1), requesting multi-hop reasoning capabilities, and human-level reading comprehension skills.

MedQA-USMLE (Jin et al., 2020) English-language subset of MedQA tailored for medical license examination in the United States. It amalgamates various knowledge facets, such as patient profiles, disease symptoms, and drug dosage requirements. Each question entails a medical case history, vital signs (e.g., blood pressure, temperature), and eventual diagnostic evaluations (e.g., CT scan). It exists in two versions: 4 and 5 options. Following Chen et al. (2023), in the supervised pipeline, we fine-tune the reader over the original 5-option train set for fair comparisons.

MedMCQA (Pal et al., 2022) Highly comprehensive dataset sourced from Indian medical school entrance exams (AIIMS, NEET-PG). It features mixed question types, such as diagnosis and mathematical problems. Due to the absence of public answer labels for the test set, we employ the eval set as the main reference, adhering to prior literature (Liévin et al., 2022; Singhal et al., 2022; Nori et al., 2023b; Chen et al., 2023; Wu et al., 2023).

MMLU-Medical (Hendrycks et al., 2021) MMLU is a multi-task zero-shot benchmark suite of 57 subsets spanning STEM, humanities, and social sciences. In the footsteps of Singhal et al. (2022); Liévin et al. (2023); Chen et al. (2023), we limit our analysis to 9 medically relevant subsets. As MMLU lacks training data, we fine-tune the FID reader over the MedMCQA train set and gauge out-of-domain generalization, securing fairness with Liévin et al. (2023); Chen et al. (2023).

Terminology. Throughout the remainder of this paper, we use the terms "MedQA" and "MMLU" to denote their medical subsets, aiming for succinctness. With regard to MedQA, unless otherwise specified, we refer to the common 4-option variant.

	MedQA	MedMCQA	MMLU
# Train	10,178	187,000	–
# Eval	11,450	4183	–
# Test	1273	–	1862
$\|\mathcal{A}\|$	4-5	4	4
# Subjects	Not labeled	21	9
# Words / q	116.6	12.7	36.2

Table 1: Summary of the medical benchmarks.

4.2 Medical-expert generator

Following exhaustive preliminary experiments (Appendix §B.1), we select PMC-LLaMA-13B as the specialized generator model (Wu et al., 2023).[2] Technically, PMC-LLaMA performs a two-step training above LLaMA (Touvron et al., 2023). First, it infuses medical knowledge by continuing the pre-training on 4.8M academic papers and 30K medical books. Second, it undergoes medical instruction tuning on question answering, rationale, and conversation, also exploiting the MedQA and MedMCQA train sets. Although the model has been exposed to the training data of some benchmarks, it is imperative to note that it was not explicitly trained to forge question-conditioned contexts. Pre-aligning the selected medical expert generator towards the benchmark data is not strictly necessary. The essential aspect is the latent knowledge encoded within the model parameters, which is frequently correlated with their quantity (Tirumala et al., 2022; Carlini et al., 2023). Nori et al. (2023b) operated CoT and self-consistency prompting in conjunction with major voting to recall knowledge within generalist foundational models and surpass domain-specialized counterparts. However, these methodologies demonstrate efficacy when applied to the largest versions of a foundational model (Singhal et al., 2022; Liévin et al., 2022; Chen et al., 2023), preferring domain-specific solutions at smaller magnitudes. In light of this, we opt for the largest medical LLM runnable on consumer GPU hardware. The average length of a single PMC-LLaMA context is ≈200 words; see Appendix §B.2 for distributional statistics.

4.3 Readers

MEDGENIE is *reader-agnostic* but, for this study, we adopt the following implementation choices.

- **ICL**. We test LLaMA-2-chat (7B) (Touvron et al., 2023)[3] and Zephyr-β (7B) (Tunstall et al., 2023),[4] both supporting a context window length of 4K tokens. Additionally, we examine more recent models with increased context windows: LLaMA-3-instruct (8B)[5] with 8K tokens and Phi-3-mini-instruct (3.8B) with 128K tokens.[6]
- **FID**. We train a FID Flan-T5-base (250M) (Chung et al., 2022).[7]

[2] axiong/PMC_LLaMA_13B
[3] meta-llama/Llama-2-7b-chat-hf
[4] HuggingFaceH4/zephyr-7b-beta
[5] meta-llama/Meta-Llama-3-8B-Instruct
[6] microsoft/Phi-3-mini-128k-instruct
[7] google/flan-t5-base

This article has evolved over two distinct versions, each exploring different LLM ICL readers:

v1 - February 2024: Experiments and ablation studies conducted utilizing LLaMA-2 and Zephyr.
v2 - June 2024: Integration of the newly released LLaMA-3 and Phi-3-mini in the core experiments, with up-to-date baselines, namely BioMistral, MeditronLlama-Slerp, LLaMA-3-Meditron, and OpenBioLLM.

4.4 Evaluation

Multiple Choice Prompting and Direct Completion In alignment with recent literature (Nori et al., 2023b; Chen et al., 2023; Liévin et al., 2022; Singhal et al., 2022), as anticipated in §3, we utilize multiple-choice prompting (MCP). In MCP, the model is presented with both $\mathbf{q}$ and $\mathcal{A}$, where each option is bound to a proxy symbol (A, B, C, D, E), and the model only has to predict a single token.[8] This diverges from earlier cloze prompting (CP) strategies (Brown et al., 2020; Lieber et al., 2021; Du et al., 2022; Smith et al., 2022; Chowdhery et al., 2023), where only $\mathbf{q}$ is passed to the model and the candidate answers are scored independently. Unlike CP, MCP (i) does not conflate the likelihood of an answer with the likelihood of its text, (ii) does not require normalization procedures, (iii) explicitly contrasts different answer options, and (iv) only requires a single forward pass; see (Robinson and Wingate, 2023). Since we focus on evaluating the pure impact of artificial contexts, we operate under the strict assumption of *directly* inferring the correct answer with this additional input signal. We do not resort to alternative schemas, such as CoT or self-consistency CoT, beyond the scope of our research work. For reproducibility, we adopt greedy decoding as in Chen et al. (2023).

Metrics We use Accuracy (% of correctly solved questions) as our main metric across all benchmarks. We resort to Recall@K for measuring the % of retrieved generated contexts rather than gold-chunked ones after top-K retrieval and reranking.

4.5 Baselines

We juxtapose MEDGENIE with two categories of models. *Closed-book.* We prioritize medium-sized LLM baselines ($\leq$8B), either pre-trained or fine-tuned on medical data, fostering an equitable and resource-aware comparison with our models. *Open-book.* We incorporate all the contributions to our knowledge, going beyond text-only grounding.

We further incorporate baselines implemented and run by us. For the *closed-book* category, we conduct few-shot inferences utilizing up-to-date generalist LLM backbones. For *open-book*, we explore augmenting our backbone models with MedWiki (Liévin et al., 2023), a collection of 293,593 medical articles from English Wikipedia, built by querying each option of MedMCQA and MedQA against the Wikipedia API.

5 Results

Our core findings are delineated in Table 2. For space reasons, we report key baselines with available results for each benchmark. To ensure complete and impartial evaluations, we also document current state-of-the-art (SOTA) models outside the intersection. We refer the reader to Appendix §E for full benchmark-specific baselines.

5.1 Artificial grounding effect

We direct our attention to the accuracy gap between MEDGENIE models and their non-grounded version (cf. teal-colored scores in Table 2).

Impact on ICL Reader Our framework significantly enhances the performance of all the considered reader models. Specifically, LLaMA-2-chat, Zephyr-β, and Phi-3-mini show average improvements of +11.7, +7.8, and +3.6 points, respectively. The impact on MedQA is particularly notable, with improvements of +15.7, +9.8, and +9.6 points, respectively. Artificial grounding further elevates Phi-3-mini, establishing a new SOTA in MedQA and delivering the best overall performance. The combination of MEDGENIE with the strong reasoning capabilities and extended context window of Phi-3 fully unlocks the potential of this lightweight model. The results for LLaMA-2-chat and Zephyr-β indicate that the advantages of MEDGENIE are particularly prominent in models without prior domain-specific knowledge. In contrast, newer models, such as LLaMA-3, already demonstrate a robust medical background, likely due to their extensive pre-training phase over 15 trillion tokens. As a result, the impact of MEDGENIE is less significant. Few-shot models struggle against fine-tuned versions on MedMCQA, highlighting the necessity of dedicated training for achieving high scores. Surprisingly, the ICL MEDGENIE models excel in both MedQA and MMLU, outperform-

[8] The prediction may be longer (e.g., "(A) Vitamin B12"), but only the mentioned symbol is considered as the answer.

Model	Ground (Source)[*]	Learning	Params	MedQA	MedMCQA	MMLU	AVG (↓)
LLaMA-3-Instruct	∅	1-shot	8B	60.6	55.7	69.8	62.0
Phi-3-mini	∅	1-shot	3.8B	55.1	53.5	70.3	59.6
MEDITRON[1]	∅	Fine-tuned	7B	52.0	**59.2**	55.6	55.6
PMC-LLaMA[1]	∅	Fine-tuned	7B	49.2	51.4	59.7	53.4
LLaMA-2[1]	∅	Fine-tuned	7B	49.6	54.4	56.3	53.4
Zephyr-β	∅	2-shot	7B	49.3	43.4	60.7	51.1
Mistral-Instruct[1]	∅	3-shot	7B	41.1	40.2	55.8	45.7
LLaMA-2-chat	∅	2-shot	7B	36.9	35.0	49.3	40.4
Codex[2]	∅	0-shot	175B	52.5	50.9	–	–
MedGENIE-Phi-3-mini	G (PMC-LLaMA)	1-shot	3.8B	**64.7** (+9.6)	54.1 (+0.6)	**70.8** (+0.5)	**63.2** (+3.6)
MedGENIE-LLaMA-3-Instruct	G (PMC-LLaMA)	1-shot	8B	<u>63.1</u> (+2.5)	56.2 (+0.5)	<u>68.9</u> (-0.9)	<u>62.7</u> (+0.7)
MedGENIE-Zephyr-β	G (PMC-LLaMA)	2-shot	7B	59.7 (+10.4)	51.0 (+7.6)	66.1 (+5.4)	58.9 (+7.8)
MedGENIE-FID-Flan-T5	G (PMC-LLaMA)	Fine-tuned	250M	53.1	52.1	59.9	55.0
Zephyr-β	R (MedWiki)	2-shot	7B	50.5	47.0	**66.9**	54.8
VOD[3]	R (MedWiki)	Fine-tuned	220M	45.8	<u>58.3</u>	56.8	53.6
MedGENIE-LLaMA-2-chat	G (PMC-LLaMA)	2-shot	7B	52.6 (+15.7)	44.8 (+9.8)	58.8 (+9.5)	52.1 (+11.7)
Mistral-Instruct	R (MedWiki)	2-shot	7B	45.1	44.3	58.5	49.3
LLaMA-2-chat	R (MedWiki)	2-shot	7B	37.2	37.2	52.0	42.1
Human (passing score)				*60.0*	*50.0*	–	
Human (expert score)				*87.0*	*90.0*	*89.8*	

[1] (Chen et al., 2023) [2] (Liévin et al., 2022) [3] (Liévin et al., 2023)

[*] Whether the input is augmented with external knowledge to ground model predictions; ∅ = No Grounding, R = Retrieved, G = Generated. While external knowledge sources for RAG (i.e., Ground=R) are the same for each benchmark, the contexts generated by an LLM (i.e., Ground=G) are specific for the questions of each benchmark.

Note: The influence (i.e., Δ Acc.) of grounding LLMs with knowledge generated under the MedGENIE framework is highlighted with colors.

Table 2: Main accuracy results of MEDGENIE (ICL with LLaMA-2-chat or Zephyr-β, and FID with Flan-T5) against open-source closed-book and open-book baselines. Evaluation on MedQA (test), MedMCQA (eval), and MMLU (test) benchmarks. Bold and underline denote the best and second-best scores.

ing fully fine-tuned LLMs. Notably, MedGENIE-LLaMA-2-chat surpasses the fine-tuned LLaMA-2-chat by 3 points in MedQA and 2.5 points in MedMCQA. Similarly, MedGENIE-Zephyr-β outperforms the strongest fine-tuned alternative, MEDITRON, by +7.7 points in MedQA and +10.5 points in MMLU. Our approach eliminates the need for expensive full fine-tuning of 7B models, which requires 56GB VRAM when using a standard AdamW optimizer. Instead, ICL MEDGENIE operates efficiently on a standard consumer GPU with 14GB VRAM in half precision.

Impact on FID Reader MedGenie-FID-Flan-T5 achieves 53.1% accuracy on MedQA, with outstanding efficiency. It outshines 7B baselines fine-tuned on the train set of target benchmarks, such as MEDITRON (+1.1), PMC-LLaMA (+3.9), LLaMA-2-chat (+3.5), and even CODEX 175B in zero-shot settings (+0.6)—leveraging up to 706× fewer parameters. In MedMCQA, it significantly exceeds ICL models, such as Zephyr-β (+8.7), Mistral-Instruct (+11.9), and LLaMA-2-chat (+17.1). It remains superior to fully fine-tuned PMC-LLaMA (+0.7) and the zero-shot CODEX 175B model (+1.2), but struggles to compete with previous SOTA holders: MEDITRON (-7.1) and VOD (-6.2). We motivate this gap with the huge difference in size and pre-knowledge, other than the more notional nature of the MedMCQA questions, where MedWiki grounding is highly effective. In MMLU, it again shows exceptional performance, outperforming all fine-tuned baselines on the MedMCQA train set. This highlights the generalization power of MEDGENIE compared to larger and more costly LLMs. Zooming out, our findings corroborate the *tiny titan* nature of Flan-T5 observed by other researchers (Fu et al., 2024).

5.2 Generated vs retrieved context

MEDGENIE tops all prior retrieve-then-read solutions across all benchmarks except for MedMCQA, where it is only rivalled by VOD. Generated contexts yield higher accuracy compared to Wikipedia and textbook chunks, as well as UMLS subgraphs. Impressively, artificial augmentation applied to LLaMA-2-chat results in +10 points (avg) compared to MedWiki augmentation.

Reranking Preference A crucial aspect of our framework is to ensure top-tier contexts. However, the use of a 13B model as the context generator raises concerns about its ability to deliver adequate quality contexts, especially compared to stronger closed and open-source alternatives such as GPT-4 and LLaMA-70B. These concerns arise from the risk of hallucinations or less accurate text compared to factual information stored in knowledge

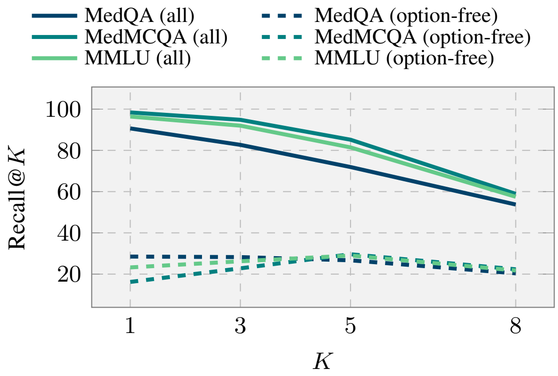


Figure 4: Percentage of multi-view generated contexts compared to MedWiki-retrieved contexts in the top-K positions of a BGE-large reranker.

bases. Human evaluation is worth exploring, but it would require considerable resources and expertise, which may not be readily accessible in our case. Hence, we propose using a SOTA reranker, BGE-large (Xiao et al., 2023),[9] to determine whether our artificial contexts are deemed more relevant than the top contexts (chunks) fetched from MedWiki. Given a question $\mathbf{q}$, we take the 5 multi-view contexts $\mathcal{C}$ generated by PMC-LLaMA. Next, we retrieve the 10 MedWiki contexts most similar to $\mathbf{q}$ (PubMedBERT (Gu et al., 2022), cosine similarity). Subsequently, we create 15 $\mathbf{q}$–context pairs to feed the reranker, registering their relevance score. Finally, we apply Recall@K to quantify the % of generated (relevant) contexts in the top-K positions according to the reranker. Figure 4 portrays the results. The reranker strongly favors our artificial contexts over the retrieved ones across all three benchmarks. Precisely, we achieve a Recall@1 of 91%, 98%, and 96% on MedQA, MedMCQA, and MMLU, respectively. Option-focused contexts contribute predominantly to this preference. However, as K rises, our recall begins to diminish, and the significance of option-free contexts amplifies, thus advocating for our multi-view approach.

Artificial Contexts for Retrieval Augmented Generation Our analysis delves deeper into assessing the efficacy of synthetic contexts by implementing them within a standard retrieve-then-read pipeline. The objective is to discern whether their combination with factual data can improve overall performance. We analyze two different settings. First, we supplement MedWiki with artificial contexts from the MedQA test set and the MedMCQA eval set. We assess whether a small portion of synthetic data may impact a much larger body of factual data. Second, we broaden the knowledge base to cover contexts generated from the MedQA and MedMCQA train set questions. Although such contexts may not pertain to the benchmark test sets, they might offer additional insights to improve accuracy. For each setting, we evaluate retrieval-augmented ICL readers on all three benchmarks. We refrain from considering MMLU contexts to ensure a complete *out-of-domain* validation, enabling us to understand the validity of artificial passages when obtained from a different domain. The results are pointed out in Table 3. Mixed generation+retrieval knowledge sources consistently improve accuracy on all benchmarks, except Mistral-Instruct on MMLU. Combined with Zephyr-β, we obtain 67.4, pushing the SOTA on MMLU. Despite these positive gains, scores fall short in competing with the generate-then-read paradigm of MEDGENIE. We attribute this discrepancy to the retriever influence and the inevitable degradation deriving from segmentation.

KB[*]	**G**[*]	**llama2**[1]	**mistral**[2]	**zephyr**[3]
		MedQA		
4.5M	–	37.2	45.1	50.4
4.5M	96K[†]	41.4 (+4.2)	45.6 (+0.5)	50.8 (+0.4)
4.5M	2M[‡]	40.8 (+3.6)	45.9 (+0.8)	51.2 (+0.8)
		MedMCQA		
4.5M	–	37.3	44.3	47.1
4.5M	96K[†]	41.8 (+4.5)	48.1 (+3.8)	50.3 (+3.2)
4.5M	2M[‡]	43.7 (+6.4)	49.8 (+5.5)	50.9 (+3.8)
		MMLU (out-of-domain)		
4.5M	–	52.0	58.5	66.9
4.5M	96K[†]	53.5 (+1.5)	58.3	67.4 (+0.5)
4.5M	2M[‡]	53.0 (+1.0)	58.4	67.0 (+0.1)

[1] LLaMA-2-chat (7B) [2] Mistral-Instruct (7B)
[3] Zephyr-β (7B) [†] MedQA+MedMCQA train set
[‡] MedQA+MedMCQA train+test set
[*] Number of chunks of MedWiki (KB) and the generated sources (G). VectorDB details in Appendix §G.
Note: The positive influence (i.e., Δ Acc.) of retrieving from artificial contexts is highlighted with colors.

Table 3: Accuracy on MedQA (test), MedMCQA (eval), and MMLU (test) with a retrieve-then-read pipeline based on ICL and MedWiki as external knowledge base, progressively enriched with artificial contexts.

RAGAS Evaluation When tasked with generating extensive medical content, the context genera-

[9] BAAI/bge-reranker-large

tor may produce significant hallucinations. Demonstrating a preference for generated contexts over those retrieved from MedWiki might not be sufficient, as the preference could be attributed to coherence rather than factual accuracy. To validate our findings further, we evaluate both generated and retrieved contexts using the RAGAS library (Es et al., 2024). RAGAS offers various reference-free metrics for RAG pipelines, following the LLM-as-a-judge paradigm. Particularly we focus on:

- **Context Recall**: Measure the extent to which the retrieved/generated context aligns with the ground-truth (GT) answer. Each sentence in the GT answer (generally one) is analyzed to determine whether it can be attributed to the retrieved context or not.

$$CR = \frac{|\text{GT sentences linked to the context}|}{|\text{GT sentences}|} \quad (4)$$

- **Context Precision**: For each chunk in retrieved/generated context, check if it is relevant or not relevant to arrive at the GT answer for the given question. Ideally, all relevant chunks should appear in the highest ranks.

$$CP@K = \frac{\sum_{k=1}^{K} P@k \cdot v_k}{|\text{Relevant chunks in the top-}K|} \quad (5)$$

$$P@k = \frac{\text{true positives}@k}{(\text{true positives}@k + \text{false positives}@k)} \quad (6)$$

where K is the total number of chunks in the context and $v_k \in \{0,1\}$ is the relevance indicator at rank k.

- **Faithfulness**: The factual consistency of the generated answer is evaluated by comparing its claims against the provided context to determine if they can be accurately inferred.

$$F = \frac{|\text{Generated answer claims implied by the context}|}{|\text{Generated answer claims}|} \quad (7)$$

Each metric gives score in a $[0,1]$ boundary; the higher, the better. We utilize `gpt-4-turbo-2024-04-09`. Due to the high API costs, we limit our evaluation to a sample of MedQA, the most relevant testbed in our study. Table 4 reports the performance of Zephyr-β on 150 randomly selected instances, where both the generated and retrieved contexts direct the LLM to produce correct answers. We also apply RAGAS to 50 random instances where both sources lead to wrong answers. The results reaffirm the superiority of artificial contexts, achieving up to +27.2 CR, +39.3 CP, +35.9 F.

	Metric	# Samples	Answer	G	R
📐	CR	150	*correct*	93.4	76.2
	CP	150	*correct*	87.9	48.6
	F	150	*correct*	59.7	23.8
	CR	50	*wrong*	59.2	32.0
	CP	50	*wrong*	55.3	29.5

Table 4: RAGAS evaluation on MedQA. For transparency, scores are linked to the LangSmith pages providing detailed run information, including selected instances, prompts, predictions, and inference time.

6 Conclusion

This paper introduces MEDGENIE, a novel fully-generative framework for medical ODQA, with a particular emphasis on resource-constrained environments. Through comprehensive experimentation with three standard medical benchmarks, MEDGENIE demonstrates substantial performance improvements over the existing closed-book and open-book methodologies. By injecting multi-view contexts from a medical LLM via ICL or lightweight fine-tuning, MEDGENIE achieves new SOTA results, even surpassing expensive fine-tuned LLMs or the largest zero-shot models with up to 706× fewer parameters. Furthermore, our research highlights the effectiveness of artificial passages in enhancing *retrieve-then-read* workflows, showcasing the potential of generated contexts to surpass or assist retrieved counterparts. From a wider angle, MEDGENIE represents a promising approach to address the intricate challenges of medical ODQA, laying the foundation for future advancements in the field.

Ethical Statement

We honor and support the ACL Code of Ethics. LLMs offer significant social benefits, but also pose potential risks. Safety and trustability are paramount in the medical domain. This paper uses knowledge embedded within LLM parameters to tackle open-domain questions. Unlike retrieve-then-read methods reliant on curated external corpora, using models to generate contextual documents may inadvertently amplify inherent biases and deviate from clinical and societal norms, potentially leading to the dissemination of convincing medical misinformation. Therefore, we advocate for a cautious approach, recommending manual scrutiny of any output by domain experts before real-world utilization. This ethical precaution is vital to avoid disseminating potentially erroneous or misleading information, especially in the clinical and scientific communities.

Limitations and Future Work

Despite achieving SOTA performance on the benchmarks presented, our MedGENIE pipeline has several limitations that warrant attention. By replacing the retrieval component with a generative one, we lose the ability to update the knowledge state, which remains frozen to that of the LLM training time. This stands in contrast to the standard retrieve-then-read approach, which allows for the incorporation of new documents as new information becomes available. In a rapidly evolving field like medicine, this ability to quickly adapt by adding temporally recent documents or documents from new domains is important to cope with scientific information overload (Landhuis, 2016). In our method, the responsibility of retaining all knowledge rests entirely on the LLM, and incorporating new knowledge would probably require retraining the context generator.

Our method generates context for any given question, even when the medical LLM lacks knowledge. This may produce a noisy context with inaccurate or irrelevant information. We retain all retrieved contexts without relevance filtering, a strategy that could notably enhance overall performance if implemented. Moreover, the efficacy of our ICL pipeline hinges on the reader's ability to process long prompts within the context window, which may not always be the case for every LLM. This limitation can impact results, particularly when dealing with complex or verbose contexts.

Recent literature suggests solutions for these limitations. Zhang et al. (2024) introduced Retrieval Augmented Fine Tuning (RAFT), a training approach that enhances the model's ability to answer questions in an *open-book* in-domain setting. RAFT trains the model to disregard irrelevant documents when given a question and a set of retrieved documents. Labruna et al. (2024) propose ADAPT-LLM, teaching an LLM to generate a <RET> token when it does not know the answer, triggering retrieval only when necessary. Interestingly, our work suggests potential improvements for both approaches. For RAFT, it might be interesting to explore whether fine-tuning can benefit from augmentation with artificially generated contexts. For the ADAPT-LLM approach, introducing a second token, <GEN>, could be valuable when the retrieved context does not contain the answer, prompting the generation of a context from an auxiliary LLM.

Previous work, such as GenRead (Yu et al., 2023), has explored the zero-shot pipeline using the same model for both generator and reader, following a self-distillation approach. This suggests that a model may indeed benefit from directly extracting knowledge embedded within its own parameters. We acknowledge the significance of assessing the performance of our generator, PMC-LLama-13B (awq), in the role of a reader. Nonetheless, we chose not to pursue this evaluation due to the model's restricted context window of merely 2,000 tokens. Expanding the input prompt with artificial contexts would rapidly exhaust this limited context capacity. However, with recent advancements in LLMs that possess a stronger medical background and an extended context window of 8K tokens (e.g., LLaMA 3 [10], OpenBioLLM [11]), this approach warrants further exploration as future work.

Finally, evaluating the effectiveness of multiple-round reasoning coupled with context generation, as opposed to other methodologies such as multiple-hop reasoning with retrieval (e.g., DSP Khattab et al. (2022)), could provide valuable insights into optimizing QA systems for the complexities of medical domain tasks.

Acknowledgements

This research is partially supported by (i) the Complementary National Plan PNC-I.1 "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, DARE—DigitAl lifelong pRevEntion initiative, code PNC0000002, CUP B53C22006450001, (ii) the PNRR—M4C2—Investment 1.3, Extended Partnership PE00000013, FAIR—Future Artificial Intelligence Research, Spoke 8 "Pervasive AI," funded by the European Commission under the NextGeneration EU program, (iii) the European Commission and the Italian MIMIT through the Chips JU TRISTAN project (G.A. 101095947). We thank Andrea Zecca, Antonio Lopez, Matteo Vannucchi, and Stefano Colamonaco for their valuable support in ablation studies with retrieve-then-read pipelines.

[10] https://www.meditron.io/

[11] aaditya/Llama3-OpenBioLLM-8B

References

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: scaling medical pretraining for large language models. *CoRR*, abs/2311.16079.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen S. Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V. Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569. PMLR.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. 2022. BioReader: a retrieval-enhanced text-to-text transformer for biomedical literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5770–5793, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xue-Yong Fu, Md. Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Bhushan TN.

2024. Tiny titans: Can smaller large language models punch above their weight in the real world for meeting summarization? *CoRR*, abs/2402.00841.

Yukang Gan, Yixiao Ge, Chang Zhou, Shupeng Su, Zhouchuan Xu, Xuyuan Xu, Quanchao Hui, Xiang Chen, Yexin Wang, and Ying Shan. 2023. Binary embedding-based retrieval at tencent. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 4056–4067. ACM.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's mergekit: A toolkit for merging large language models. *CoRR*, abs/2403.13257.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.*, 3(1):2:1–2:23.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *CoRR*, abs/2310.05694.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Kung-Hsiang Huang, Siffi Singh, Xiaofei Ma, Wei Xiao, Feng Nan, Nicholas Dingwall, William Yang Wang, and Kathleen McKeown. 2023. SWING: Balancing coverage and faithfulness for dialogue summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 512–525, Dubrovnik, Croatia. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. AMMU: A survey of transformer-based biomedical pretrained language models. *J. Biomed. Informatics*, 126:103982.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *CoRR*, abs/2212.14024.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *CoRR*, abs/2402.10373.

Tiziano Labruna, Jon Ander Campos, and Gorka Azkune. 2024. When to retrieve: Teaching llms to utilize information retrieval effectively. *CoRR*, abs/2404.19705.

Esther Landhuis. 2016. Scientific literature: Information overload. *Nature*, 535(7612):457–458.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Yoav Levine, Itay Dalmedigos, Ori Ram, Yoel Zeldes, Daniel Jannai, Dor Muhlgay, Yoni Osin, Opher Lieber, Barak Lenz, Shai Shalev-Shwartz, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2022. Standing on the shoulders of giant frozen language models. *CoRR*, abs/2204.10019.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *CoRR*, abs/2303.14070.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. Technical report, AI21 Labs.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *CoRR*, abs/2207.08143.

Valentin Liévin, Andreas Geert Motzfeldt, Ida Riis Jensen, and Ole Winther. 2023. Variational open-domain question answering. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 20950–20977. PMLR.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2023. AWQ: activation-aware weight quantization for LLM compression and acceleration. *CoRR*, abs/2306.00978.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *CoRR*, abs/2307.03172.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *CoRR*, abs/2308.09442.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of GPT-4 on medical challenge problems. *CoRR*, abs/2303.13375.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolò 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. 2023b. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *CoRR*, abs/2311.16452.

Philhoon Oh and James Thorne. 2023. Detrimental contexts in open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11589–11605, Singapore. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138.

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 Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *CoRR*, abs/2305.09617.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zheng, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using deepspeed and megatron to train megatron-turing NLG 530b, A large-scale generative language model. *CoRR*, abs/2201.11990.

Dan Su, Mostofa Patwary, Shrimai Prabhumoye, Peng Xu, Ryan Prenger, Mohammad Shoeybi, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2023. Context generation improves open domain question answering. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 793–808, Dubrovnik, Croatia. Association for Computational Linguistics.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *CoRR*, abs/2211.09085.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022,* *NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of LM alignment. *CoRR*, abs/2310.16944.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2023. Augmenting black-box llms with medical textbooks for clinical question answering. *CoRR*, abs/2309.02233.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *CoRR*, abs/2304.14454.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *CoRR*, abs/2309.07597.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023a. A survey for efficient open domain question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14447–14465, Toronto, Canada. Association for Computational Linguistics.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. RAFT: adapting language model to domain specific RAG. *CoRR*, abs/2403.10131.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models for question answering. *CoRR*, abs/2201.08860.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023b. Merging generated and retrieved knowledge for open-domain QA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4710–4728, Singapore. Association for Computational Linguistics.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, Zheng Li, and Fenglin Liu. 2023. A survey of large language models in medicine: Progress, application, and challenge. *CoRR*, abs/2311.05112.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *CoRR*, abs/2101.00774.

A Implementation Details

Hardware Configuration We ran each experiment on an in-house workstation having one Nvidia GeForce RTX3090 GPU with 24GB of dedicated memory, 64GB of RAM, and an Intel® Core™ i9-10900X1080 CPU @ 3.70GHz.

Checkpoints All the models trained are open source and have permissive licenses for this study: Apache-2.0 (PMC-LLaMA, Mistral-Instruct, Flan-T5), MIT (Zephyr-β), Meta License (LLaMA-2). We initialized them with the corresponding pre-trained versions available in the HuggingFace Transformers library.[12]

Medical-Expert Generator To accommodate memory limitations, we applied post-training Activation-aware Weight Quantization (AWQ) (Lin et al., 2023) in 4 bits to PMC-LLaMA 13B.

Inference For context generation and ICL experiments, we utilized the vLLM library.[13] Since vLLM does not currently support inference for sequence-to-sequence models, we employed the Transformer library for the FID Flan-t5-base. Specifically, generating $\mathcal{C}$, we set a random decoding strategy with a temperature of 0.9, a frequency penalty of 1.95, and a maximum of 512 new tokens. We returned 3 output contexts for a given option-focused prompt and 2 output contexts for a given option-free prompt. We increased the inference batch size to 5 and generated all contexts in parallel with one inference call to the LLM. Thus, the overall latency remains the same as when using a single context. To ensure reproducibility, during the evaluation of both ICL and supervised readers, we adopted a greedy decoding strategy with the random state set to 0.

FID Training In our implementation, we successfully scaled up to 5 independently encoded question–context pairs on 24GB VRAM. For each pair, we used a maximum context window of 1024 tokens for MedQA and 600 tokens for MedMCQA. During training, for MedQA, we chose a batch size of 1 with 4 accumulation steps, totaling 40,712 training steps, and executed evaluations every 10,178 steps. Similarly, for MedMCQA, we employed a GPU batch size of 2 with 2 accumulation steps, trained over 182,816 steps, and evaluated every 22,852 steps. For both benchmarks, we employed a linear scheduler with a warmup phase comprising 10% of the total training steps. The learning rate was set to 5e-5, using an AdamW optimizer with a weight decay of 0.01.

[12] https://github.com/huggingface/transformers

[13] https://github.com/vllm-project/vllm

B Context Generation Insights

B.1 Human Evaluation of Context Generator

We compare three LLMs as candidate generators, two medical (PMC-LLaMA-13B, BioMedGPT-7B) and one generalist (LLaMA-2-Chat-13B), by enlisting the expertise of five external medical professionals with NLP competencies. We exclude MEDITRON due to the release of pre-trained checkpoints only. Motivated by (Fabbri et al., 2019; Huang et al., 2023; Narayan et al., 2018), we use a direct comparison strategy, proven to be more reliable, sensitive, and less labor-intensive than rating scales. We consider a pool of 100 questions sampled in a balanced manner from MedQA, MedMCQA, and MMLU. For each, experts are presented with contexts inferred by 2 out of 3 LLMs. We generate one context for question, allocating 60% of the evaluation to option-focused and 40% to option-free. In order to have also insights about the quality of the contexts generated, we ask experts to select the best all-around context considering relevance, factuality and succinctness, optionally providing a text comment; a "tie" is declared if an expert perceives the two contexts as of equal quality. This approach inherently assesses context quality, with high-quality contexts being those that fully satisfy our evaluation criteria. Finally, the total number of mandatory preference labels per participant is 300. We randomize the order of pairs and context sources per question to guard the rating against being gamed. The final score of each LLM is the percentage of times that its contexts are selected as the best, minus the percentage of times they are not. The results indicate PMC-LLaMA as the best choice, garnering 41% preference, surpassing LLaMA-2-Chat-13B with 36% and BioMedGPT-7B with 33%. The average Kendall's τ coefficients among all inter-annotator agreements is 0.72. Table 5 provides an annotation example.

B.2 Artificial Context Distributional Statistics

Figure 5 plots the length distribution of the artificial contexts used in this study, focusing on the test sets of our benchmarks (eval set for MedMCQA). We tokenize contexts into words using NLTK (Bird and Loper, 2004). In summary, the average number

MedMCQA[84d89a4f-db04-402d-88ab-ca39f6eee8bd]	
Question	Respiratory rhythm generation center is located at: A. Dorsal respiratory group **B. Pre-Botzinger complex** C. Ventral respiratory neurons D. Pneumotaxic center
PMC-LLaMA ***(C1)*** Rank: #1	Respiratory rhythm generation occurs at the Pre-Botzinger complex, located on either side of the medulla between the nucleus ambiguus and the lateral reticular nucleus. This complex is primarily involved in inspiration/inhalation. It fires bursts of action potentials at a frequency that oscillates with inspiration, controlling respiratory rate. The dorsal respiratory group (DRG), ventral respiratory neurons (VR), and pneumotaxic center also play a role in regulating breathing but not specifically associated with rhythm generation as Pre-Botzinger complex is responsible for generating inspiratory drive to spinal motor circuits controlling thoracic movements associated with breathing.
BioMedGPT-LM-7B ***(C2)*** Rank: #3	The pre-Botzinger complex is located in the caudal medulla and plays a critical role in the generation of respiratory rhythm. It receives input from the pontine respiratory groups and integrates them into a steady discharge pattern. This discharge pattern is used to control the duration and frequency of the respiratory cycle. The pre-Botzinger complex is a group of cells that reside in the caudal medulla and are responsible for generating the respiratory rhythm. This complex is also known as the p̈re-BotCör p̈re-Botzinger,n̈amed after the German anatomist and physiologist, Botzinger. The pre-BotC receives input from the pneumotaxic center and integrates it with the input from the pontine respiratory groups. The pre-Botzinger complex is a key component of the respiratory rhythm generation center, located in the caudal medulla. The question in the quiz is asking for the location of the pre-Botzinger complex, which is a group of neurons in the caudal medulla that play a critical role in respiration.
LLaMA-2-chat-13B ***(C3)*** Rank: #2	The Pre-Botzinger complex (PBC) is a group of neurons located in the medulla oblongata that play a crucial role in generating the respiratory rhythm. The PBC receives inputs from various sources, including sensory information from the body and inputs from the central nervous system, and integrates this information to generate the rhythmic activity that drives breathing. The dorsal respiratory group (DRG) is a group of neurons located in the dorsal part of the medulla oblongata that are involved in regulating breathing, but they are not the primary source of respiratory rhythm generation. The ventral respiratory neurons are also involved in breathing regulation, but they are not the main source of rhythm generation.The pneumotaxic center is a region in the medulla oblongata that helps regulate breathing by integrating sensory information and inputs from the central nervous system, but it is not the primary source of respiratory rhythm generation.
Expert judge	C1 stands out as the most relevant and concise response, directly stating the location of the Pre-Botzinger complex and contrasting its role with other potential options. This response effectively addresses the question with clarity. C3 offers a comprehensive explanation of the Pre-Botzinger complex's role in respiratory rhythm generation, though it is slightly longer and lacks the direct comparison found in C1. C2, while informative, is less effective in directly addressing the question due to its narrower focus on the Pre-Botzinger complex without explicitly comparing its function with other potential locations. In summary, C1 provides the most direct and succinct response, followed by C3 for its thoroughness, and then C2 for its informative content but lack of direct comparison.

Table 5: Example of medical expert evaluation on the contexts generated by three LLMs.

of words for MedQA-4opt, MedQA-5opt, MedMCQA, and MMLU is 207, 207, 203, and 202, respectively. The maximum number of words is 442, 452, 460, and 453.

B.3 Number and Type of Generated Contexts

Table 6 displays the tangible impact of incorporating option-free contexts alongside option-focused contexts. Consistently, this integration enhances the performance of every model considered across all benchmarks, with the total number of contexts playing an essential role. We observe that given more contexts, our framework generally achieves greater performance. Particularly noteworthy is the experiment conducted on MedQA with the ICL MedGENIE models. Here, we demonstrate that employing 3 option-focused + 2 option-free contexts yields superior accuracy compared to using only 5 option-focused contexts. Specifically, there's an accuracy enhancement of +1.4 and +0.6 for MedGENIE-LLaMA-2-chat and MedGENIE-Zephyr-β, respectively.

C In-Context Learning Sample Selection

Selecting appropriate samples for the prompts is essential to generate high-quality contexts. Recent studies suggest strategies that aim to uncover more effective contextual demonstrations by retrieving examples that closely match the semantics of each input test (Liu et al., 2022). However, we are constrained by the need to first consider each demonstration as a context–question pair, rather than solely focusing on the question. Secondly, we need to ensure that the context is not too long to avoid (i) saturation of the context window when appending the new test input, which is the result of concatenating five other contexts; (ii) the "lost in the middle" phenomenon; (iii) excessive increase in inference costs. As a result, we employ a strategy similar to that pursued by Chen et al. (2023) for

k (↑)	option-focused	option-free	MedQA (5opt)	MedQA	MedMCQA	MMLU
			MEDGENIE-FID-FLAN-T5			
1	✔	✗	40.69	45.8	46.2	51.5
2	✔	✗	42.58	48.0	48.5	56.0
3	✔	✗	44.07	50.8	50.6	58.5
4	✔	✔	44.62	52.1	51.6	60.0
5	✔	✔	45.56	53.1	52.1	59.9
			MEDGENIE-LLAMA-2-CHAT			
1	✔	✗	42.2	48.5	41.9	54.6
2	✔	✗	43.8	50.0	43.3	55.4
3	✔	✗	45.7	51.1	44.1	55.6
4	✔	✔	44.7	51.1	44.3	58.0
5	✔	✗	-	51.2	-	-
5	✔	✔	46.0	52.6	44.8	58.8
			MEDGENIE-ZEPHYR-β			
1	✔	✗	50.2	56.7	48.7	62.9
2	✔	✗	52.2	58.0	49.9	61.3
3	✔	✗	52.5	58.8	50.8	63.7
4	✔	✔	54.6	57.5	50.8	65.3
5	✔	✗	-	59.2	-	-
5	✔	✔	54.9	59.7	51.0	66.1

Table 6: Performance variation across *k* different numbers of contexts during evaluation. For MEDGENIE-FID-FLAN-T5, contexts are intended to be combined with the input question and encoded independently of each other. Conversely, for MEDGENIE-LLAMA 2-CHAT and MEDGENIE-ZEPHYR-β, they are passed within the input prompt as a single concatenated context before appending the test question. The first three contexts are always to be considered only *option-focused*, while the fourth and fifth ones may be *option-free*. Therefore, when $k = 4$, only the fourth context is option-free. When $k = 5$ and *option-free* is flagged, it implies that both the fourth and fifth contexts are *option-free*.

MEDITRON, adapted based on the characteristics of the reader model considered. For models with lower context window capability, such as Zephyr-β and LLaMA-2-chat, we follow these steps:

- We sample three questions from the training set and generate a context with PMC-LLaMA for each of them.
- We ensure accurate filtering of only relevant content from the generated text.
- Additionally, we involve a medical expert to craft pairs of questions with relevant contexts for each benchmark considered.
- Consequently, we end up with four different pairs of shots for each model in each benchmark.
- We consider two shots as in-context demonstrations.

Tested pairs for MedQA and MedMCQA are documented in Table 7 and Table 8, respectively. Additionaly, Table 9 presents how the performances of each ICL MEDGENIE model are affected by each pair, unveiling the accuracy effectiveness of artificial examples.

Finally, For LLMs with higher context window capacity, such as LLaMA-3-instruct and Phi-3-mini-instruct, our approach is slightly different:

- We sample three questions from the training set and generate a context with PMC-LLaMA for each of them.
- We consider only one long-context artificial shot as in-context demonstration.

This refined strategy ensures effective handling of context constraints while generating accurate and informative demonstrations for each model and benchmark scenario.

D Clustering-Based Prompting

The cluster-based methodology by Yu et al. (2023) can be detailed as follows:

1. ask an LLM to generate a background context for each question in the training split (alternatively, we can retrieve a context from a knowledge source), obtaining a set of question–context pair;
2. perform inference with the LLM by leveraging the generated context, and maintain from

	MEDQA	
Context	**Shot 1**	**Shot 2**
Human (H)	CONTEXT: Nitrofurantoin is a commonly used antibiotic for the treatment of uncomplicated urinary tract infections (UTIs) in pregnant women. It is considered safe during pregnancy and is effective against common pathogens causing UTIs. Ampicillin and ceftriaxone are not the first-line choices for treating uncomplicated UTIs, and doxycycline is contraindicated in pregnancy due to potential adverse effects on fetal development. QUESTION: A 23-year-old pregnant woman... Which of the following is the best treatment for this patient? A. Ampicillin, B. Ceftriaxone, C. Doxycycline, **D. Nitrofurantoin**	CONTEXT: Placing the infant in a supine position on a firm mattress while sleeping is the recommended precaution to reduce the risk of sudden infant death syndrome (SIDS). This position helps maintain clear airways and minimizes the risk of suffocation. Avoiding practices such as covering the infant excessively, using devices to maintain sleeping position, and prohibiting pacifier use during sleep are not recommended and may pose additional risks. QUESTION: A 3-month-old baby died suddenly... Which of the following precautions could have prevented the death of the baby? **A. Placing the infant in a supine position on a firm mattress while sleeping**, B. Keeping the infant covered and maintaining a high room temperature, C. Application of a device to maintain the sleeping position, D. Avoiding pacifier use during sleep
Artificial (A1)	CONTEXT: Most outpatient physicians treat asymptomatic bacteriuria with sulfate-based cephalosporins such as nitrofurantoin (100 mg BID for 7 days) or cephalexin (500mg tid for 7 days). Both drugs are considered safe during pregnancy. QUESTION: A 23-year-old pregnant woman... Which of the following is the best treatment for this patient? A. Ampicillin, B. Ceftriaxone, C. Doxycycline, **D. Nitrofurantoin**	CONTEXT: Sudden infant death syndrome (SIDS) is the unexpected, sudden death of a child under one year old. An autopsy does not show an explainable cause of death in cases with SIDS. Placing the child in a supine position on a firm mattress while sleeping decreases the risk of SIDS by preventing potential hazards such as soft bedding material or entrapment risks that could compromise respiration. QUESTION: A 3-month-old baby died suddenly... Which of the following precautions could have prevented the death of the baby? **A. Placing the infant in a supine position on a firm mattress while sleeping**, B. Keeping the infant covered and maintaining a high room temperature, C. Application of a device to maintain the sleeping position, D. Avoiding pacifier use during sleep
Artificial (A2)	CONTEXT: Acute hypoxic respiratory failure in the setting of recent surgery for femur fracture suggests pulmonary embolism as the most likely pathogenesis. The histologic section demonstrates a thromboembolus lodged in the lumen of a pulmonary artery. Thrombotic or embolic phenomenon has occurred which led to sudden cardiac arrest (pulmonary0̆0a0passive congestion, ischemia, and hypertension are unlikely given that no CAD event or myocardial infarction preceded this acute event). QUESTION: A pulmonary autopsy specimen... Which of the following is the most likely pathogenesis for the present findings? **A. Thromboembolism**, B. Pulmonary ischemia, C. Pulmonary hypertension, D. Pulmonary passive congestion	CONTEXT: Women with von Willebrand disease (vWD) often present with menorrhagia and easy bruising. The platelet count is usually normal, but the bleeding time and PTT are prolonged. Hemophilia A, lupus anticoagulant, protein C deficiency, or factor V deficiency would not present with these findings on the PTT test. QUESTION: A 20-year-old woman... Which of the following is the most likely cause of this patient2̆019s symptoms? A. Hemophilia A, B. Lupus anticoagulant, C. Protein C deficiency, **D. Von Willebrand disease**
Artificial (A3)	CONTEXT: The presence of splenomegaly and the finding of immature granulocytic cells in the bone marrow are consistent with this diagnosis. Chronic myeloid leukemia (CML) is characterized by an abnormality involving the ABL1 gene on chromosome 9q, which results in unregulated tyrosine kinase activity. The JAK-STAT pathway, loss of function of the APC gene, altered expression of retinoic acid receptor genes, or induced expression PDGFRA are not associated with CML; these abnormalities can be seen in other types of leukemia or myelodysplastic syndromes (MDS). QUESTION: A 52-year-old... Which of the following mechanisms is most likely responsible for this patient's condition? A. Cytokine-independent activation of the JAK-STAT pathway, B. Loss of function of the APC gene, C. Altered expression of the retinoic acid receptor gene, **D. Unregulated expression of the ABL1 gene**	CONTEXT: Post-translational modifications (PTMs) are covalent modifications to a polypeptide following its synthesis by the ribosome. The chemically-tagged protein mentioned in the question acts as an E3 ubiquitin ligase by catalyzing the attachment of ubiquitin molecules to lysine residues on targeted proteins, marking them for degradation. Glycosylation involves adding sugar molecules; phosphorylation/dephosphorylation adds or removes phosphate groups and carboxylation involves adding carbon dioxide. Ubiquitination modifies a protein through addition of small, globular proteins called ubiquitins through isopeptide bonds. QUESTION: An investigator is studying... Which of the following post-translational modifications has most likely occurred? A. Glycosylation, B. Phosphorylation, **C. Ubiquitination**, D. Carboxylation

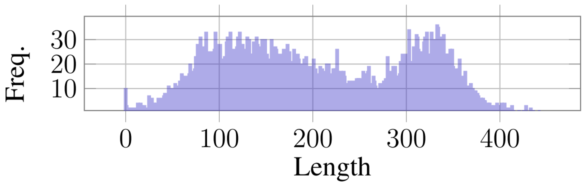
Table 7: Tested 2-shot in-context learning examples for MedQA. The prompts marked with (H) are created by a medical expert, while A* refers to relevant portions of contexts generated by PMC-LLaMA.

MEDMCQA and MMLU		
Context	**Shot 1**	**Shot 2**
Human (H)	CONTEXT: Persistent obstruction of urine flow, as seen in benign prostatic hyperplasia, can result in increased pressure within the urinary system. This elevated pressure may eventually lead to atrophy of the kidney parenchyma. Atrophy refers to the shrinking or reduction in size of an organ or tissue, and in this context, it reflects the consequence of long-term obstruction on the affected kidney. QUESTION: Chronic urethral obstruction due to benign prismatic hyperplasia can lead to the following change in kidney parenchyma A. Hyperplasia, B. Hyperophy, **C. Atrophy**, D. Dyplasia	CONTEXT: Vitamin B12, also known as cobalamin, is primarily found in animal products such as meat, fish, eggs, and dairy. It is not naturally present in significant amounts in plant-based foods, making it essential for individuals following a vegetarian or vegan diet to obtain this vitamin through fortified foods or supplements. In contrast, Vitamin C, Vitamin B7 (biotin), and Vitamin D can be obtained from both animal and plant sources. QUESTION: Which vitamin is supplied from only animal source: **A. Vitamin B12**, B. Vitamin B7, C. Vitamin C, D. Vitamin D
Artificial (A1)	CONTEXT: Hyperviscosity is a condition where the blood becomes abnormally thick, hindering its ability to flow properly. Cryoglobulinemia is a condition characterized by abnormal antibodies in the blood (antibodies are specialized cells that recognize and attack foreign invaders). These abnormal antibodies become solid at cold temperatures and lead to clumping of red blood cells, an increase in viscosity, and subsequent obstruction of small vessels. QUESTION: Hyper viscosity is seen in **A. Cryoglobulinemia**, B. Multiple myeloma, C. MGUS, D. Lymphoma	CONTEXT: Inversion of the foot refers to a foot in which its sole faces medially. Since the plantar surface of the foot is in contact with the ground at all times, this condition occurs when one or more muscles responsible for moving it become tight or weak, resulting in an alteration in normal alignment. QUESTION: In inversion of the foot, the sole will face: A. Upwards, B. Downwards, C. Laterally, **D. Medially**
Artificial (A2)	CONTEXT: Antibiotic prophylaxis for infective endocarditis is indicated in individuals with predisposing cardiac conditions. In this scenario, determining if an isolated secundum ASD and mitral valve prolapse without regurgitation are associated with the potential risk of developing infective endocarditis requires further information. The presence of a prior coronary aery bypass graft and coarctation of aoa are both established indications for antibiotic prophylaxis due to their association with infective endocarditis risk. QUESTION: Antiboiotic Prophylaxis for infective endocarditis is indicated in: A. Isolated secundum ASD, B. Mitral valve prolapse without regurgitation, C. Prior coronary aery bypass graft, **D. Coarctation of aoa**	CONTEXT: The Anterolateral portal is also known as the lateral portal. It is used for viewing the patellofemoral joint, inserting probe or laser for soft-tissue procedures. QUESTION: Anterolateral ahroscopy of knee is for: **A. To see patellofemoral aiculation**, B. To see the posterior cruciate ligament, C. To see the anterior poion of lateral meniscus, D. To see the periphery of the posterior horn of medial meniscus
Artificial (A3)	CONTEXT: Ligamentum teres is a degenerative string of tissue that exists in the fetal remnant of umbilical vein. In adults, it runs along the inferior margin of the liver and functions as both an anatomic landmark and as part of a ligamentous structure that connects the falciform ligament with the round, triangular, and coronary ligaments. QUESTION: Ligament teres is a remnant of ? A. Ductus aeriosus, B. Umbilical aery, **C. Umbilical vein**, D. Ductus venosus	CONTEXT: The Magic syndrome refers to the presence of mouth and genital ulcers. Although initially thought to represent a distinct disease entity, it is now recognized as a subset of Behcet disease. QUESTION: Magic syndrome is seen in: **A. Behcet disease**, B. Aphthous major, C. Herpetiform, D. Bloom syndrome

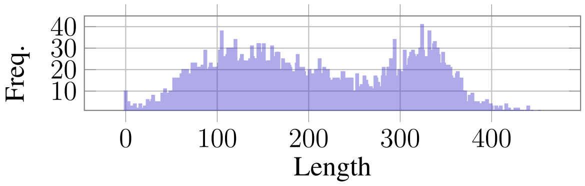
Table 8: Tested 2-shot in-context learning examples for MedMCQA and MMLU. The prompts marked with (H) are created by a medical expert, while A* refers to relevant portions of contexts generated by PMC-LLaMA.

the initial set of pairs only those for which the LLM answered correctly;

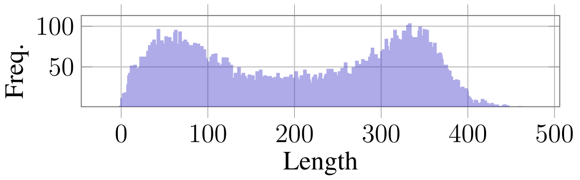
3. encode each question–context pair from the filtered set;
4. use K-means to cluster all the embedding vectors obtained into K classes, where the number of classes matches the number of documents that need to be generated eventually;
5. randomly select n question–context pairs from each cluster;
6. present each different n question–context pairs to the LLM as in-context demonstra-



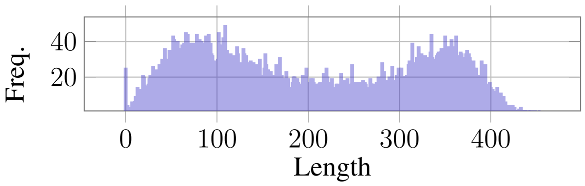
(a) MedQA (4 options), test



(b) MedQA (5 options), test



(c) MedMCQA, eval



(d) MMLU, test

Figure 5: Word-level length distribution of PMC-LLaMA artificial contexts.

Dataset	**H**	**A1**	**A2**	**A3**
MedGENIE-LLaMA-2-chat				
MedQA	52.4	**52.6**	51.5	51.6
MedMCQA	44.1	**44.8**	44.3	44.7
MMLU	56.5	58.2	**58.8**	56.6
MedGENIE-Zephyr-β				
MedQA	**59.7**	58.9	58.8	57.8
MedMCQA	50.5	50.8	**51.0**	50.8
MMLU	65.1	**66.1**	66.0	65.3

Table 9: Effect of each tested pair of shots on the accuracy of each ICL MEDGENIE model. H, A1, A2, A3 refer to the shot examples provided in Table 7 and Table 8 for corresponding benchmarks. Best results are in bold.

tions for generating a context from a given test question.

This approach ensures that the LLM is exposed to different distributions of examples, resulting in generated contexts that cover various (uncontrolled) perspectives. We investigate the potential benefits deriving from the integration of this methodology into MEDGENIE ICL pipeline, tested on MedQA. For point (1), we implement the strategy using both contexts retrieved from MedWiki and contexts artificially generated with PMC-LLaMA. We set K=5 and n=3. The results are summarized in Table 10. Integration leads to a notable improvement of up to +2.4 points in accuracy. However, it comes with a considerably higher cost compared to MEDGENIE alone, primarily due to the additional operations required at points (1) and (2). These operations are unnecessary for MEDGENIE, making it a more cost-effective option, especially in scenarios with larger training sets, such as MedMCQA.

Retrieval source	**Cluster-based**	**LLaMA-2-chat**	**Zephyr-β**
–	✗	52.6	59.7
MedWiki	✓	52.3	59.8 (+0.1)
PMC-LLaMA	✓	**54.3** (+1.7)	**62.1** (+2.4)

Note: The positive influence (i.e., Δ Acc.) of incorporating clustering-based prompting within the MEDGENIE is highlighted using colors.

Table 10: Comparison of clustering-based prompting for context generation with the standard full-generative MEDGENIE ICL pipeline on the MedQA test set.

E Full Baselines

In Table 11, Table 12, and Table 13, we present a fine-grained comparison of MedGENIE models with closed and open-book alternatives known in the literature, both fine-tuned and in zero/few-shot modes. It should be noted that the baseline results provided by Liévin et al. (2022) for MedMCQA should only be regarded as a reference point, as they relate to a subset of 1K samples from the eval set.

F Robustness

F.1 Sensitivity to Question Subject

In Figure 6, we present a detailed analysis of the *per-subject* performance of the MEDGENIE models across MedMCQA (21 medical subjects) and MMLU (9 health and biology subjects). For MedMCQA, the analysis is done in comparison with our baselines: Zephyr-β and LLaMA-2-chat in 2-shot prompting. For MMLU, we extend the analysis to known results from MEDITRON, PMC-LLaMA, and VOD, all fine-tuned over the train set of MedMCQA. In this case, we consider Zephyr-β in 3-shot prompting (Chen et al., 2023) due to its higher accuracy. Please note that the subcategory `professional_medicine` of MMLU corresponds to the MedQA questions.

F.2 Sensitivity to Option Order

Table 14 illustrates the frequencies of the predicted answers after *option shuffling* with 10 distinct random seeds, testing MedGENIE-FID-Flan-T5 in all benchmarks. Note that our seed base is 0. The results highlight a classification bias. In MedQA, the model is inclined to default the last option, discouraging the first, while in MedMCQA, it underestimates D. Similarly, Table 15 examines the robustness of the model to shuffled options with an unsupervised pipeline. Shuffling can impact MedGENIE-Zephyr-β up to 4.4 accuracy points, and MedGENIE-LLaMA-2-chat up to 5.9, registering more oscillations in MMLU.

G Artificial Context Augmentation for RAG

Table 16 shows the detailed results of our experiments aimed at evaluating the effectiveness of combining artificial data with factual data sourced from MedWiki. We compare two approaches: (i) a standard RAG pipeline, where we retrieve the top k=5 relevant contexts for a given query, and (ii) a modified approach where we retrieve the top k=10 most relevant contexts and rerank them to consider only the top-5 reranked contexts.

Both artificial context and MedWiki data are segmented into chunks using the `RecursiveCharacterTextSplitter` with parameters `chunk_size=1000` and `chunk_overlap=200` from LangChain. Each chunk is encoded using `pubmedbert-base-embeddings`.[14] The embeddings of each chunk are stored using Milvus as vectorDB. To determine the similarity between vectors, we employ cosine similarity. Additionally, we use BGE-large[15] as reranker.

H Relation between Context Window and Context Length

To ensure the feasibility and effectiveness of our approach, we implement measures to maintain manageable length of generated contexts. Primarily, our methodology involves crafting optimized prompts and selecting relevant shot examples to guide the LLM, as illustrated in Figure 7 and Figure 8. This approach combines prompt engineering techniques with expert insights from medical professionals. Additionaly, ss detailed in Section A, we establish a maximum generative hyperparameter of 512 new tokens, resulting in a maximum context length of 2560 tokens when concatenated with four other contexts. This deliberate design choice ensures compatibility with the 4096-token context window of both LLaMA-2 and Zephyr-β models. By avoiding longer context lengths, we prevent exceeding the maximum context length in a significant proportion of test samples, particularly considering the question lengths in benchmarks like MedQA. By imposing these constraints, we guarantee that our models can effectively process and generate responses within the specified bounds.

I MedWiki as Knowledge Base for RAG

While we acknowledge that PubMed [16] and similar sources offer a more comprehensive repository of medical knowledge, it is important to consider practical constraints, especially in low-resource settings. Hosting such extensive datasets can be prohibitive in terms of infrastructure and cost. While exploring API services is possible, they would lack the control and flexibility needed for an optimized retrieval pipeline. It is worth noting that previous work, notably by Liévin et al. (2023), has demonstrated the quality and effectiveness of MedWiki as a retrieval corpus. As Table 12 shows, MedWiki-augmented VOD achieves SOTA accuracy scores for openbook models on MedMCQA, largely surpassing other small pre-trained language models like PubmedBERT, SciBERT, and BioBERT augmented with the full PubMed dataset. One reason for this lies in the type of knowledge required to achieve high results on MedQA, MedMCQA, and MMLU-Medical benchmarks. In fact, the resolution of the targeted questions benefits particularly from the use of broad notional knowledge, rather than chunks of text extrapolated from highly specialized scientific publications, which are not always directly transferable and generalizable to other contexts, risking to introduce noise. To further validate our choice of MedWiki as the knowledge base, we conducted additional experiments using Textbooks (the medical text corpus attached to MedQA) as an alternative knowledge base for RAG. As shown

[14] `https://huggingface.co/NeuML/pubmedbert-base-embeddings`

[15] `https://huggingface.co/BAAI/bge-reranker-large`

[16] https://pubmed.ncbi.nlm.nih.gov/

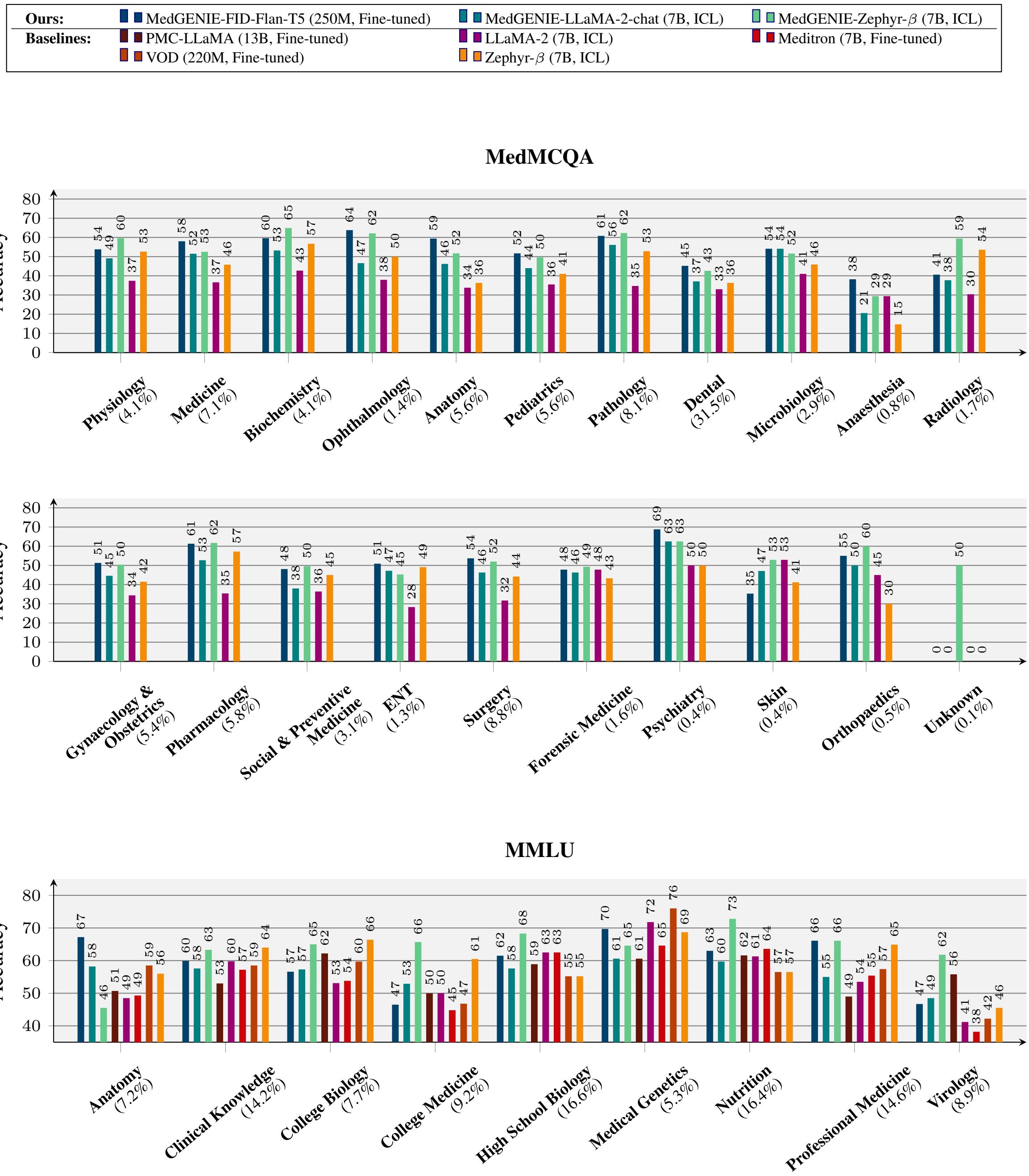


Figure 6: Fine-grained MedGENIE performance on MedMCQA (21 medical subjects) and MMLU-Medical (9 health and biology subjects). The frequency percentage for each subject is shown in brackets. In MedMCQA, Zephyr-β denotes our experiment with 2-shot prompting. Conversely, for MMLU, Zephyr-β refers to performance from (Chen et al., 2023) in 3-shot.

in Table 11, the results demonstrate that LLMs augmented with Textbooks only exhibit a marginal increase in performance with respect to WedWiki, corroborating the effectiveness of the latter. In light of these findings, we opted to maintain MedWiki as the knowledge base for all benchmarks, thereby providing a common reference point and ensuring consistency across our experiments.

J Qualitative Examples

Table 17 and Table 18 showcase qualitative examples of contexts generated by PMC-LLaMA on MedQA and MedMCQA, respectively. Concept links are highlighted with colors, while text spans that are highly effective in deducing the correct answer option are bolded.

K Prompt Templates

Figure 7 and Figure 8 report the constructed prompts for option-focused and option-free artificial context generation, with expert-crafted examples. Figure 9 and Figure 10 illustrate the prompt template used to perform 1-shot inference on MedQA with LLaMA-3-instruct and Phi-3-mini-instruct for artificial grounding (MedGENIE). Figure 11 and Figure 12 show the prompt template used to perform 2-shot inference on MedQA with Zephyr-β and LLaMA-2-chat for both artificial grounding and MedWiki grounding. Figure 13, Figure 14 and Figure 15 instead show how Zephyr-β, LLaMA-2-chat and PMC-LLaMA perform 2-shot inference on MedQA without grounding, respectively. These prompt templates are also applicable to MedMCQA and MMLU, with adjustments made by interchanging the provided 2-shot examples. Figure 16 and Figure 17 display the templates adopted with Zephyr-β and LLaMA-2-chat for performing 2-shot inference with either artificial grounding (MedGENIE) or MedWiki grounding on MedMCQA. The templates utilized for MMLU are presented in Figure 18 and Figure 19 instead.

Model	Ground (Source)	Learning	Params	MedQA (↓)
LLaMA-3-Instruct	∅	1-shot	8B	60.6
LLaMA-3-Meditron[14]	∅	?[**]	8B	60.6
OpenBioLLM[15]	∅	?[**]	8B	59.0
Phi-3-mini	∅	1-shot	3.8B	55.1
Codex[1]	∅	0-zhot	175B	52.5
MEDITRON[2]	∅	Fine-tuned	7B	52.0
BioMistral DARE[12]	∅	Fine-tuned	7B	51.1
BioMistral[12]	∅	Fine-tuned	7B	50.6
BioMedGPT[3]	∅	k-shot[**]	10B	50.4
BioMedLM[4]	∅	Fine-tuned	2.7B	50.3
PMC-LLaMA[*]	∅	Fine-tuned	13B	50.2
LLaMA-2[2]	∅	Fine-tuned	7B	49.6
Zephyr-β	∅	2-shot	7B	49.6
Zephyr-β[2]	∅	3-shot	7B	49.2
PMC-LLaMA[2]	∅	Fine-tuned	7B	49.2
Flan-PaLM[4]	∅	5-shot	62B	46.1
InstructGPT[1]	∅	0-shot	175B	46.0
Vicuna 1.3[1]	∅	0-shot	33B	45.2
BioLinkBERT[4]	∅	Fine-tuned	340M	45.1
Galactica	∅	0-shot	120B	44.4
LLaMA-2[1]	∅	0-shot	70B	43.4
Guanaco[1]	∅	0-shot	33B	42.9
LLaMA-2-chat[1]	∅	0-shot	70B	42.3
Vicuna 1.5[1]	∅	0-shot	65B	41.6
Mistral-Instruct[2]	∅	3-shot	7B	41.1
PaLM[4]	∅	5-shot	62B	40.9
Guanaco[1]	∅	0-shot	65B	40.8
MeditronLlama-Slerp[13]	∅	k-shot[**]	7B	39.20
Falcon-Instruct[1]	∅	0-shot	40B	39.0
Vicuna 1.3[1]	∅	0-shot	13B	38.7
PubMedBERT[4]	∅	Fine-tuned	110M	38.1
LLaMA-2-chat	∅	2-shot	7B	37.2
BioBERT[5]	∅	Fine-tuned	110M	36.7
MTP-Instruct[1]	∅	0-shot	30B	35.1
GPT-Neo[4]	∅	Fine-tuned	2.5B	33.3
MedGENIE-Phi-3-mini	G (PMC-LLaMA)	1-shot	3.8B	**64.7**
MedGENIE-LLaMA-3-Instruct	G (PMC-LLaMA)	1-shot	8B	<u>63.1</u>
MedGENIE-Zephyr-β	G (PMC-LLaMA)	2-shot	7B	59.7
MedGENIE-FID-Flan-T5	G (PMC-LLaMA)	Fine-tuned	250M	53.1
MedGENIE-LLaMA-2-chat	G (PMC-LLaMA)	2-shot	7B	52.6
Codex[1]	R (Wikipedia)	0-shot	175B	52.5
GPT-3.5-Turbo[6]	R (Wikipedia)	k-shot[**]	–	52.3
Zephyr-β	R (Textbooks)	2-shot	7B	51.4
Zephyr-β	R (MedWiki)	2-shot	7B	50.4
DRAGON[7]	R (UMLS)	Fine-tuned	360M	47.5
InstructGPT[1]	R (Wikipedia)	0-shot	175B	47.3
VOD[8]	R (MedWiki)	Fine-tuned	220M	45.8
Mistral-Instruct	R (MedWiki)	2-shot	7B	45.1
BioReader[9]	R (PubMed-RCT)	Fine-tuned	230M	43.0
GreaseLM[10]	R (UMLS)	Fine-tuned	359M	38.5
QA-GNN[11]	R (UMLS)	Fine-tuned	360M	38.0
LLaMA-2[6]	R (Wikipedia)	k-shot[**]	13B	37.6
LLaMA-2-chat	R (Textbooks)	2-shot	7B	37.9
LLaMA-2-chat	R (MedWiki)	2-shot	7B	37.2

[1] (Liévin et al., 2022) [2] (Chen et al., 2023) [3] (Luo et al., 2023) [4] (Singhal et al., 2022)
[5] (Lee et al., 2020) [6] (Wang et al., 2023) [7] (Yasunaga et al., 2022) [8] (Liévin et al., 2023)
[9] (Frisoni et al., 2022) [10] (Zhang et al., 2022) [11] (Yasunaga et al., 2021)
[12] (Labrak et al., 2024) [13] (Goddard et al., 2024) [14] https://www.meditron.io
[15] aaditya/Llama3-OpenBioLLM-8B
[*] AWQ 4-bit inference. [**] Lack of inference details.

Table 11: Accuracy comparison on MedQA. Bold and underline denote the best and second-best scores.

Model	Ground (Source)	Learning	Params	MedMCQA (↓)
MEDITRON[1]	∅	Fine-tuned	7B	**59.2**
LLaMA-3-Meditron[9]	∅	?[**]	8B	58.4
PMC-LLaMA[*]	∅	Fine-tuned	13B	57.6
OpenBioLLM[10]	∅	?[**]	8B	56.9
LLaMA-3-Instruct	∅	1-shot	8B	55.7
LLaMA-2[1]	∅	Fine-tuned	7B	54.4
Phi-3-mini	∅	1-shot	3.8B	53.5
Galactica[2]	∅	Fine-tuned	120B	52.9
PMC-LLaMA[1]	∅	Fine-tuned	7B	51.5
BioMedGPT[3]	∅	Fine-tuned	10B	51.4
Codex[4]	∅	0-shot	175B	50.9[**]
BioMistral DARE[7]	∅	Fine-tuned	7B	48.7
BioMistral[7]	∅	Fine-tuned	7B	48.1
Flan-PaLM[2]	∅	5-shot	62B	46.2
InstructGPT[4]	∅	0-shot	175B	44.0[**]
PaLM[2]	∅	5-shot	62B	43.4
Zephyr-β[1]	∅	3-shot	7B	43.0
Llama-2[4]	∅	0-shot	70B	42.8[**]
Zephyr-β	∅	2-shot	7B	42.5
Llama-2-chat[4]	∅	0-shot	70B	41.8[**]
Vicuna 1.5[4]	∅	0-shot	13B	41.5[**]
Mistral-Instruct[1]	∅	3-shot	7B	40.2
Vicuna 1.3[4]	∅	0-shot	65B	38.3[**]
Vicuna 1.3[4]	∅	0-shot	65B	38.0[**]
Guanaco[4]	∅	0-shot	33B	37.4[**]
MeditronLlama-Slerp[8]	∅	k-shot[**]	7B	36.9
Guanaco[4]	∅	0-shot	65B	36.7[**]
LLaMA-2-chat[4]	∅	0-shot	13B	36.6[**]
LLaMA-2-chat	∅	2-shot	7B	35.1
MPT-Instruct[4]	∅	0-shot	20B	34.6[**]
LLaMA-2[4]	∅	0-shot	13B	31.7[**]
Falcon-Instruct[4]	∅	0-shot	20B	30.0[**]
GPT-NeoX[4]	∅	0-shot	20B	27.8[**]
VOD [5]	R (MedWiki)	Fine-tuned	220M	58.3
MedGENIE-LLaMA-3-Instruct	G (PMC-LLaMA)	1-shot	8B	56.2
MedGENIE-Phi-3-mini	G (PMC-LLaMA)	1-shot	3.8B	54.1
MedGENIE-FID-Flan-T5	G (PMC-LLaMA)	Fine-tuned	250M	52.1
MedGENIE-Zephyr-β	G (PMC-LLaMA)	2-shot	7B	51.0
Zephyr-β	R (MedWiki)	2-shot	7B	47.1
InstructGPT [4]	R (Wikipedia)	0-shot	175B	46.7[**]
MedGENIE-LLaMA-2-chat	G (PMC-LLaMA)	2-shot	7B	44.8
Mistral-Instruct	R (MedWiki)	2-shot	7B	44.3
PubmedBERT [6]	R (Pubmed)	Fine-tuned	110M	43.0
SciBERT [6]	R (Pubmed)	Fine-tuned	110M	41.0
BioBERT [6]	R (Pubmed)	Fine-tuned	110M	39.0
LLaMA-2-chat	R (MedWiki)	2-shot	7B	37.3

[1] (Chen et al., 2023) [2] (Taylor et al., 2022) [3] (Luo et al., 2023) [4] (Liévin et al., 2022)
[5] (Liévin et al., 2023) [6] (Pal et al., 2022) [7] (Labrak et al., 2024) [8] (Goddard et al., 2024)
[9] https://www.meditron.io [10] aaditya/Llama3-OpenBioLLM-8B
[*] AWQ 4-bit inference. [**] Estimated using 1k samples.

Table 12: Accuracy comparison on MedMCQA. Bold and underline denote the best and second-best scores.

	Model	Ground (Source)	Learning	Params	MMLU (↓)
	Phi-3-mini	∅	1-shot	3.8B	70.3
	LLaMA-3-Instruct	∅	1-shot	8B	69.8
	Zephyr-β[1]	∅	3-shot	7B	63.3
	Galactica[2]	∅	Fine-tuned*	120B	61.8
	Zephyr-β	∅	2-shot	7B	60.5
	PMC-LLaMA[1]	∅	Fine-tuned*	7B	59.7
	LLaMA-2[1]	∅	Fine-tuned*	7B	56.3
	Mistral-Instruct[1]	∅	3-shot	7B	55.8
	MEDITRON[1]	∅	Fine-tuned*	7B	55.6
	LLaMA-2-chat	∅	2-shot	7B	49.4
	MedGENIE-Phi-mini	G (PMC-LLaMA)	1-shot	8B	**70.8**
	MedGENIE-LLaMA-3-Instruct	G (PMC-LLaMA)	2-shot	7B	68.9
	Zephyr-β	R (MedWiki)	2-shot	7B	66.9
	MedGENIE-FID-Flan-T5	G (PMC-LLaMA)	Fine-tuned*	250M	59.9
	MedGENIE-LLaMA-2-chat	G (PMC-LLaMA)	2-shot	7B	58.8
	Mistral-Instruct	R (MedWiki)	2-shot	7B	58.5
	VOD[2]	R (MedWiki)	Fine-tuned*	220M	56.8
	LLaMA-2-chat	R (MedWiki)	2-shot	7B	52.0

[1] (Chen et al., 2023) [2] (Liévin et al., 2023)
* Fine-tuned on the MedMCQA train set.

Table 13: Accuracy comparison on MMLU-Medical. Bold and underline denote the best and second-best scores.

MedQA (4 options)

Seed	A	B	C	D	E	Acc.	p-value
no shuffle	294▲	340▲	339	297▲	3	53.1	$7 \cdot 10^{-4}$
data	**353**	**309**	**346**	**265**	**0**		
4	201▼	254	427▲	389▲	2	51.1	$< 10^{-10}$
data	**309**	**265**	**353**	**346**	**0**		
11	170▼	343	283	475▲	2	51.9	$< 10^{-10}$
data	**265**	**346**	**309**	**353**	**0**		
13	233▼	316	330	469▲	1	51.1	$< 10^{-10}$
data	**346**	**309**	**353**	**265**	**0**		
40	171▼	346	402▲	351▲	3	52.3	$6 \cdot 10^{-10}$
data	**265**	**346**	**353**	**309**	**0**		
41	159▼	306	424▲	382▲	2	52.6	$< 10^{-10}$
data	**265**	**309**	**353**	**346**	**0**		
42	303▼	268	297	401▲	4	52.2	$9 \cdot 10^{-4}$
data	**353**	**265**	**309**	**346**	**0**		
43	298▼	334	244	395▲	2	52.9	$2 \cdot 10^{-4}$
data	**353**	**309**	**265**	**346**	**0**		
45	237▼	306	247	482▲	1	51.5	$< 10^{-10}$
data	**346**	**309**	**265**	**353**	**0**		
47	229▼	453▲	294	293▲	4	51.5	$< 10^{-10}$
data	**346**	**353**	**309**	**265**	**0**		
50	157▼	316	330	469▲	1	52.0	$< 10^{-10}$
data	**265**	**309**	**346**	**353**	**0**		

MedQA (5 options)

Seed	A	B	C	D	E	Acc.	p-value
no shuffle	231▼	274	221▼	266	281▲	45.6	$5 \cdot 10^{-9}$
data	**273**	**277**	**252**	**269**	**202**		
4	157▼	244	329▲	212	331▲	44.1	$< 10^{-10}$
data	**277**	**252**	**273**	**202**	**269**		
11	144▼	171▼	203▼	405▲	250	43.3	$< 10^{-10}$
data	**269**	**202**	**277**	**273**	**252**		
13	138▼	160▼	215▼	377▲	383▲	44.1	$< 10^{-10}$
data	**252**	**202**	**269**	**273**	**277**		
40	149▼	324▲	148▼	288	364▲	44.4	$< 10^{-10}$
data	**269**	**273**	**202**	**277**	**252**		
41	150▼	239	305▲	197	382▲	44.6	$< 10^{-10}$
data	**269**	**252**	**273**	**202**	**277**		
42	224▼	173▼	221▼	310▲	345▲	45.6	$3 \cdot 10^{-9}$
data	**273**	**202**	**252**	**277**	**269**		
43	226▼	239	215▼	194	399▲	45.3	$< 10^{-10}$
data	**273**	**252**	**269**	**202**	**277**		
45	145▼	220▼	193▼	194	521▲	42.3	$< 10^{-10}$
data	**252**	**269**	**277**	**202**	**273**		
47	149▼	341▲	239▼	200	344▲	44.9	$< 10^{-10}$
data	**252**	**273**	**277**	**202**	**269**		
50	142▼	231	209▼	413▲	278▲	44.4	$< 10^{-10}$
data	**269**	**252**	**277**	**273**	**202**		

MedMCQA

Seed	A	B	C	D	Acc.	p-value
no shuffle	1481	1072	899	731▼	52.1	$1 \cdot 10^{-5}$
data	**1348**	**1085**	**925**	**825**		
4	933▼	859	1582▲	809▼	51.2	$< 10^{-10}$
data	**1085**	**825**	**1348**	**925**		
11	810	941	1021	1411	51.2	$6 \cdot 10^{-2}$
data	**825**	**925**	**1085**	**1348**		
13	855	1042	1565▲	721▼	51.7	$< 10^{-10}$
data	**925**	**1085**	**1348**	**825**		
40	776	949	1591▲	867▼	51.0	$< 10^{-10}$
data	**825**	**925**	**1348**	**1085**		
41	787	1010	1584▲	802▼	51.4	$< 10^{-10}$
data	**825**	**1085**	**1348**	**925**		
42	1438	893	1058	794▼	51.1	$9 \cdot 10^{-7}$
data	**1348**	**825**	**1085**	**925**		
43	1493▲	1068	834	788▼	51.9	$6 \cdot 10^{-8}$
data	**1348**	**1085**	**825**	**925**		
45	883	1085	829	1386	51.2	$3 \cdot 10^{-1}$
data	**925**	**1085**	**825**	**1348**		
47	866	1603▲	1011	703▼	51.0	$< 10^{-10}$
data	**925**	**1348**	**1085**	**825**		
50	802	1053	921	1407	51.5	$2 \cdot 10^{-1}$
data	**825**	**1085**	**925**	**1348**		

MMLU-Medical

Seed	A	B	C	D	Acc.	p-value
no shuffle	490▲	441	471	460▼	59.9	$1 \cdot 10^{-9}$
data	**402**	**454**	**434**	**572**		
4	400▼	456▼	597▲	409	59.1	$< 10^{-10}$
data	**454**	**572**	**402**	**434**		
11	445▼	420	473	524▲	58.8	$< 10^{-10}$
data	**572**	**434**	**454**	**402**		
13	393	418	589▲	462▼	59.5	$< 10^{-10}$
data	**434**	**454**	**402**	**572**		
40	435▼	425	584▲	418	57.5	$< 10^{-10}$
data	**572**	**434**	**402**	**454**		
41	428▼	417	601▲	416	58.5	$< 10^{-10}$
data	**572**	**454**	**402**	**434**		
42	472▲	457▼	502▲	431	59.7	$8 \cdot 10^{-9}$
data	**402**	**572**	**454**	**434**		
43	497▲	425	511▼	429	59.9	$9 \cdot 10^{-7}$
data	**402**	**454**	**572**	**434**		
45	395	431	506▼	530▲	58.8	$< 10^{-10}$
data	**434**	**454**	**572**	**402**		
47	403	507▲	488	464▼	60.0	$< 10^{-10}$
data	**434**	**402**	**454**	**572**		
50	426▼	427	467	542▲	58.5	$< 10^{-10}$
data	**572**	**454**	**434**	**402**		

Table 14: Frequencies of predicted answers after option shuffling with distinct random seeds. Classification bias of MedGENIE-FID-Flan-T5 on MedQA (4 and 5 options), MedMCQA, and MMLU-Medical benchmarks. We highlight labels that are under-estimated using the color blue ▼ and over-estimated using the color red ▲ (±10% of the gold label frequency). Using the χ^2 test, we report the p-value for the null hypothesis *"the predictive distribution of the model equals the empirically observed one."*

Shuffling Seed	**4**	**11**	**13**	**40**	**41**	**42**	**43**	**45**	**47**	**50**	**Base (0)**	**AVG**
						MedGENIE-Zephyr-β						
MedQA	58.9	60.3	59.1	58.8	**60.6**	58.1	58.8	59.3	57.8	60.5	59.7	59.3
MedQA (5opt)	50.5	52.2	52.4	52.2	50.7	51.9	52.4	54.8	51.3	51.8	**54.9**	52.3
MedMCQA	51.0	50.7	**52.6**	51.5	52.0	49.6	50.2	51.5	51.0	51.1	51.0	51.1
MMLU	64.1	61.9	65.1	63.2	64.0	63.9	64.9	62.9	64.9	62.4	**66.1**	63.9
						MedGENIE-LLaMA-2-chat						
MedQA	50.4	**52.6**	51.9	51.1	51.5	51.8	51.6	51.8	52.2	51.2	**52.6**	51.7
MedQA (5opt)	46.1	42.4	44.6	44.1	44.1	45.8	**46.2**	45.9	45.2	44.3	46.0	45.0
MedMCQA	43.5	43.9	**45.1**	44.3	44.5	42.9	42.9	42.4	44.9	43.7	44.8	43.9
MMLU	54.4	53.1	54.8	53.7	52.9	56.7	56.6	54.1	56.1	53.4	**58.8**	55.0

Table 15: Classification bias of MedGENIE-Zephyr-β and MedGENIE-LLaMA-2-chat after option shuffling with distinct random seeds. Bold and underline denote the best and second-best scores for each model.

R	**G**	**Rerank**	**LLaMA-2-chat (7B)**	**mistral-instruct (7B)**	**Zephyr-β (7B)**
			MedQA		
4.5M	-	✗	37.2	45.1	50.4
4.5M	-	✔	36.3	44.6	50.5
4.5M	96K	✗	40.2	44.9	50.5
4.5M	96K	✔	41.4	45.6	50.8
4.5M	2M	✗	40.8	44.4	51.0
4.5M	2M	✔	40.5	45.9	51.2
			MedQA (5opt)		
4.5M	-	✗	32.8	35.1	44.0
4.5M	-	✔	32.2	36.8	44.7
4.5M	96K	✗	36.5	37.6	47.8
4.5M	96K	✔	35.8	37.9	47.5
4.5M	2M	✗	33.5	37.2	47.9
4.5M	2M	✔	36.3	37.9	47.8
			MedMCQA		
4.5M	-	✗	35.8	44.2	47.0
4.5M	-	✔	37.3	44.3	47.1
4.5M	96K	✗	40.9	47.4	50.1
4.5M	96K	✔	41.8	48.1	50.3
4.5M	2M	✗	43.1	48.6	50.9
4.5M	2M	✔	43.7	49.8	50.7
			MMLU (out-of-domain)		
4.5M	-	✗	51.8	58.5	66.9
4.5M	-	✔	52.0	58.3	66.4
4.5M	96K	✗	52.3	57.9	66.2
4.5M	96K	✔	53.5	58.3	67.4
4.5M	2M	✗	51.5	58.1	66.5
4.5M	2M	✔	53.0	58.4	67.0

Table 16: Results of experiments conducted to evaluate the effectiveness of combining artificial data with factual data sourced from MedWiki. In the absence of reranking, the top-5 retrieved context chunks are concatenated. Conversely, with reranking, we initially retrieve 10 chunks and subsequently narrow down to the top-5 through reranking. Context relevance is determined using cosine similarity. Here, R represents the total number of MedWiki chunks, while G represents the number of artificial context chunks considered.

MedQA [test:50]	
Question	A 65-year-old male is treated for anal carcinoma with therapy including external beam radiation. How does radiation affect cancer cells?
Options	A. Induces the formation of thymidine dimers B. Induces the formation of disulfide C. Induces deamination of cytosine **D. Induces breaks in double-stranded DNA** (✓)
#1 Option-focused context	Radiation therapy for cancer works by **damaging cellular DNA** through the formation of free radicals, which **can cause breaks in double-stranded DNA**. It **does not** induce the formation of thymidine dimers or disulfide bonds. Deamination is a chemical reaction that affects cytosine residues in DNA and RNA, but **this does not result from radiation**.
#1 Option-free context	External beam radiation therapy (EB) describes the delivery of ionizing radiation to a target volume in the body using externally generated X-rays or gamma rays. Radiation is most often used for its **cytotoxic effects** on cancer cells, and generally works by **damaging cellular DNA so that it can no longer function properly**. Normal healthy cells are also affected by radiotherapy, however, and EB may cause acute side effects as well as possibly contribute to late side effects that can appear months or even years after treatment has concluded. The extent of these normal tissue/organ effects is dependent on several factors including: total dose administered; fractionation schedule (i.e., number of treatments delivered and their frequency); overall time course over which therapy is delivered; region(s) of body being irradiated; inherent radiosensitivity of organs in close proximity to target tumor; underlying comorbidities such as diabetes mellitus type II or coronary artery disease which may independently increase risk for complication development following exposure to ionizing radiation).

Table 17: Examples of contexts generated by PMC-LLaMA for a MedQA test set instance.

MedMCQA [eval:c6365cce-507c-40f6-90a2-46b867f47b6e]	
Question	Axonal transport is:
Options	A. Antegrade B. Retrograde **C. Antegrade and retrograde** (✓) D. None
#1 Option-focused context	Axons are long tubular structures that conduct nerve impulses. **Antegrade** transport occurs in the direction of the axon's tip, **whereas retrograde** transport is in the opposite direction, toward the cell body.
#2 Option-focused context	The question is asking about the **direction** of transport along axons. Axons are specialized structures that conduct electrical impulses in nerve cells and can extend up to a meter or more in length. Materials such as proteins, organelles, and other cellular components need to be transpoed along the axons between the cell body and the synapses (ends of neuron that connect with those of other neurons). **Transport may occur in either direction**: **Antegrade** transport refers to movement from the cell body toward the synapse; **retrograde** transport is movement from synaptic end back toward the cell body.
#1 Option-free context	Axonal transport is a part of the intracellular transport system in which substances are carried from the cell body along microtubules to the axon terminals. **It includes:** - Transports towards axon terminal (**antegrade**) - Transports back to cell body (**retrograde**) One example of antegrade movement is when materials, such as vesicles containing neurotransmitters, are carried from the soma through the axon and out its endings where they fuse with plasma membrane for release into synaptic cleft. An example of retrograde movement would be when sugar molecules and proteins are returned to their origin within cells. In this way, neurons can communicate information about incoming stimuli using chemical signals called neurotransmitters that act upon receptors found on dendrites or cell bodies. *[...]*

Table 18: Examples of contexts generated by PMC-LLaMA for a MedMCQA test set instance.

```
### Question:
Chronic urethral obstruction due to benign prismatic hyperplasia can lead to the following
change in kidney parenchyma:
- Hyperplasia
- Hyperophy
- Atrophy
- Dyplasia

### Context:
In the scenario of chronic urethral obstruction caused by benign prostatic hyperplasia (BPH),
it's essential to consider the impact on the kidney parenchyma. The continuous blockage of the
urethra, a tube responsible for carrying urine from the bladder, can result in a condition known
as hydronephrosis.

Chronic Urethral Obstruction:
- This condition involves a persistent blockage in the urethra, commonly caused by the
non-cancerous enlargement of the prostate gland, known as benign prostatic hyperplasia (BPH).

Kidney Parenchyma and Hydronephrosis:
- The kidney parenchyma is the functional tissue of the kidney responsible for filtration and
urine production.
- Hydronephrosis is the swelling or enlargement of the kidney due to the backup of urine
caused by the obstruction. The increased pressure within the kidney can lead to changes in its
structure and function.

Understanding the Terms:
- Hyperplasia: Refers to an increase in the number of cells. Consider whether this is the likely
change in the kidney parenchyma due to chronic urethral obstruction.
- Hyperophy: This seems to be a misspelling. It might be intended as "hypertrophy," which refers
to an increase in the size of cells. Consider if this is a probable outcome in the context of
the described condition.
- Atrophy: Describes a reduction in the size or function of an organ or tissue. Consider whether
atrophy is a likely change in the kidney parenchyma due to prolonged obstruction.
- Dyplasia: Likely a typographical error. The correct term is "dysplasia," which refers to
abnormal development or growth of cells. Consider whether dysplasia is a probable outcome in
the kidney parenchyma.

By understanding the impact of chronic urethral obstruction on the kidney and considering the
definitions of the provided terms, you can deduce the potential change in kidney parenchyma.

### Question:
Which vitamin is supplied from only animal source:
- Vitamin C
- Vitamin B7
- Vitamin B12
- Vitamin D

### Context:
Vitamins are essential micronutrients required by the human body for various physiological
functions. They are classified into different groups, and each vitamin plays a specific role
in maintaining health. The question pertains to identifying the vitamin that is exclusively
supplied from animal sources.
- Vitamin C (Ascorbic Acid): This vitamin is found in various fruits and vegetables, particularly
citrus fruits, berries, and leafy greens. It is not exclusive to animal sources.
- Vitamin B7 (Biotin): Biotin is a water-soluble vitamin found in a variety of foods, including
meat, fish, eggs, nuts, seeds, and certain vegetables. While it is present in some animal
products, it is not exclusively derived from animals.
- Vitamin B12 (Cobalamin): Vitamin B12 is unique in that it is primarily found in animal-based
sources. It plays a crucial role in neurological function and the formation of red blood cells.
Dietary sources include meat, fish, eggs, and dairy products. Vegetarians and vegans may need
to supplement or rely on fortified foods to meet their B12 requirements since plant-based foods
generally lack this vitamin.
- Vitamin D: Vitamin D is synthesized in the skin upon exposure to sunlight and is also found
in some food sources. While animal products such as fatty fish, liver, and egg yolks contain
vitamin D, it can also be obtained from fortified foods and supplements. Therefore, vitamin D
is not exclusively derived from animal sources.

### Question:
{{new_question}}

### Context:
```


Figure 7: Prompt template for generating an *option-focused* context with PMC-LLaMA.

```
### Question:
Chronic urethral obstruction due to benign prismatic hyperplasia can lead to the following
change in kidney parenchyma:

### Context:
Benign prostatic hyperplasia (BPH) is a common condition in aging men characterized by the
non-malignant enlargement of the prostate gland. The prostate surrounds the urethra, and its
enlargement can lead to various urinary symptoms such as difficulty in urination, incomplete
emptying of the bladder, and increased frequency of urination. When BPH causes chronic urethral
obstruction, it can have implications for the kidneys and their parenchyma. The term "parenchyma"
refers to the functional tissue of an organ, and in the case of the kidneys, it includes the
renal cortex and medulla, where vital functions such as filtration, reabsorption, and secretion
occur. Chronic urethral obstruction can create back pressure on the urinary system, impacting
the flow of urine from the kidneys to the bladder. This increased pressure in the urinary tract
can lead to several changes in the kidney parenchyma, collectively referred to as obstructive
nephropathy. Some of the key changes include:
Hydronephrosis: The prolonged obstruction of urine flow can cause the renal pelvis and calyces
to dilate, a condition known as hydronephrosis. This dilation is a result of the accumulation
of urine upstream of the obstruction, causing stretching and expansion of the renal structures.
Interstitial fibrosis: Chronic obstruction may lead to inflammation and fibrosis in the
interstitium of the kidney. Fibrosis is the excessive formation of connective tissue, and in
this context, it can replace normal kidney tissue, impairing its function.
Renal atrophy: Prolonged obstruction and the associated changes can lead to the atrophy of renal
tubules and glomeruli. This atrophy is a consequence of the reduced blood flow and the pressure
exerted on the kidney tissues.
Impaired renal function: Over time, the structural changes in the kidney parenchyma can result
in impaired renal function. The ability of the kidneys to filter waste products, regulate
electrolytes, and maintain fluid balance may be compromised.
Renal failure: In severe cases, chronic urethral obstruction due to BPH can progress to renal
failure, where the kidneys are no longer able to adequately perform their vital functions. This
is a serious and potentially life-threatening condition that may require medical intervention,
such as surgery to relieve the obstruction.

### Question:
Which vitamin is supplied from only animal source:

### Context:
Vitamin B12, also known as cobalamin, is the vitamin that is primarily supplied from only animal
sources. Unlike many other vitamins that can be obtained from both plant and animal sources,
vitamin B12 is unique in its occurrence primarily in animal-derived foods. Vitamin B12 plays a
crucial role in various physiological processes, including the formation of red blood cells,
neurological function, and DNA synthesis. It is essential for maintaining the health of nerve
cells and aiding in the production of DNA and RNA.
Common sources of vitamin B12 from animal products include:
- Meat: Particularly, beef, pork, and lamb are good sources of vitamin B12.
- Poultry: Chicken and turkey also contain vitamin B12.
- Fish: Fatty fish such as salmon, trout, and tuna are good sources.
- Shellfish: Clams, oysters, and mussels are rich in vitamin B12.
- Dairy products: Milk, cheese, and eggs contain vitamin B12, although in smaller amounts
compared to meat and fish.
Since vitamin B12 is not found in significant amounts in plant foods, individuals following
a strict vegetarian or vegan diet may be at risk of B12 deficiency and may need to consider
supplementation or fortified foods to meet their dietary requirements. It's an essential
nutrient for overall health, and a deficiency can lead to various health issues, including
anemia and neurological problems.

### Question:
{{new_question}}

### Context:
```

Figure 8: Prompt template for generating an *option-free* context with PMC-LLaMA.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a medical expert. Your task is to answer questions related to medical exams based on a given context by selecting the correct option. Return as output only the selected option.<|eot_id|><|start_header_id|>user<|end_header_id|>

Select only one option. Don't explain your choice.

Context: The scenario suggests sudden infant death syndrome (SIDS). SIDS is the unexpected, sudden death of an infant that cannot be explained by history or autopsy findings. This condition occurs when infants are placed to sleep on their stomachs and are not covered by a firm bedding. Placing babies on a firm mattress in a supine position reduces the risk of SIDS significantly.
Sudden infant death syndrome (SIDS) is defined as the sudden and unexplained death of an appearance healthy infant younger than one year old. Risk factors include sleeping in the prone position, soft surfaces on which to sleep, bed sharing with adults, late or no immunization for infectious diseases like HIV.
Sudden infant death syndrome (SIDS) is the unexpected, sudden death of a child under one year old. An autopsy does not show an explainable cause of death in cases with SIDS. Placing the child in a supine position on a firm mattress while sleeping decreases the risk of SIDS by preventing potential hazards such as soft bedding material or entrapment risks that could compromise respiration. The remaining options do not significantly reduce SIDS risks and are therefore incorrect answers to this question.
Sudden infant death syndrome (SIDS) is the sudden, unexplained death of an apparently healthy baby. Evidence suggests that the risk of SIDS can be reduced by:
- Placing babies in a supine position (on their backs) for sleep
- Using a firm sleep surface, such as a crib mattress covered with a fitted sheet
- Maintaining soft bedding and loose clothing tucked around the baby to prevent facility to wedge herself between two surfaces or get trapped or wedged between beds, furniture or other objects. [...]

Question: A 3-month-old baby died suddenly at night while asleep. His mother noticed that he had died only after she awoke in the morning. No cause of death was determined based on the autopsy. Which of the following precautions could have prevented the death of the baby?
A. Placing the infant in a supine position on a firm mattress while sleeping
B. Keeping the infant covered and maintaining a high room temperature
C. Application of a device to maintain the sleeping position
D. Avoiding pacifier use during sleep
Answer:<|eot_id|><|start_header_id|>assistant<|end_header_id|>

A. Placing the infant in a supine position on a firm mattress while sleeping.
<|eot_id|><|start_header_id|>user<|end_header_id|>

Now help me with another question. Just select only one option as you did so far. Don't explain your choice.

Context: {{new_context}}

Question: {{new_question}}
{{new_option_set}}
Answer:<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

Figure 9: Prompt template for question answering with LLaMA-3-Instruct on MedQA by leveraging artificial grounding.

```
<s><|user|>
You are a medical expert. Your task is to answer questions related to medical exams based on a
given context by selecting the correct option. Return as output only the selected option.

Context: The scenario suggests sudden infant death syndrome (SIDS). SIDS is the unexpected,
sudden death of an infant that cannot be explained by history or autopsy findings. This condition
occurs when infants are placed to sleep on their stomachs and are not covered by a firm bedding.
Placing babies on a firm mattress in a supine position reduces the risk of SIDS significantly.
Sudden infant death syndrome (SIDS) is defined as the sudden and unexplained death of an
appearance healthy infant younger than one year old. Risk factors include sleeping in the prone
position, soft surfaces on which to sleep, bed sharing with adults, late or no immunization for
infectious diseases like HIV.
Sudden infant death syndrome (SIDS) is the unexpected, sudden death of a child under one year
old. An autopsy does not show an explainable cause of death in cases with SIDS. Placing the
child in a supine position on a firm mattress while sleeping decreases the risk of SIDS by
preventing potential hazards such as soft bedding material or entrapment risks that could
compromise respiration. The remaining options do not significantly reduce SIDS risks and are
therefore incorrect answers to this question.
Sudden infant death syndrome (SIDS) is the sudden, unexplained death of an apparently healthy
baby. Evidence suggests that the risk of SIDS can be reduced by:
- Placing babies in a supine position (on their backs) for sleep
- Using a firm sleep surface, such as a crib mattress covered with a fitted sheet
- Maintaining soft bedding and loose clothing tucked around the baby to prevent facility to
wedge herself between two surfaces or get trapped or wedged between beds, furniture or other
objects. [...]

Question: A 3-month-old baby died suddenly at night while asleep. His mother noticed that he
had died only after she awoke in the morning. No cause of death was determined based on the
autopsy. Which of the following precautions could have prevented the death of the baby?
A. Placing the infant in a supine position on a firm mattress while sleeping
B. Keeping the infant covered and maintaining a high room temperature
C. Application of a device to maintain the sleeping position
D. Avoiding pacifier use during sleep

Answer by returning only the correct letter among A, B, C or D. Don't explain your choice.
Answer:<|end|>
<|assistant|>
A<|end|>
<|user|>
Context: {{new_context}}

Question: {{new_question}}
{{new_option_set}}

Answer by returning only the correct letter among A, B, C or D. Don't explain your choice.
Answer:<|end|>
<|assistant|>
```

Figure 10: Prompt template for question answering with Phi-3-mini on MedQA by leveraging artificial grounding.

```
<|system|>
You are a medical expert. Your task is to answer questions related to medical exams based on a
given context by selecting the correct option. Return as output only the selected option.</s>
<|user|>
Select only one option. Don't explain your choice.

### Context:
Nitrofurantoin is a commonly used antibiotic for the treatment of uncomplicated urinary tract
infections (UTIs) in pregnant women. It is considered safe during pregnancy and is effective
against common pathogens causing UTIs. Ampicillin and ceftriaxone are not the first-line
choices for treating uncomplicated UTIs, and doxycycline is contraindicated in pregnancy due to
potential adverse effects on fetal development.

### Question:
A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She
states it started 1 day ago and has been worsening despite drinking more water and taking
cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her
temperature is 97.7°F (36.5°C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations
are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for an absence
of costovertebral angle tenderness and a gravid uterus. Which of the following is the best
treatment for this patient?
(A) Ampicillin
(B) Ceftriaxone
(C) Doxycycline
(D) Nitrofurantoin</s>
<|assistant|>
(D) Nitrofurantoin.</s>
<|user|>
Select only one option. Don't explain your choice.

### Context:
Placing the infant in a supine position on a firm mattress while sleeping is the recommended
precaution to reduce the risk of sudden infant death syndrome (SIDS). This position helps
maintain clear airways and minimizes the risk of suffocation. Avoiding practices such as
covering the infant excessively, using devices to maintain sleeping position, and prohibiting
pacifier use during sleep are not recommended and may pose additional risks.

### Question:
A 3-month-old baby died suddenly at night while asleep. His mother noticed that he had died only
after she awoke in the morning. No cause of death was determined based on the autopsy. Which of
the following precautions could have prevented the death of the baby?
(A) Placing the infant in a supine position on a firm mattress while sleeping
(B) Keeping the infant covered and maintaining a high room temperature
(C) Application of a device to maintain the sleeping position
(D) Avoiding pacifier use during sleep</s>
<|assistant|>
(A) Placing the infant in a supine position on a firm mattress while sleeping.</s>
<|user|>
Now help me with another question. Just select only one option as you did so far. Don't explain
your choice.

### Context:
{{new_context}}

### Question:
{{new_question}}</s>
<|assistant|>
```

Figure 11: Prompt template for question answering with Zephyr-β on MedQA by leveraging artificial/MedWiki grounding. We include only the best 2-shot pairs as per Table 9. It should be noted that for MedWiki grounding, these pairs always correspond to the human-made ones (H) according to Table 7.

```
[INST] «SYS»
You are a medical expert. Your task is to answer questions related to medical exams based on a
given context. Answer as concise as possible. Your answer must be always a string of one line
starting with "The answer is", followed by your final choice. Nothing more.
«/SYS»
Make a choice based on the context and question. Take the following two questions as examples.

# Example 1
### Context:
Most outpatient physicians treat asymptomatic bacteriuria with sulfate-based cephalosporins
such as nitrofurantoin (100 mg BID for 7 days) or cephalexin (500mg tid for 7 days). Both drugs
are considered safe during pregnancy.

### Question:
A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She
states it started 1 day ago and has been worsening despite drinking more water and taking
cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her
temperature is 97.7°F (36.5°C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations
are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for an absence
of costovertebral angle tenderness and a gravid uterus. Which of the following is the best
treatment for this patient?
(A) Ampicillin
(B) Ceftriaxone
(C) Doxycycline
(D) Nitrofurantoin

The answer is (D) Nitrofurantoin.

# Example 2
### Context:
Sudden infant death syndrome (SIDS) is the unexpected, sudden death of a child under one year
old. An autopsy does not show an explainable cause of death in cases with SIDS. Placing the
child in a supine position on a firm mattress while sleeping decreases the risk of SIDS by
preventing potential hazards such as soft bedding material or entrapment risks that could
compromise respiration.

### Question:
A 3-month-old baby died suddenly at night while asleep. His mother noticed that he had died only
after she awoke in the morning. No cause of death was determined based on the autopsy. Which of
the following precautions could have prevented the death of the baby?
(A) Placing the infant in a supine position on a firm mattress while sleeping
(B) Keeping the infant covered and maintaining a high room temperature
(C) Application of a device to maintain the sleeping position
(D) Avoiding pacifier use during sleep

The answer is (A) Placing the infant in a supine position on a firm mattress while sleeping.

Now, help me with this question. Remember to answer with just a string of one line starting
with "The answer is" as shown by the previous examples.

### Context:
{{new_context}}

### Question:
{{new_question}} [/INST]
```

Figure 12: Prompt template for question answering with LLaMA-2-chat on MedQA by leveraging artificial/MedWiki grounding. We include only the best 2-shot pairs as per Table 9. It should be noted that for MedWiki grounding, these pairs always correspond to the human-made ones (H) according to Table 7.

```
<|system|>
You are a medical expert. Your task is to answer questions related to medical
exams by selecting the correct option.  Return as output only the selected
option.</s>
<|user|>
Select only one option. Don't explain your choice.

### Question:
A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon
urination.  She states it started 1 day ago and has been worsening despite
drinking more water and taking cranberry extract. She otherwise feels well and
is followed by a doctor for her pregnancy. Her temperature is 97.7°F (36.5°C),
blood pressure is 122/77 mmHg, pulse is 80/min, respirations are 19/min, and
oxygen saturation is 98% on room air. Physical exam is notable for an absence
of costovertebral angle tenderness and a gravid uterus. Which of the following
is the best treatment for this patient?
(A) Ampicillin
(B) Ceftriaxone
(C) Doxycycline
(D) Nitrofurantoin</s>
<|assistant|>
(D) Nitrofurantoin.</s>
<|user|>
Select only one option. Don't explain your choice.

### Question:
A 3-month-old baby died suddenly at night while asleep.  His mother noticed
that he had died only after she awoke in the morning.  No cause of death was
determined based on the autopsy. Which of the following precautions could have
prevented the death of the baby?
(A) Placing the infant in a supine position on a firm mattress while sleeping
(B) Keeping the infant covered and maintaining a high room temperature
(C) Application of a device to maintain the sleeping position
(D) Avoiding pacifier use during sleep</s>
<|assistant|>
(A) Placing the infant in a supine position on a firm mattress while
sleeping.</s>
<|user|>
Now help me with another question. Just select only one option as you did so
far. Don't explain your choice.

### Question:
{{new_question}}</s>
<|assistant|>
```

Figure 13: Prompt template for question answering with Zephyr-β **without grounding** on MedQA.

```
[INST]«SYS»
You are a medical expert. Your task is to answer questions related to medical exams. Answer as concise as possible. Your answer must be always a string of one line starting with "The answer is", followed by your final choice. Nothing more.
«/SYS»
# Example 1
### Question:
A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She states it started 1 day ago and has been worsening despite drinking more water and taking cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her temperature is 97.7°F (36.5°C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for an absence of costovertebral angle tenderness and a gravid uterus. Which of the following is the best treatment for this patient?
A. Ampicillin
B. Ceftriaxone
C. Doxycycline
D. Nitrofurantoin

### Answer:
D. Nitrofurantoin

# Example 2
### Question:
A 3-month-old baby died suddenly at night while asleep. His mother noticed that he had died only after she awoke in the morning. No cause of death was determined based on the autopsy. Which of the following precautions could have prevented the death of the baby?
(A) Placing the infant in a supine position on a firm mattress while sleeping
(B) Keeping the infant covered and maintaining a high room temperature
(C) Application of a device to maintain the sleeping position
(D) Avoiding pacifier use during sleep

The answer is (A) Placing the infant in a supine position on a firm mattress while sleeping.

Now, help me with this question. Remember to answer with just a string of one line starting with "The answer is" as shown by the previous examples.

### Question:
{{new_question}}[/INST]
```

Figure 14: Prompt template for question answering with LLaMA-2-chat **without grounding** on MedQA.

```
### Instruction:
Make a choice based on the question and options. Take the following two questions as examples.

# Few-shot Example 1
### Question:
A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She
states it started 1 day ago and has been worsening despite drinking more water and taking
cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her
temperature is 97.7°F (36.5°C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations
are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for an absence
of costovertebral angle tenderness and a gravid uterus. Which of the following is the best
treatment for this patient?
A. Ampicillin
B. Ceftriaxone
C. Doxycycline
D. Nitrofurantoin

### Answer:
D. Nitrofurantoin

# Few-shot Example 2
### Question:
A 3-month-old baby died suddenly at night while asleep. His mother noticed that he had died only
after she awoke in the morning. No cause of death was determined based on the autopsy. Which of
the following precautions could have prevented the death of the baby?
A. Placing the infant in a supine position on a firm mattress while sleeping
B. Keeping the infant covered and maintaining a high room temperature
C. Application of a device to maintain the sleeping position
D. Avoiding pacifier use during sleep

### Answer:
A. Placing the infant in a supine position on a firm mattress while sleeping

Now help me with another question

### Question:
{{new_question}}

### Answer:
```

Figure 15: Prompt template for question answering with PMC-LLaMA **without grounding** on MedQA.

```
<|system|>
You are a medical expert. Your task is to answer questions related to medical exams based on a
given context by selecting the correct option. Return as output only the selected option.</s>
<|user|>
Select only one option. Don't explain your choice.

### Context:
Antibiotic prophylaxis for infective endocarditis is indicated in individuals with predisposing
cardiac conditions. In this scenario, determining if an isolated secundum ASD and mitral valve
prolapse without regurgitation are associated with the potential risk of developing infective
endocarditis requires further information. The presence of a prior coronary aery bypass graft
and coarctation of aoa are both established indications for antibiotic prophylaxis due to their
association with infective endocarditis risk.

### Question:
Antiboiotic Prophylaxis for infective endocarditis is indicated in:
(A) Isolated secundum ASD
(B) Mitral valve prolapse without regurgitation
(C) Prior coronary aery bypass graft
(D) Coarctation of aoa</s>
<|assistant|>
(D) Coarctation of aoa.</s>
<|user|>
Select only one option. Don't explain your choice.

### Context:
The Anterolateral portal is also known as the lateral portal.  It is used for viewing the
patellofemoral joint, inserting probe or laser for soft-tissue procedures.

### Question:
Anterolateral ahroscopy of knee is for:
(A) To see patellofemoral aiculation
(B) To see the posterior cruciate ligament
(C) To see the anterior poion of lateral meniscus
(D) To see the periphery of the posterior horn of medial meniscus</s>
<|assistant|>
(A) To see patellofemoral aiculation.</s>
<|user|>
Now help me with another question. Just select only one option as you did so far. Don't explain
your choice.

### Context:
{{new_context}}

### Question:
{{new_question}}
```

Figure 16: Prompt template for question answering with Zephyr-β on MedMCQA by leveraging artificial/MedWiki grounding. We include only the best 2-shot pairs as per Table 9. It should be noted that for MedWiki grounding, these pairs always correspond to the human-made ones (H) according to Table 7.

```
[INST] «SYS»
You are a medical expert. Your task is to answer questions related to medical exams based on a
given context. Answer as concise as possible. Your answer must be always a string of one line
starting with "The answer is", followed by your final choice. Nothing more.
«/SYS»
Make a choice based on the context and question. Take the following two questions as examples.

# Example 1
### Context:
Hyperviscosity is a condition where the blood becomes abnormally thick, hindering its ability
to flow properly. Cryoglobulinemia is a condition characterized by abnormal antibodies in the
blood (antibodies are specialized cells that recognize and attack foreign invaders). These
abnormal antibodies become solid at cold temperatures and lead to clumping of red blood cells,
an increase in viscosity, and subsequent obstruction of small vessels.

### Question:
Hyper viscosity is seen in
(A) Cryoglobulinemia
(B) Multiple myeloma
(C) MGUS
(D) Lymphoma

The answer is (A) Cryoglobulinemia.

# Example 2
### Context:
Inversion of the foot refers to a foot in which its sole faces medially. Since the plantar
surface of the foot is in contact with the ground at all times, this condition occurs when one
or more muscles responsible for moving it become tight or weak, resulting in an alteration in
normal alignment.

### Question:
In inversion of the foot, the sole will face:
(A) Upwards
(B) Downwards
(C) Laterally
(D) Medially

The answer is (D) Medially.

Now, help me with this question. Remember to answer with just a string of one line starting
with "The answer is" as shown by the previous examples.

### Context:
{{new_context}}

### Question:
{{new_question}}[/INST]
```

Figure 17: Prompt template for question answering with LLaMA-2-chat on MedMCQA by leveraging artificial/MedWiki grounding. We include only the best 2-shot pairs as per Table 9. It should be noted that for MedWiki grounding, these pairs always correspond to the human-made ones (H) according to Table 7.

```
<|system|>
You are a medical expert. Your task is to answer questions related to medical exams based on a
given context by selecting the correct option. Return as output only the selected option.</s>
<|user|>
Select only one option. Don't explain your choice.

### Context:
Hyperviscosity is a condition where the blood becomes abnormally thick, hindering its ability
to flow properly. Cryoglobulinemia is a condition characterized by abnormal antibodies in the
blood (antibodies are specialized cells that recognize and attack foreign invaders). These
abnormal antibodies become solid at cold temperatures and lead to clumping of red blood cells,
an increase in viscosity, and subsequent obstruction of small vessels.

### Question:
Hyper viscosity is seen in
(A) Cryoglobulinemia
(B) Multiple myeloma
(C) MGUS
(D) Lymphoma</s>
<|assistant|>
(A) Cryoglobulinemia.</s>
<|user|>
Select only one option. Don't explain your choice.

### Context:
Inversion of the foot refers to a foot in which its sole faces medially. Since the plantar
surface of the foot is in contact with the ground at all times, this condition occurs when one
or more muscles responsible for moving it become tight or weak, resulting in an alteration in
normal alignment.

### Question:
In inversion of the foot, the sole will face:
(A) Upwards
(B) Downwards
(C) Laterally
(D) Medially</s>
<|assistant|>
(D) Medially.</s>
<|user|>
Now help me with another question. Just select only one option as you did so far. Don't explain
your choice.

### Context:
{{new_context}}

### Question:
{{new_question}}
```

Figure 18: Prompt template for question answering with Zephyr-β on MMLU-Medical by leveraging artificial/MedWiki grounding. We include only the best 2-shot pairs as per Table 9. It should be noted that for MedWiki grounding, these pairs always correspond to the human-made ones (H) according to Table 7.

```
[INST] «SYS»
You are a medical expert. Your task is to answer questions related to medical exams based on a
given context. Answer as concise as possible. Your answer must be always a string of one line
starting with "The answer is", followed by your final choice. Nothing more.
«/SYS»
Make a choice based on the context and question. Take the following two questions as examples.

# Example 1
### Context:
Antibiotic prophylaxis for infective endocarditis is indicated in individuals with predisposing
cardiac conditions. In this scenario, determining if an isolated secundum ASD and mitral valve
prolapse without regurgitation are associated with the potential risk of developing infective
endocarditis requires further information. The presence of a prior coronary aery bypass graft
and coarctation of aoa are both established indications for antibiotic prophylaxis due to their
association with infective endocarditis risk.

### Question:
Antiboiotic Prophylaxis for infective endocarditis is indicated in:
(A) Isolated secundum ASD
(B) Mitral valve prolapse without regurgitation
(C) Prior coronary aery bypass graft
(D) Coarctation of aoa

The answer is (D) Coarctation of aoa.

# Example 2
### Context:
The Anterolateral portal is also known as the lateral portal.  It is used for viewing the
patellofemoral joint, inserting probe or laser for soft-tissue procedures.

### Question:
Anterolateral ahroscopy of knee is for:
(A) To see patellofemoral aiculation
(B) To see the posterior cruciate ligament
(C) To see the anterior poion of lateral meniscus
(D) To see the periphery of the posterior horn of medial meniscus

The answer is (A) To see patellofemoral aiculation.

Now, help me with this question.  Remember to answer with just a string of one line starting
with "The answer is" as shown by the previous examples.

### Context:
{{new_context}}

### Question:
{{new_question}}[/INST]
```

Figure 19: Prompt template for question answering with LLaMA-2-chat on MMLU-Medical. We include only the best 2-shot pairs as per Table 9. It should be noted that for MedWiki grounding, these pairs always correspond to the human-made ones (H) according to Table 7.