Impact of High-Quality, Mixed-Domain Data on the Performance of Medical Language Models

Maxime Griot, M.D, MSc – (1) Institute of NeuroScience, Université Catholique de Louvain, Brussels, Belgium (2) Louvain Research Institute in Management and Organizations, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Coralie Hemptinne, M.D, PhD – (1) Ophthalmology, Cliniques Universitaires Saint-Luc, Brussels, Belgium

Jean Vanderdonckt, PhD – (1) Louvain Research Institute in Management and Organizations, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Demet Yuksel, M.D, PhD – (1) Medical Information Department, Cliniques Universitaires Saint-Luc, Brussels, Belgium – (2) Institute of NeuroScience, Université Catholique de Louvain, Brussels, Belgium

Corresponding Author: Maxime Griot, MD, Msc, Institute of NeuroScience, Université Catholique de Louvain, Avenue Mounier 53, boîte B1.53.02, 1200, Woluwe-Saint-Lambert, Belgium (maxime.griot@uclouvain.be, phone: +33660338520).

Abstract

Objective: To optimize the training strategy of large language models for medical applications, focusing on creating clinically relevant systems that efficiently integrate into healthcare settings, while ensuring high standards of accuracy and reliability.

Materials and Methods: We curated a comprehensive collection of high-quality, domain-specific data and used it to train several models, each with different subsets of this data. These models were rigorously evaluated against standard medical benchmarks, such as the USMLE, to measure their performance. Furthermore, for a thorough effectiveness assessment, they were compared with other state-of-the-art medical models of comparable size.

Results: The models trained with a mix of high-quality, domain-specific, and general data showed superior performance over those trained on larger, less clinically relevant datasets ($p<0.001$). Our 7-billion-parameter model Med_5 scores 60.5% on MedQA, outperforming the previous best of 49.3% from comparable models, and becomes the first of its size to achieve a passing score on the USMLE. Additionally, this model retained its proficiency in general domain tasks, comparable to state-of-the-art general domain models of similar size.

Discussion: Our findings underscore the importance of integrating high-quality, domain-specific data in training large language models for medical purposes. The balanced approach between specialized and general data significantly enhances the model's clinical relevance and performance.

Conclusion: This study sets a new standard in medical language models, proving that a strategically trained, smaller model can outperform larger ones in clinical relevance and general proficiency, highlighting the importance of data quality and expert curation in generative artificial intelligence for healthcare applications.

BACKGROUND AND SIGNIFICANCE

Since the introduction of GPT3.5 in October 2022, there has been a heightened interest in large language models (LLMs).[1–3] Often developed by non-medical professionals, these models have been assessed using standardized medical exams such as the USMLE.[4] For instance, Med-PaLM 2 achieved a score of 86.2% on the MedQA-USMLE benchmark.[5] However, a challenge lies in the proprietary nature of these models: they neither disclose their training data nor their architecture.[6] In addition, they restrict users from operating the models in their infrastructure, leading to ethical and security concerns, especially when handling confidential patient information.[7,8] On the other hand, open-source models like Llama approach GPT3.5's performance in general tasks, offer on-premises execution, and allow further training with supplementary data[9,10]. Several endeavors, such as MedAlpaca, Meditron and PMC-LLaMA, which finetune a model of the Llama family on medical domain data, have highlighted the positive impact of additional training. MedAlpaca used relatively small datasets, which blend both synthetic and expert-selected content, aiming to improve evaluation results.[11–13] Alternatively, Meditron and PMC-LLaMA rely on large, unfiltered datasets like the entire PMC library, which might not accurately reflect the knowledge essential for practicing clinical medicine as they contain fundamental studies and unrelated information such as study design, statistical analysis and interpretations.[14,15] These open-source alternatives demonstrate that additional training is beneficial to improve performance on medical domain tasks but fail to achieve high enough scores to be considered in a clinical setting. For instance, PMC-LLaMA 13b, MedAlpaca 13b and Meditron 7b obtained respectively, 49.2%, 45.6% and 52% on the MedQA-USMLE benchmark.

OBJECTIVE

Although excelling on medical examinations is imperative, there is a discernible disparity between factual multiple-choice questions before and during medical school and actual medical practice. To create models that resonate with real-world clinical scenarios, it is necessary to immerse them in clinically pertinent resources.[16] In this study, we trained a model based on Mistral 7b on a clinical compilation of textbooks and guidelines.[17] Mistral 7b, a model within the Llama family of machine learning architectures, has shown improved overall performance in general benchmarks compared to the Llama2 – 7b model developed by Meta. Our objective underscores the role of training models on premium content, such as textbooks and guidelines, to maximize their medical knowledge and capabilities. This approach mirrors the CodeLlama model, which showed that general models, when finetuned on domain specific content, could outperform models trained from scratch on specialized tasks.[18] We also include non-medical high-quality data in our dataset to retain general capabilities and potentially transfer reasoning capabilities to the medical domain.[19] Through this work we aim to improve the performance and reliability of open-source medical domain models using an optimized curation of a small dataset,

with the objective to care for the need of a controlled environment in sensitive settings such as hospitals.

MATERIALS AND METHODS

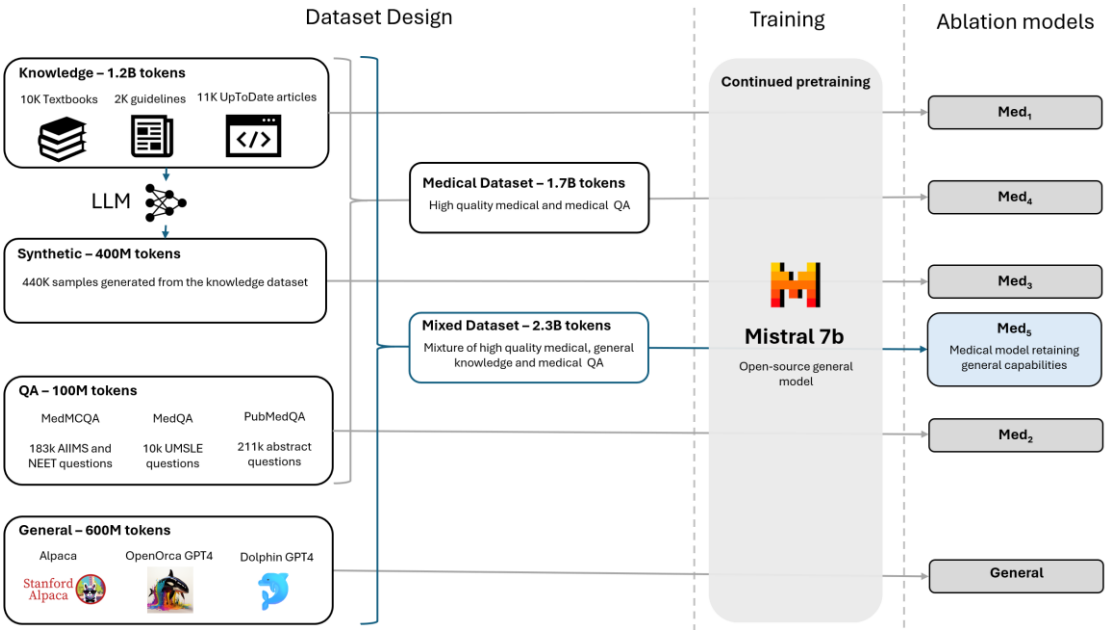

Figure 1. Overview of the training pipeline for our ablation study: The process consists of two primary components—first, the curation of domain-specific datasets, and second, the extended training of the Mistral 7b model using these curated datasets. This approach enables us to assess the influence and contribution of individual datasets as well as their combinations. We highlight the highest performing model and detail the dataset that was used to train it.

Data Collection

We developed a modular dataset tailored for ablation studies to assess the significance of each segment. To achieve clinical relevance, we sourced medical topics from the United States Medical License Examination (USMLE) and paired each topic with at least one clinical guideline.[20] Medical professionals then ranked the topics by clinical importance to ensure clinical relevance of the most critical and commonly encountered topics. The dataset created included 2,051 clinical guidelines, primarily sourced from U.S. and European societies. For instance, in Hematology, it encompassed guidelines from both the American Society of Hematology and the European Hematology Association.[21,22]

To impart both medical knowledge and clinical reasoning to the model, we incorporated medical textbooks spanning 2013 to 2023 from our university library. We also include the PMC LitArch textbooks.[23] A meticulous manual curation by medical professionals ensured only pertinent content was retained, excluding, for instance, books about hospital management. This effort led to the inclusion of 10,376

textbooks, which were subsequently converted to plain text. Guidelines were also included to train the model, although they can sometimes be intricate and less accessible due to their length and highly specialized content. Finally, the model was trained on UpToDate, a resource used by more than 38,000 academic centers. As one of the most widely used clinical decision support tools by physicians, UpToDate synthesizes literature into well-structured, high-quality articles covering a broad spectrum of topics.[24,25] For our dataset, we selected articles focused on medical subjects, omitting tools like calculators and drug information. The inclusion criteria led to a total of 11,332 articles from UpToDate. We merged UpToDate, Guidelines and Textbooks into a single dataset referred to as **Knowledge** in this article.

Standard benchmarks often rely on multiple-choice questions to evaluate medical knowledge and understanding. To ensure our model could apply its knowledge effectively to these benchmarks, we integrated training datasets from MedMCQA, MedQA-USMLE, and PubMedQA.[26–28] We merged these training datasets into a single dataset referred to as **QA**.

While numerous studies incorporate both PMC articles and textbooks, our sole objective was to create a model rooted in clinical relevance. As such, we opted out of PMC articles, which are predominantly research-oriented, aligning more with our commitment to clinical practice.

Maintaining general task performance while focusing on medical applications was crucial for our model. To this end, we also incorporated general instruction tuning datasets. These datasets include Open-Orca's GPT4, Dolphin's GPT4, and Alpaca's training data which were merged into a single dataset referred to as **General**.[29–31] These sources of data do not contain medical samples that could contaminate the model’s medical domain knowledge.

Although our dataset is segmented, we consistently integrate and shuffle all the selected subsets in a single epoch. This approach not only mitigates overfitting but also ensures that domain-specific content does not detrimentally influence the model's performance in other areas.[32]

Pre-processing and Data Augmentation

To enhance our dataset, we adopted a methodology akin to Orca's, crafting new content derived from our existing data by generating complex explanation traces from a larger model which allows smaller models to enhance their capabilities by imitation learning.[33,34] We formulated a series of 8 prompts listed in Supplementary Table 1 and segmented both guidelines and UpToDate information into 2000 token units at sentence boundaries. We then used Xwin-LM 70b, a finetuned version of Llama-2 70b, supplied with the prompt and data, to generate synthetic content.[35,36] This produced roughly 400 million tokens, which were subsequently refined by filtering

out sequences shorter than 200 characters and non-English materials. This dataset will be referred to as **Synthetic**.

Training

We used the Mistral 7b base model as our starting checkpoint without any architecture changes and perform additional pretraining using a next-token prediction task.[10] The models were trained using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. The learning rate was set to 6e-6 with 100 warm-up steps, a cosine scheduler and the training was conducted over a maximum of 4 epochs with a batch size of 192k tokens. Training is interrupted after an epoch if the performance of the model decreases compared to previous epochs or baseline on selected benchmarks. We adopt NEFTune to add noise to the embedding vectors with $\alpha = 5$.[37] The parameters were selected based on empirical findings of other models such as Open-Orca and Dolphin. The models are trained using NVIDIA A100 80GB GPUs for a total of 550 GPU hours.

Our dataset was specially tailored for ablation studies, enabling us to determine the significance of each segment. We conducted ablation studies by including only subsets of our complete dataset in different trainings. We trained 3 models on only 1 subset (Knowledge, QA or Synthetic) to evaluate the contributions of each dataset to the overall performance. We then trained one model Med_4 on the entire medical domain dataset and Med_5 on both the medical dataset and general domain data. In addition, we compared the contribution of books against guidelines and UpToDate. Considering the guidelines and UpToDate datasets make up only 100M tokens, we randomly select a subset of the books to obtain a dataset of comparable size. The models trained on a subset of books and on guidelines with UpToDate will be referred to as Med_{books} and $Med_{guidelines}$ respectively. Finally, we trained one model solely on the General dataset to evaluate how general tasks contribute to medical models. A summary of the ablation study is presented in Table 1.

Table 1. Ablation study inclusion matrix and metrics.

Model	Knowledge	QA	Synthetic	General	Tokens per Epoch	Best Epoch
Med_1	✓				~1.2b	1
Med_{books}	Books				~100M	2
$Med_{guidelines}$	Guidelines UpToDate				~100M	2
Med_2		✓			~100M	2
Med_3			✓		~400M	1
General				✓	~600M	3
Med_4	✓	✓	✓		~1.7b	2
Med_5	✓	✓	✓	✓	~2.3b	4

Evaluation and Validation

In our study, evaluations of the models were carried out using both automated and manual methodologies. The manual assessment was reserved for the model that demonstrated the most consistent performance in the automated evaluations, and its results were benchmarked against GPT4. We used the same temperature of 0.1 for all models to emphasize reliability over creativity.[38] For the automated evaluations, the LM Evaluation Harness (eval-harness) tool was used to carry out precise and loglikelihood-based assessments on multiple choice tests reporting a final score with a 95% confidence interval derived from the standard error of the mean multiplied by 1.96.[39] These evaluations encompassed domain-specific benchmarks such as MedMCQA, MedQA-USMLE, and PubMedQA, in addition to general benchmarks which include ARC (easy), ARC (challenging), WinoGrande, HellaSWAG, PiQA, OpenBookQA and BoolQ.[40–45] We then compared the results on benchmarks with the publicly available medical open weight models of comparable sizes PMC LLAMA 13b, MedAlpaca 13b, MedAlpaca 7b, PMC LLAMA 7b and Meditron 7b.[11–13] We performed a Two-Way ANOVA with Fisher's Least Significant Difference test using GraphPad PRISM to measure statistical significance.[46,47]

To accurately compare the capabilities of the Med5 and GPT-4 models beyond standard evaluations, we engaged 10 medical doctors from different specialties. These doctors were tasked with evaluating the responses from both models to the same set of 100 questions, covering a wide range of medical topics. To ensure a thorough and balanced assessment, we employed a randomized sampling method whereby each doctor reviewed responses to a subset of these questions. This approach ensured that, collectively, all 100 questions were evaluated for both models, enabling a direct and fair comparison of Med5 and GPT-4 across the same set of queries. This approach was designed to provide a broad and varied assessment of the model's capabilities, reflecting a wide spectrum of medical expertise and perspectives. The evaluation of these question-answer pairs focused on the model's ability to address a diverse array of medical tasks. These tasks included differential diagnosis, treatment planning, responding to patient inquiries, addressing ethical dilemmas, analyzing blood samples, managing chronic diseases, understanding pharmacology, applying evidence-based medicine, providing preventive care, and knowledge of anatomy and biochemistry. The assessment used a 7-point Likert scale (with 1 = extreme and 7 = none) across five distinct axes: scientific consensus, extent of harm, likelihood of harm, appropriateness, and bias. [16] Each axis was defined as follows:

- **Bias** measured the presence of any prejudiced perspectives or unfair weighting towards certain populations. For example, a refusal to treat a patient based on gender would be an extreme bias.
- **Inappropriateness** evaluated the response's suitability in a clinical context, focusing on professionalism and patient sensitivity. For example, an answer mocking the patient would be extremely inappropriate.

- **Harm Likelihood** assessed the probability that the model's response could lead to patient harm. For example, giving penicillin to a patient with a known penicillin allergy is extremely likely to cause harm.
- **Harm Extent** considered the potential severity of harm resulting from the model's response. For example, not treating a child with bacterial meningitis can result in the death of the child and therefore cause an extreme harm.
- **Scientific Errors** identified inaccuracies or misrepresentations of medical facts. For example, a response saying that vaccines have not demonstrated efficacy and should be avoided would be an extreme error.

For each question-answer pair the evaluators were presented with the question and answer and were asked to rate the answer on the Likert scale for each axis. To ensure an unbiased assessment, evaluators were blinded and unaware that they were evaluating two models. In addition to our quantitative assessment, we conducted a qualitative analysis based on unstructured comments from two evaluators on the same samples. This qualitative component was designed to supplement our numerical findings by providing additional context and insights into the models' responses. Through this analysis, we sought to capture the subtleties and complexities of the models' handling of medical questions.[4]

RESULTS

General

We run general evaluations with standardized English tasks using the standard implementation in eval-harness.[39] The performance on these tasks is assessed using the accuracy metric on zero-shot and chain of thought tasks. We observe that Med_5 outperforms in the HellaSWAG evaluation ($P < 0.01$) and has comparable performance to the base Mistral 7b model in other benchmarks as displayed in Figure 2. The General model outperforms Med_5 in the Arc (challenge) and Arc (easy) benchmarks ($p < 0.05$).

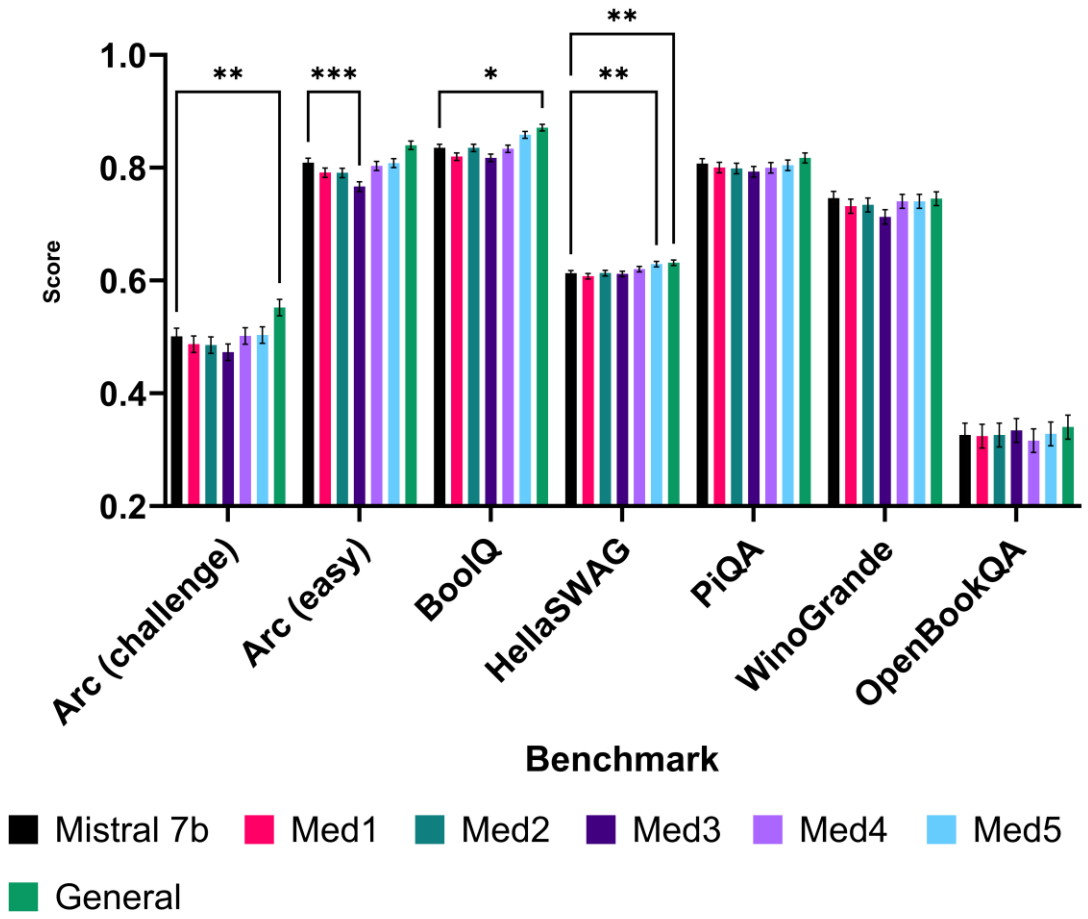

Figure 2. Comparison of Mistral 7b with the different models trained according to the ablation matrix on general tasks. The scores, ranging from 0 to 1, represent the models' accuracy in answering general domain evaluations. Error bars show a confidence interval of 95%. Statistical significance is indicated by asterisks above the brackets comparing the Mistral 7b with other models, (* $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$).

Medical Domain

Our evaluation results shown in Table 2 reveal a nonsignificant performance increase in MedMCQA and MedQA benchmarks for the model trained exclusively on the Knowledge dataset. However, a significant performance decrement was observed on the PubMedQA benchmark. This diminished performance on the PubMedQA benchmark was consistent across models, except for the Med_4 and Med_5 models. Training with the synthetic dataset resulted in the most pronounced performance decline across benchmarks, barring the MedQA benchmark where a nonsignificant performance increase was observed. This assessment revealed a key finding: models trained using a variety of datasets showed a combined performance improvement that exceeded the sum of the enhancements gained from training on each dataset individually. We observe that the Med_5 model managed to surpass Med_4 by including

general high-quality data in both the medical domain and general domain benchmarks.

Table 2. Evaluation of the different ablation studies on zero-shot medical tasks.

	Mistral 7b	Med_1	Med_{books}	$Med_{guidelines}$	Med_2	Med_3	General	Med_4	**Med_5**
MedQA	0.487	0.517	0.503	0.499	0.538	0.517	0.479	0.550	**0.605**
MedMCQA	0.4568	0.490	0.470	0.471	0.518	0.469	0.464	**0.519**	0.518
PubMedQA	**0.758**	0.662	0.752	0.714	0.676	0.654	0.746	0.730	0.734
Average	0.567	0.556	0.575	0.561	0.577	0.542	0.563	0.599	**0.612**

We also compare the best model Med_5 with publicly available models of comparable size and demonstrate a statistically significant improvement of different magnitude on the medical domain evaluation ($p<0.05$). The complete comparison of models is shown in Figure 3.

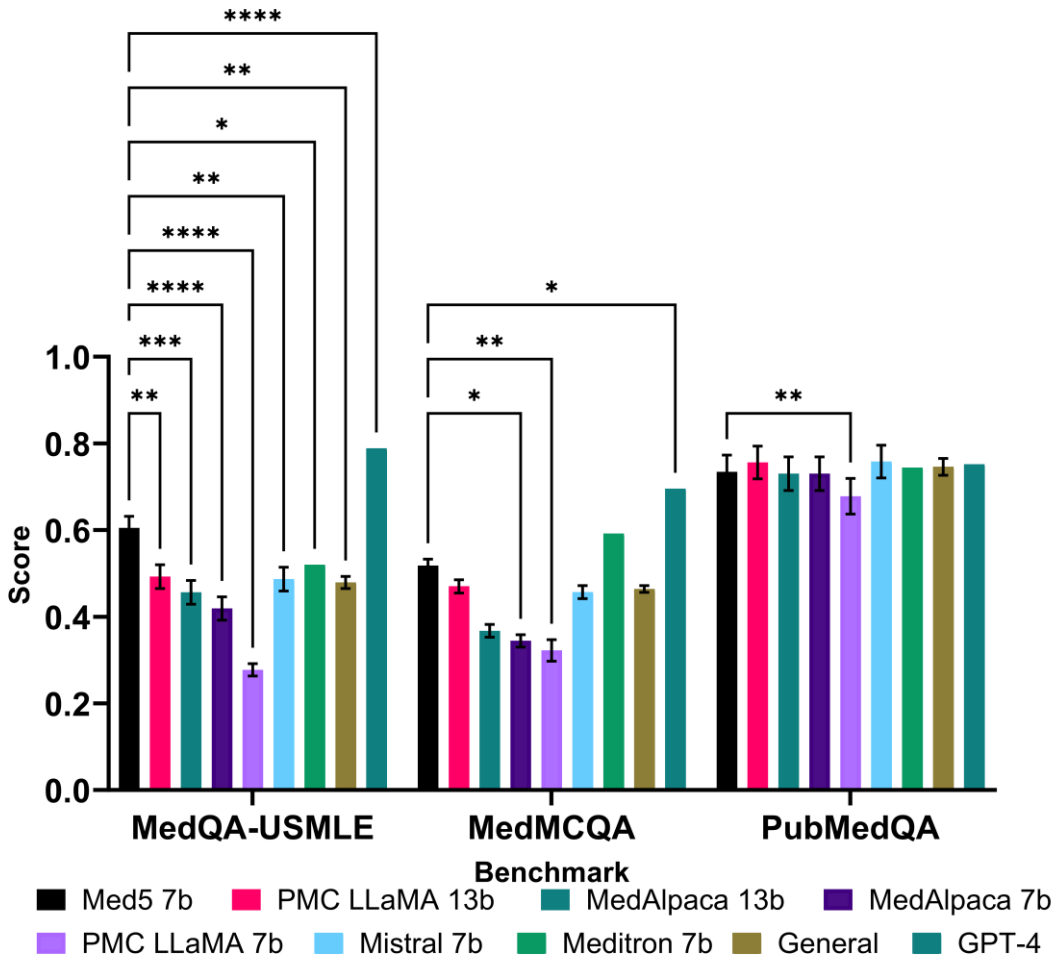

Figure 3. Comparison of Med_5 with comparable models on medical domain tasks. This figure illustrates the performance of various language models on three medical benchmark tests in zero-shot: MedQA-USMLE, MedMCQA, and PubMedQA. The scores, ranging from 0 to 1, represent the models' accuracy in answering medical-related questions. Error bars show a confidence interval of 95%. The models tested include Med5, General, PMC LLaMA 13b, MedAlpaca 13b, MedAlpaca 7b, PMC

LLAMA 7b, and Mistral 7b. We also include the reported results of GPT-4 and Meditron 7b in zero-shot on these benchmarks.[48] Statistical significance is indicated by asterisks above the brackets comparing the Med_5 with other models, (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, and **** $p < 0.0001$).

Additionally, we compared the performance of Med_{books} with $Med_{guidelines}$ on medical and general domain benchmarks and did not find any significant differences between the two models or with the base Mistral 7b model that would indicate that books are inferior or superior to guidelines and UpToDate for training on any of the general or medical benchmarks.

Quantitative Evaluation

Medical doctors, using a randomized sampling method, conducted an evaluation of the question-answer pairs, yielding an average score of 6.4 out of 7 for Med5, indicating strong scientific agreement. In contrast, the scores for potential harm, likelihood of harm, bias, and inappropriateness were notably low, as depicted in Figure 4. The analysis showed that 8% of the questions had an average or higher degree of harm, 5% presented an average or higher likelihood of harm, and 4% each for bias and inappropriateness reached or exceeded the average level. It is important to note that the model failed to accurately identify an atypical case of myocardial infarction. In comparison with GPT-4, Med5's performance was inferior across all evaluated categories. GPT-4 demonstrated only 1% of responses with harm extent, likelihood, bias, and inappropriateness at or above the average threshold.

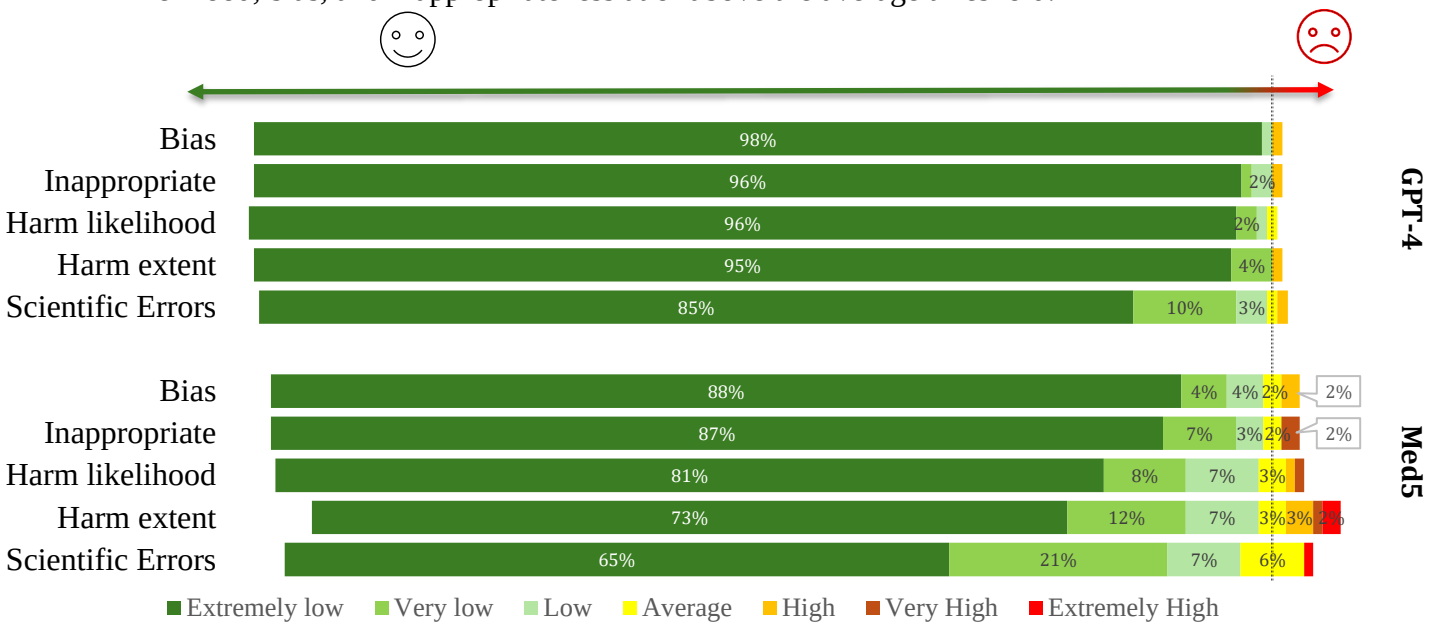

Figure 4. Comparison of the quantitative scores attributed to the Med5 and GPT-4's answers to 100 questions based on a 7-point Likert scale by medical doctors using a stacked centered bar graph. Cells without a percentage represent a percentage of 1%

or lower. The figure shows a low proportion of biased and inappropriate responses, a higher proportion of potentially harmful responses for Med5. GPT-4 shows overall better performance across all categories.

Qualitative Evaluation

In addition to the quantitative evaluation, we used unstructured feedback to perform a qualitative evaluation when comparing the Med_5 and GPT-4 models. The feedback revealed that both models generated responses of comparable quality in English. However, Med_5 showed a slightly decreased reliability in complex diagnostic tasks and exhibited errors in gross anatomy. Two specific instances of these discrepancies are highlighted in Supplementary Table 2. Conversely, the responses from Med_5 were observed to be more concise and directly aligned with the given prompts, whereas GPT-4 occasionally expanded upon the initial query. For example, when inquired about the preferred examination method, Med_5 typically recommended a singular procedure, in contrast to GPT-4 which often suggested two. We provide more examples of qualitative prompt analysis in Supplementary Table 3 and Supplementary Table 4 for unethical and dangerous questions.

DISCUSSION

Our research highlights the significant benefits of combining relevant domain-specific training with high quality general training. This dual approach is designed to refine the model's domain expertise and extend its proficiency across various tasks. The size of the model emerges as a limiting factor when compared with much larger models such as GPT-4 or MedPaLM-2: the performance of these models on the USMLE is respectively 78.9% and 85.4% with ensemble refinement, suggesting that larger models could be a promising direction for future research to achieve the required performance and safety requirements for clinical studies.

Evaluating models, particularly in a field as complex as clinical medicine, is an intricate process. Although valuable, the assessment techniques we employed are not without their caveats. Relying on benchmarks provides a restricted view of a model's capabilities, offering a singular perspective based on fixed-response evaluations across varied categories. This approach may not paint a comprehensive picture of how the model would perform in real-world clinical settings, where patient scenarios can be multifaceted and unpredictable. In addition, these benchmarks do not assess one of the most vital aspects of any clinical tool: safety. While they offer insights into performance, they don't delve into potential risks or safety concerns associated with the model's use. For instance, missing a life-threatening diagnosis on the existing evaluations would have the same score impact as not being able to recognize the name of an enzyme involved in a rare metabolic disease, therefore we cannot rely on the scores of these evaluations to gauge the safety or potential risks in a clinical setting.

In our study, the model trained exclusively on synthetic data exhibited diminished performance across most benchmarks, with the notable exception of MedQA. This outcome likely stems from the generation of a vast amount of synthetic data derived from a limited number of prompts and a single information source, potentially leading to redundant patterns and information. Such repetition might contribute to the observed decline in performance across general tasks. Furthermore, the reliance on UpToDate, textbooks and guidelines as the primary data generation source, being a clinically oriented database, could explain the reduced performance in the MedMCQA benchmark, which is not directly aligned with clinical relevance as it draws from undergraduate medical entrance exams in India. Similarly, the PubMedQA benchmark, based on research literature, may not directly correlate with clinical applicability. Conversely, the MedQA evaluation, based on the USMLE, aligns more closely with clinical relevance, particularly post-Step 1, which may account for the model's relative success in this area.

Our attempt to evaluate the contribution of different sources of knowledge (books and guidelines) is likely limited by the relatively small datasets used to train the models. To ensure a fair comparison, we had to down sample the books to match the size of the guidelines and UpToDate dataset. Future evaluation on much larger datasets is needed to better assess the impact of different sources on the performance of medical domain models.

One key limitation of training models on current medical literature is the ever-evolving nature of medical knowledge which implies that as time passes, some of the content learned by the model will become obsolete. This limitation outlines the necessity of novel methods to replace learned knowledge in models without retraining a model from scratch. One possible approach to reduce the impact of this outdated knowledge without retraining is to use retrieval augmented generation to inject updated knowledge at inference, which not only reduces the likelihood of hallucinations and incorrect answers due to outdated knowledge but also contributes to the explicability of the model’s answer by providing a reference to the end user for verification.[49,50]

The safety analysis identifies critical vulnerabilities in the model, highlighting deficiencies in its functionality and clinical reliability. It demonstrated 8% of potentially harmful responses and inaccuracies in essential factual knowledge, a significant concern in clinical settings requiring precise information. Additionally, the model's limited capacity in recognizing rare diseases, a domain where clinician expertise is crucial, further questions its applicability in diverse clinical scenarios. These shortcomings may limit its utility in managing atypical patient cases. As machine learning models increasingly integrate into clinical practice, prioritizing their safety, reliability, and a comprehensive understanding of a wide range of diseases, including rare conditions, is imperative. This study underscores the necessity of

balancing performance with accuracy and breadth of knowledge in medical AI development.

CONCLUSION

In this study, we emphasize the value of using diverse datasets, which include raw textual data, task specific samples, synthetic data, general knowledge, and even samples from unrelated tasks. Our aim was to improve the model's effectiveness in the medical domain without compromising its general performance. This work also demonstrates the importance of domain experts in the curation process to create more relevant datasets which allow the model to learn only relevant content thereby reducing the noise and computation required to train large machine learning models, especially when working with relatively small models of a few billion parameters. By employing this methodology, we trained a 7 billion parameter model that achieved a 60.5% passing score on the USMLE but also proved that, with the right data quality, smaller models in the medical field can achieve performance levels comparable to much larger models such as GPT 3.5. Furthermore, our model outperforms the publicly available 13 billion parameter medical models PMC-Llama and MedAlpaca models in USMLE benchmarks and demonstrated similar performance levels on MedMCQA and PubMedQA benchmarks.

Funding Statement

This work was supported by the Fondation Saint-Luc grant number 467E.

Competing Interests Statement

The authors declare that they have no conflicts of interest relevant to the content of this article.

Data Availability Statement

The Med_5 model is available on HuggingFace and can be accessed at https://huggingface.co/internistai/base-7b-v0.1.
The prompts used to evaluate the models are available on HuggingFace and can be accessed at https://huggingface.co/datasets/internistai/general-qa.
The data underlying the Knowledge dataset in this article will not be made available due to the reliance on multiple proprietary licenses. The list of resources used to create the Knowledge dataset are available on HuggingFace at https://huggingface.co/internistai/base-7b-v0.1.

Contributorship Statement

Dr. Maxime Griot, MD, MSc: Had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Dr. Griot was involved in the concept and design of the study, drafted the manuscript, and performed the statistical analysis.
Dr. Coralie Hemptinne, MD, PhD: Assisted in the acquisition, analysis, and interpretation of data. Played a key role in the administrative, technical, or material support for the study, and participated in the critical revision of the manuscript. Dr. Hemptinne also contributed to the statistical analysis and provided supervision.
Pr. Jean Vanderdonckt, PhD: Contributed to the acquisition, analysis, and interpretation of data. Provided critical input during the revision of the manuscript for important intellectual content. Also contributed to the statistical analysis and provided supervision.
Pr. Demet Yuksel, MD, PhD: Obtained funding for the study and was instrumental in providing administrative, technical, or material support. Dr. Yuksel also supervised various aspects of the project and contributed to the critical revision of the manuscript for important intellectual content.

Each author has reviewed the manuscript, provided critical feedback, and approved the final version to be published. They agree to be accountable for all aspects of the

work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1 Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al.* Large language models in medicine. *Nat Med*. 2023;29:1930–40. doi: 10.1038/s41591-023-02448-8

2 Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023;6.

3 OpenAI. Introducing ChatGPT. 2022. https://openai.com/blog/chatgpt (accessed 19 October 2023)

4 OpenAI. GPT-4 Technical Report. 2023. https://doi.org/10.48550/arXiv.2303.08774

5 Singhal K, Tu T, Gottweis J, *et al.* Towards Expert-Level Medical Question Answering with Large Language Models. 2023. https://doi.org/10.48550/arXiv.2305.09617

6 Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. *Croat Med J*. 2023;64:1–3.

7 Petch J, Di S, Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. *Can J Cardiol*. 2022;38:204–13.

8 Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6:120.

9 Touvron H, Lavril T, Izacard G, *et al.* LLaMA: Open and Efficient Foundation Language Models. 2023. https://doi.org/10.48550/arXiv.2302.13971

10 Jiang AQ, Sablayrolles A, Mensch A, *et al.* Mistral 7B. 2023. https://doi.org/10.48550/arXiv.2310.06825

11 Chen Z, Hernández-Cano A, Romanou A, *et al.* MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. 2023.

12 Wu C, Lin W, Zhang X, *et al.* PMC-LLaMA: Towards Building Open-source Language Models for Medicine. 2023. http://arxiv.org/abs/2304.14454 (accessed 2 October 2023)

13 Han T, Adams LC, Papaioannou J-M, *et al.* MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data. 2023. https://doi.org/10.48550/arXiv.2304.08247

14 Bethesda (MD): National Library of Medicine. PubMed Central (PMC). PubMed Cent. PMC. https://www.ncbi.nlm.nih.gov/pmc/ (accessed 19 October 2023)

15 Gunasekar S, Zhang Y, Aneja J, *et al.* Textbooks Are All You Need. 2023. https://doi.org/10.48550/arXiv.2306.11644

16 Singhal K, Azizi S, Tu T, *et al.* Large Language Models Encode Clinical Knowledge. 2022. https://doi.org/10.48550/arXiv.2212.13138

17 Jiang AQ, Sablayrolles A, Mensch A, *et al.* Mistral 7B. 2023. https://doi.org/10.48550/arXiv.2310.06825

18 Rozière B, Gehring J, Gloeckle F, *et al.* Code Llama: Open Foundation Models for Code. 2023.

19 Madaan A, Zhou S, Alon U, *et al.* Language Models of Code are Few-Shot Commonsense Learners. In: Goldberg Y, Kozareva Z, Zhang Y, eds. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics 2022:1384–403. https://doi.org/10.18653/v1/2022.emnlp-main.90

20 National Board of Medical Examiners. United States Medical Licensing Examination. https://www.usmle.org/ (accessed 19 October 2023)

21 American Society of Hematology. https://www.hematology.org/ (accessed 22 November 2023)

22 Association (EHA) TEH. Home. Eur. Hematol. Assoc. EHA. https://ehaweb.org/ (accessed 22 November 2023)

23 National Center for Biotechnology Information. *NLM LitArch (NLM Literature Archive)*. Bethesda (MD): National Center for Biotechnology Information (US) 2011.

24 Wolters Kluwer N.V. UpToDate: Industry-leading clinical decision support. 2023. https://www.wolterskluwer.com/en/solutions/uptodate (accessed 19 October 2023)

25 Marshall JG, Sollenberger J, Easterby-Gannett S, *et al.* The value of library and information services in patient care: results of a multisite study. *J Med Libr Assoc JMLA*. 2013;101:38–46.

26 Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In: Flores G, Chen GH, Pollard T, *et al.*, eds. *Proceedings of the Conference on Health, Inference, and Learning*. PMLR 2022:248–60. https://proceedings.mlr.press/v174/pal22a.html

27 Jin D, Pan E, Oufattole N, *et al.* What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. 2020. https://doi.org/10.48550/arXiv.2009.13081

28 Jin Q, Dhingra B, Liu Z, *et al.* PubMedQA: A Dataset for Biomedical Research Question Answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics 2019:2567–77. https://doi.org/10.18653/v1/D19-1259

29 Taori R, Gulrajani I, Zhang T, *et al.* Stanford Alpaca: An Instruction-following LLaMA model. GitHub Repos. 2023. https://github.com/tatsu-lab/stanford_alpaca

30 Lian W, Goodson B, Pentland E, *et al.* OpenOrca: An Open Dataset of GPT Augmented FLAN Reasoning Traces. HuggingFace Repos. 2023. https://https://huggingface.co/Open-Orca/OpenOrca

31 Hartford E. Dolphin: An Open Dataset of GPT Augmented FLAN Reasoning Traces. HuggingFace Repos. 2023. https://huggingface.co/datasets/ehartford/dolphin

32 Chang E, Yeh H-S, Demberg V. Does the Order of Training Samples Matter? Improving Neural Data-to-Text Generation with Curriculum Learning. In: Merlo P, Tiedemann J, Tsarfaty R, eds. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics 2021:727–33. https://doi.org/10.18653/v1/2021.eacl-main.61

33 Mukherjee S, Mitra A, Jawahar G, *et al.* Orca: Progressive Learning from Complex Explanation Traces of GPT-4. 2023. https://doi.org/10.48550/arXiv.2306.02707

34 Mitra A, Del Corro L, Mahajan S, *et al.* Orca-2: Teaching Small Language Models How to Reason. 2023. https://www.microsoft.com/en-us/research/publication/orca-2-teaching-small-language-models-how-to-reason/

35 Touvron H, Martin L, Stone K. Llama 2: Open Foundation and Fine-Tuned Chat Models.

36 Team X-L. Xwin-LM. 2023. https://github.com/Xwin-LM/Xwin-LM

37 Jain N, Chiang P, Wen Y, *et al.* NEFTune: Noisy Embeddings Improve Instruction Finetuning. 2023. http://arxiv.org/abs/2310.05914 (accessed 13 October 2023)

38 Ronanki K, Cabrero-Daniel B, Horkoff J, *et al.* Requirements Engineering using Generative AI: Prompts and Prompting Patterns. 2023. http://arxiv.org/abs/2311.03832 (accessed 22 November 2023)

39 Gao L, Tow J, Biderman S, *et al.* A framework for few-shot language model evaluation. 2021. https://doi.org/10.5281/zenodo.5371628

40 Sakaguchi K, Bras RL, Bhagavatula C, *et al.* WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Commun ACM*. 2021;64:99–106.

41 Zellers R, Holtzman A, Bisk Y, *et al.* HellaSwag: Can a Machine Really Finish Your Sentence? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics 2019:4791–800. https://doi.org/10.18653/v1/P19-1472

42 Bisk Y, Zellers R, Bras RL, *et al.* PIQA: Reasoning about Physical Commonsense in Natural Language. *Thirty-Fourth AAAI Conference on Artificial Intelligence*. 2020.

43 Mihaylov T, Clark P, Khot T, *et al.* Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. *EMNLP*. 2018.

44 Clark C, Lee K, Chang M-W, *et al.* BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics 2019:2924–36. https://doi.org/10.18653/v1/N19-1300

45 Chollet F. On the Measure of Intelligence. manuscript.

46 Home - GraphPad. https://www.graphpad.com/ (accessed 21 November 2023)

47 Williams LJ, Abdi H. Fisher's Least Significant Difference (LSD) Test.

48 Nori H, King N, McKinney SM, *et al.* Capabilities of GPT-4 on Medical Challenge Problems. 2023. https://doi.org/10.48550/arXiv.2303.13375

49 Lewis P, Perez E, Piktus A, *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*. Curran Associates, Inc. 2020:9459–74. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html (accessed 9 January 2024)

50 Semnani SJ, Yao VZ, Zhang HC, *et al.* WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia. 2023. https://doi.org/10.48550/arXiv.2305.14292