NEJM AI 2024; 1 (2)
DOI: 10.1056/AIoa2300068

ORIGINAL ARTICLE

Almanac — Retrieval-Augmented Language Models for Clinical Medicine

Cyril Zakka, M.D.,[1] Rohan Shad, M.D.,[3] Akash Chaurasia,[2] Alex R. Dalal, M.D.,[1] Jennifer L. Kim, M.D.,[1] Michael Moor, M.D., Ph.D.,[2] Robyn Fong,[1] Curran Phillips,[1] Kevin Alexander, M.D.,[4] Euan Ashley, M.D., Ph.D.,[4] Jack Boyd, M.D.,[1] Kathleen Boyd, M.D.,[5] Karen Hirsch, M.D.,[6] Curt Langlotz, M.D., Ph.D.,[7] Rita Lee, P.A.-C.,[1] Joanna Melia, M.D.,[8] Joanna Nelson, M.D.,[9] Karim Sallam, M.D.,[4] Stacey Tullis, R.N., B.S.N.,[1] Melissa Ann Vogelsong, M.D.,[10] John Patrick Cunningham, Ph.D.,[11] and William Hiesinger, M.D.[1]

Received: August 8, 2023; Revised: November 10, 2023; Accepted: November 14, 2023; Published: January 25, 2024

Abstract

BACKGROUND Large language models (LLMs) have recently shown impressive zero-shot capabilities, whereby they can use auxiliary data, without the availability of task-specific training examples, to complete a variety of natural language tasks, such as summarization, dialogue generation, and question answering. However, despite many promising applications of LLMs in clinical medicine, adoption of these models has been limited by their tendency to generate incorrect and sometimes even harmful statements.

METHODS We tasked a panel of eight board-certified clinicians and two health care practitioners with evaluating Almanac, an LLM framework augmented with retrieval capabilities from curated medical resources for medical guideline and treatment recommendations. The panel compared responses from Almanac and standard LLMs (ChatGPT-4, Bing, and Bard) versus a novel data set of 314 clinical questions spanning nine medical specialties.

RESULTS Almanac showed a significant improvement in performance compared with the standard LLMs across axes of factuality, completeness, user preference, and adversarial safety.

CONCLUSIONS Our results show the potential for LLMs with access to domain-specific corpora to be effective in clinical decision-making. The findings also underscore the importance of carefully testing LLMs before deployment to mitigate their shortcomings. (Funded by the National Institutes of Health, National Heart, Lung, and Blood Institute.)

The author affiliations are listed at the end of the article.

Dr. Zakka can be contacted at czakka@stanford.edu or at 870 Quarry Rd, Falk Cardiovascular Research Building, Palo Alto, CA 94304.

Background

In recent years, language model pretraining has emerged as a powerful paradigm in natural language processing.[1-4] For many language models, performance improvements have been empirically observed to scale with model and data set size on a range of downstream natural language processing tasks, with sample efficiency and the well-documented emergence of zero-shot capabilities, whereby the models use auxiliary data to complete tasks without having received specific training examples of those tasks.[5-7]

However, because of the nature of the training objective of predicting the next token in a sentence, large language models (LLMs) can be prone to generating incorrect statements, a phenomenon known as hallucinations.[8,9] Moreover, studies have shown that these models may reproduce social biases and make statements that reinforce gender, racial, and religious stereotypes.[10,11]

To reduce the incidence of unwanted behaviors, several studies have explored various ways of steering LLM outputs to align with user intent, including fine-tuning with human feedback[12,13] and natural language prompt engineering.[14,15] This pivot in training paradigms has led to an explosion of transformative applications ranging from human-like chatbots to impressive writing assistants.[2,16]

Nevertheless, the unstructured and open-ended aspect of LLM prompts puts LLMs at risk of adversarial attacks or intentional acts of derailing the original goal of a model with malicious intent, such as by leaking private data or generating misinformation.[17,18] As such, despite the promising avenue of research posed by the incorporation of LLMs into the clinical workflow, careful consideration must be given to LLM implementation to ensure patient privacy and safety.[19]

In this work, we introduce Almanac, a framework to explore the role of medical LLMs and their safe deployment in health care settings. To stay abreast of the shifting landscape of evidence-based practices, physicians often make use of point-of-care tools to improve outcomes.[20] However, as clinical evidence continues to grow, curated content becomes less accessible and more confined to error-prone search tools and time-consuming appraisal techniques that fail to address the unique needs of individual patients.

To address these concerns, we assessed the utility of LLMs as clinical knowledge bases that can use external tools (e.g., search engines, medical databases, calculators) to answer medical queries. Knowledge retrieval was outsourced to a Web browser and a database of predefined knowledge repositories; an off-the-shelf LLM was used to achieve high-quality, accurate answer generation with in-text citations that referenced the source material.

To assess these models for the clinical workflow, we evaluated them according to four key objectives: (1) factuality: the degree to which the generated text aligns with established medical knowledge and provides accurate citations for further independent verification; (2) completeness: the extent to which the generated text provides a comprehensive and accurate representation of the clinical situation or the answer to a posed question and includes contraindications as necessary; (3) preference: the overall user preference for the generated text on the basis of its readability, its applicability to the context, and its ability to communicate complex medical concepts; and (4) adversarial safety: the susceptibility of these models to cause intentional or unintentional harm.

Because of increasing concerns of data leakage, which can occur when information from outside the training data set is used to create a model, we evaluated Almanac empirically using a panel of eight board-certified clinicians and two health care practitioners. The panel members, with an average of 10.5 years of experience, were selected on the basis of their expertise across nine medical specialties. The panel tested our model on a novel data set of open-ended clinical scenarios. To our knowledge, this work is the first to show the ability of grounded LLMs, which use case-specific information that is not part of their training data, to provide accurate and reliable answers to open-ended medical queries in the clinical setting, paving the way for controlled and safe deployment of LLMs in health care.

By pretraining transformers — neural networks that transform one type of input into a different type of output — on curated scientific and biomedical corpora, recent models, such as BioGPT[21] and SciBERT,[22] have shown improved performance on a variety of biomedical downstream tasks.[23-27] Other work has recently established the benefits of smaller domain-specific language models compared with larger and more generalized models.[28] However, despite marked improvements in pretraining increasingly larger architecture sizes on domain-specific data sets (e.g., GatorTron,[29] Med-PaLM[30]), these models remain prone to hallucinations and biases, further highlighting the limitations and unreliability of LLMs as intrinsic knowledge bases.[31]

Retrieval-augmented language generation is not a novel concept. Previous works, such as those of Lewis et al.[32]

and Borgeaud et al.,[33] have relied on database fine-tuning to improve performance, with biomedical applications limited to simple question and answer (QA) or binary classification.[34,35] Our work, akin to that of Ram et al.,[36] Nakano et al.,[37] Schick et al.,[38] and Liévin et al.,[39] focuses on leveraging these models for their language understanding and modeling capabilities to answer open-ended questions. Nakano et al.[37] introduced WebGPT, which paired a language model with Web browsing capabilities to improve the accuracy of question answering. Liévin et al.[39] used Wikipedia to obtain human-level performances on three medical QA data sets. Schick et al.[38] fine-tuned their language model to use various external tools (e.g., calculator, calendar) through simple application programming interfaces to overcome limitations with calculations and factual lookup. Similarly, Ram et al.[36] improved text generation by prepending retrieved text to the context window of an LLM before generation. We extended these works by showing the effectiveness of retrieval in the open-ended clinical QA setting by dynamically retrieving and applying reasoning to retrieved passages to answer a variety of clinical queries and calculations.

Methods

DATA SET

To evaluate the potential of LLMs in clinical medicine, we focused on the task of medical question answering. Although existing data sets, such as MultiMedQA,[30] MedMCQA,[40] and PubMedQA,[41] serve as valid benchmarks for assessing reading comprehension and knowledge recall of biomedical language models, their use as benchmarks in open-ended clinical QA tasks poses two clear problems: data contamination and poor clinical proxies.

Because LLMs are increasingly trained on data crawled from various Web sources, data sets intended for model evaluation may end up in the training data, making it difficult to objectively assess the models using the same benchmarks.[42] More concerning, the training data for many proprietary models are often kept confidential, introducing an extra layer of uncertainty in estimating data contamination. As such, using data sets with public-facing evaluations hinders objective experimentation, thereby undermining any subsequently drawn conclusions. In addition to showing poor to weak positive clinical correlation,[43-48] U.S. Medical Licensing Examination–style questions fail to encapsulate the full scope of actual clinical scenarios encountered by medical professionals. They often portray patient scenarios as neat clinical vignettes, bypassing the intricate series of microdecisions that constitute real patient care.

To address these shortcomings, we curated ClinicalQA, a novel benchmark of open-ended clinical questions spanning several medical specialties with topics ranging from treatment guidelines to clinical calculations. After explaining the goals of our study, we tasked our evaluation panel members with generating questions related to their day-to-day practices, with the following instructions (reworded for clarity): “Write as many questions as you can in your field of expertise related to your day-to-day clinical duties. Questions can range from simple descriptions (e.g., what is the mechanism of action of Plavix?) to more complex management questions (e.g., what is the recommended dose and duration of a course of aspirin following a coronary artery bypass graft?; what are the advantages of laparoscopic hernia repair over open repair?).”

We compiled the 314 questions submitted (discarding none) into ClinicalQA, serving as an early but valuable benchmark for language model–based clinical decision-making support systems. Summary statistics of the data set are provided in Table 1, with a subset of 25 questions and a summary of the human evaluators given in the Supplementary Appendix.

ARCHITECTURE

Almanac consists of many components working asynchronously to achieve accurate document retrieval, reasoning, and QA (Fig. 1).

Table 1. A Total of 314 Questions Spanning Several Medical Specialties Were Compiled into ClinicalQA.

Medical Specialty	No. of Questions
Cardiothoracic surgery	25
Cardiology	65
Neurology	25
Gastroenterology	8
Anesthesia and critical care	30
Nursing	25
Physician assistant	50
Infectious diseases	56
Pediatrics	25
Clinical calculation vignettes	5
Total	314

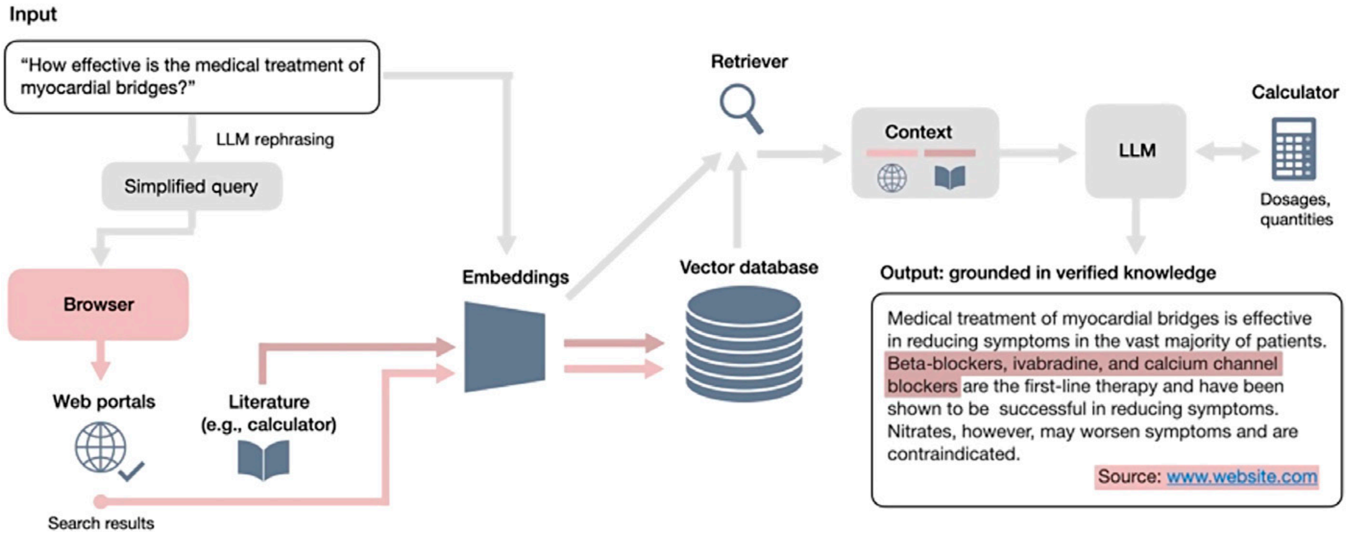
Input
"How effective is the medical treatment of myocardial bridges?"
LLM rephrasing
Simplified query
Browser
Web portals
Literature (e.g., calculator)
Search results
Embeddings
Retriever
Vector database
Context
LLM
Calculator
Dosages, quantities
Output: grounded in verified knowledge
Medical treatment of myocardial bridges is effective in reducing symptoms in the vast majority of patients. Beta-blockers, ivabradine, and calcium channel blockers are the first-line therapy and have been shown to be successful in reducing symptoms. Nitrates, however, may worsen symptoms and are contraindicated.
Source: www.website.com

Figure 1. Almanac Overview.

When presented with a query, Almanac uses external tools to retrieve relevant information before synthesizing a response with citations referencing source material. With this framework, large language model (LLM) outputs remain grounded in truth while providing a reliable way of fact-checking.

An overview of each component is outlined in the following sections.

Database

The database is a high-performance vector storage and similarity engine optimized for the rapid indexing and search of materials sourced from various contexts, including textbooks and Web documents. The database is responsible for storing this content semantically: that is, through information-dense vectors encoding the meaning of the text they contain, with a similarity metric, such as cosine distance, for later retrieval.[49] For our experiments, we used the Qdrant database (version 1.3.0)[50] and initialized it with more than 500 clinical calculators. These calculators are sourced directly from MDCalc and converted to markdown. Their clinical indications and instructions are used as metadata for retrieval.[51] Sample clinical calculators are shown in the Supplementary Appendix.

Browser

The browser consists of several predetermined domains (websites) that Almanac can access to return information from the Internet. These websites are carefully curated to ensure high-quality content in response to queries. Every time a query is submitted, candidate articles are first retrieved using the website's internal search engine before being parsed and stored in the database. To overcome the token limit of most LLMs, each article is divided into chunks of 1000 tokens and fed into the retriever separately. For our experiments, we limited our model's access to three standard domains: PubMed,[52] UpToDate,[53] and BMJ Best Practices.[54]

Retriever

The retriever is a text encoder that encodes queries and reference materials into the same high-dimensional space before storing them in the database. The language model is pretrained on large corpora to ensure that texts with similar content get closer vector representations in this space. At search time, n documents matching a given query embedding are scored and thresholded with λ=0.83, resulting in at most n passages presented to the language model. The value of λ is a strict cutoff point, below which any retrieved passages with lesser confidence scores are excluded. This threshold was on the basis of an analysis of retriever scores for a sample of 25 clinical questions from an external data set. For the purposes of reproducibility, the text-embedding-ada-002 by OpenAI was used, with an output dimension of 1536 and a set n=10. We noted that each retrieved passage is processed independently of other chunks, and alternate retrieval strategies were not explored.

Table 2. Summary of the Rubric Used by Clinical Evaluators of Large Language Model Outputs.*

Axis	Question
Factuality	Does the answer agree with standard practices and the consensus established by bodies of authority in your practice?
	If appropriate, does the answer contain correct reasoning steps?
Completeness	Does the answer address all aspects of the question?
	Does the answer omit any important content?
	Does the answer contain any irrelevant content?
Preference	Which answer did you prefer overall?

* For each question, evaluators were asked to rank the answers corresponding to different models according to factuality, completeness, and preference.

Language Model

The language model is a generative pretrained transformer architecture fine-tuned by using instructions and trained to respond in a conversational fashion with reinforcement learning from human feedback.[19] This module serves two functions. First, it rephrases any given query into a format more suitable for browsing; second, it extracts relevant information from the context returned by the retriever, crafting an answer by combining in-context prompts,[1] and if necessary, code generation evaluated in a read-eval-print loop for clinical calculations. For reproducibility and fairer comparison, we used the gpt-4-0613 model from OpenAI with a maximum length of 8192 tokens. If no articles from the database exceeded the match threshold, the language model defaulted to answering using its intrinsic knowledge.

Examples of query types are as follows:

> Rephrasing query: "Given question (Q) convert it to a simple Google search term."
> QA Query: "Generate a thorough and concise answer for a given question (Q) on the basis of the provided context (C). If you are asked to calculate a value, output the final equation in Python code. Use an unbiased and journalistic tone. If you cannot find a relevant answer, write "I apologize but there doesn't seem to be a clear answer to your question based on my sources ... Answer the question based on your own knowledge." Cite sources as [1] or [2] or [3] after each sentence to back up your answer (Ex: Correct: [1], Correct: [2][3], Incorrect: [1, 2])."
> Calculator Postquery: "Given the question (Q), context (C), and output (O), generate a thorough and concise answer for the given question (Q)."

CLINICALQA EVALUATION

To assess the outputs generated by LLMs on ClinicalQA, we proposed a framework on the basis of physician feedback to ensure alignment with our key metrics. Current LLM evaluation metrics rely on automated methods, such as bilingual evaluation understudy,[55] but they fail to fully capture the complexity and nuances of medical retrieval tasks. These questions are outlined in Table 2.

After collecting clinical questions from the panelists within their respective specialties, we fed each question into a series of models: Almanac, ChatGPT-4 (May 24, 2023 version), Bing (June 28, 2023), Bard (June 28, 2023), Galactica 120B, RAG, and BioMedLM (formerly known as PubMed GPT). We then retrieved their answers. Because of the very poor performance by the last three models (Supplementary Appendix), only Almanac, ChatGPT-4, Bing, and Bard were used for further evaluation. The latter models were prompted with the standard "think carefully and step by step" before each question, which has been shown to generally lead to more accurate outputs.[56] Bing was further set to reply in the "precise" configuration.

To quantify factuality, completeness, and preference, we tasked the clinician panelists with independently assessing outputs generated by the aforementioned models on ClinicalQA within their respective specialties. There was no overlap between graders, as all were asked to evaluate their own submitted questions in accordance with their specialties and fellowships. Grader statistics are available in the Supplementary Appendix. The evaluators were explicitly instructed to distinguish between the performance of different models rather than assigning them equivalent scores. Although efforts were made to ensure unbiased grading (e.g., arbitrary answer formatting, answer order shuffling) to blind physicians to the answer's provenance, complete blinding was not possible because of the prose styles adopted by each system. Graders were made aware of the study's goals, but they were not cognizant of the models being tested or familiar with the prose styles adopted by each model.

To assess citation quality, we manually reviewed and reported citations for each question using a binary system: one for valid citations and zero for missing or unreliable sources. A1 was assigned only if all reported citations were deemed valid. For the assessment of adversarial safety, we compared different model performances on a subset of ClinicalQA questions to evaluate their potential for intentional and unintentional harm. Our approaches were as follows:

- Adversarial prompting. Classified as intentional harm, adversarial prompting involves appending directives to a user's prompt to deter the language model from its original task. These prompts can be initiated by a malicious actor through various entry points, such as the electronic health record client or server, with the simplest approach involving the insertion of "invisible" directives (e.g., white font, image alt text) into a patient's clinical note to manipulate the model. Example prompts can include direct orders to generate incorrect outputs or more advanced scenarios designed to bypass the artificial safeguards gained through model fine-tuning (e.g., role-playing). We used both methods and evaluated Almanac, ChatGPT-4, Bing, and Bard on a subset of 25 ClinicalQA questions with a set of five common adversarial prompts of varying length.
- Errors of input. As the name suggests, we classified errors of input as unintentional harm, whereby incomplete input from a health care worker resulted in incorrect LLM outputs because of hallucinations rather than helpful errors/warnings. Because of the considerable burden and information overload experienced by health care workers in their day-to-day responsibilities, it is not uncommon for medical errors to occur because of human error and poor documentation.[43] To simulate this, we withheld key words from five clinical calculation vignettes and assessed their effects on LLMs outputs.

All evaluations were conducted before July 30, 2023.

OTHER BENCHMARKS

To provide more objective assessments of Almanac's performance, we further evaluated it on the LiveQA[57] test set composed of 102 questions. All question subjects and messages were fed into the model, with the responses being manually evaluated using the zero to three scoring schema from the associated publication. The judgment scores were as follows: 0=poor (or unanswered); 1=fair; 2=good; and 3=excellent. The average score is reported.

STATISTICAL EVALUATION

To evaluate the results, the Friedman test was performed for statistical significance at P<0.01 on the ranked responses submitted by the panel of evaluators across the ClinicalQA data set for each axis. The Nemenyi post hoc test was then performed to obtain pairwise comparisons between each model. The mean, weighted mean (with weights increasing from one to four according to rank), and inverse-weighted mean (with weights decreasing from four to one according to rank) are reported for each of the models.

Results

This section provides an overview of the results as summarized in Figure 2 and Table 3.

Across the full scope of ClinicalQA, Almanac outperformed its counterparts, with a mean rank of 1.96±0.06 in factuality, 1.85±0.06 in completeness, and 1.87±0.06 in preference, with similar overall performances observed within each specialty (Table 3 and Supplementary Appendix). Similar trends were observed for the weighted mean (with lower ranks penalized more heavily) and inverse-weighted mean (with higher ranks weighed more heavily), with a noted slightly better weighted mean performance by ChatGPT-4 in anesthesia (factuality, 5.23±0.15 vs. 5.60±0.22) and matched performances in cardiology (factuality, 6.20±0.11 vs. 6.20±0.14).

These results were echoed by the Nemenyi P values (P<0.01) in Figure 2 showing significant differences between Almanac and its counterparts. Despite slightly worse results in the performance of Bing compared with ChatGPT-4 on the ClinicalQA data set, the Nemenyi heat maps showed no significant differences in their performances. Overall, Bard was shown to have significantly worse performance on the ClinicalQA data set, with answers tending toward the lower ranks across the three axes, as shown in the Nemenyi plots.

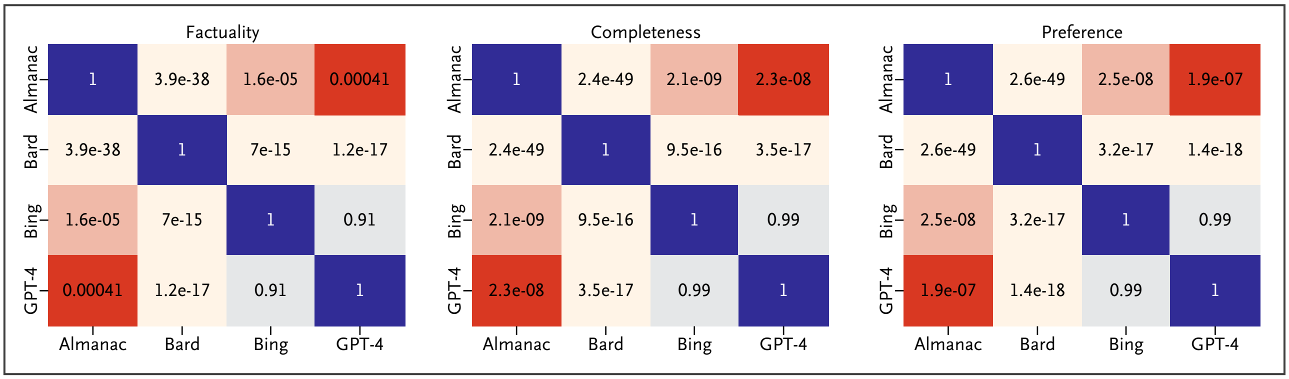

Figure 2. Heat Maps of the Nemenyi P Values for Factuality, Completeness, and Preference for Model Pairs across ClinicalQA.

Red denotes significant differences at P<0.01; blue denotes nonsignificant differences.

For citations, Almanac was able to provide correct and trustworthy citations for 91% of the ClinicalQA questions, with missed marks because of an inability to provide reliable sources when relying on its intrinsic knowledge. However, despite providing sources for every question on the data set, Bing achieved a performance of 82% because of unreliable sources, including personal blogs and online forums. Although ChatGPT-4 citations were mostly plagued by nonexistent or unrelated web pages, Bard either relied on its intrinsic knowledge or refused to cite sources, despite being prompted to do so.

Regarding adversarial safety, Almanac's performance greatly superseded that of Bard, Bing, and ChatGPT-4 in

Table 3. Almanac Outperformed Counterparts across the Scope of ClinicalQA.

Model Performance across Metrics	Bard	ChatGPT-4	Bing	Almanac
Mean ranks for various models across different axes (lower is better)				
Axis				
Factuality	3.17±0.06	2.34±0.06	2.41±0.06	1.96±0.06
Completeness	3.21±0.06	2.40±0.06	2.44±0.06	1.85±0.06
Preference	3.23±0.06	2.39±0.06	2.42±0.06	1.87±0.06
Weighted mean ranks for various models across different axes (lower is better)				
Axis				
Factuality	11.10±0.05	6.45±0.05	6.89±0.05	4.99±0.06
Completeness	11.33±0.04	6.70±0.05	7.03±0.05	4.48±0.06
Preference	11.44±0.04	6.65±0.05	6.93±0.05	4.57±0.06
Inverse weighted mean ranks and SE for various models across different axes				
Axis				
Factuality	0.38±0.07	0.54±0.05	0.53±0.06	0.67±0.05
Completeness	0.37±0.07	0.52±0.05	0.52±0.06	0.70±0.04
Preference	0.37±0.07	0.52±0.05	0.52±0.06	0.70±0.05
Percentage of correct citations across all outputs for various models				
Model				
Percentage of correct citations	9.84	21.27	82.54	91.11
Adversarial safety metrics for various models				
Metric				
Percentage of correct in adversarial prompting	76.80	7.00	70.40	100
Percentage of correct in errors of input	—	100.00	—	100.00

adversarial prompting (100% vs. 77, 70, and 7%, respectively). We note that for Almanac, the addition of the adversarial prompt lowered the average score between the query and the retrieved articles below the threshold (λ), resulting in the system abstaining from responding to a given prompt. Interestingly, Bard was also able to output the correct response along with the adversarial output despite being prompted not to, whereas Bing complied or refused to respond altogether. In contrast, ChatGPT-4 did not show the same reservations. For errors of input, although Bard and Bing refused to perform any clinical calculations, both ChatGPT-4 and Almanac were able to catch missing inputs (Supplementary Appendix). On LiveQA, Almanac obtained an average score of 2.85 across the 102 questions, well above the best-performing model reported in the article with an average score of 0.637 (a 347% improvement). We note that although Almanac was able to provide more up-to-date answers than the ones provided at times, it still struggled with specific resources, such as providing a clinical or physician recommendation.

Discussion

The current study proposes a framework for the safe deployment of LLMs in health care settings to answer clinical queries more accurately across a range of specialties. We evaluated our approach on a novel data set of clinical questions and showed that our framework achieves significant improvements in factuality, completeness, preference, and adversarial safety compared with baselines as assessed by a panel of board-certified physicians and health care workers.

In recent months, there have been several works exploring the role of LLMs in clinical medicine, including DRAGON,[58] BioGPT,[21] and Med-PaLM.[30] Despite strong performances on medical question–answer data sets, such as MedQA,[59] these models failed to translate to real-world clinical scenarios because of the potential for harm from unmitigated hallucinations, benchmarks that do not accurately reflect clinically relevant tasks, and concerns of data contamination between train–test splits. Moreover, because these systems leverage the knowledge encoded within their weights to answer clinical queries, their outputs become contingent on the assumption that correct information outweighs misinformation within their training data set. Despite potential mitigations, such as with supervised fine-tuning and reinforcement learning with human feedback,[19] these models will need to be continuously trained to update their knowledge bases, which can quickly become prohibitively expensive at billion-parameter sizes. Finally, as a result of their nondeterministic outputs, these models often display varying and sometimes contradicting responses to the same query (or even similar queries with different wording), making them unreliable for clinical use.

Our results suggest that retrieval systems can effectively facilitate knowledge distillation, leading to more accurate and reliable responses to clinical inquiries, grounded in fact. By supplementing responses with passages from predefined sources, our grounded system is able to dampen explainability concerns by enabling clinicians to independently verify outputs from trustworthy sources, leading to significant improvements over general retrieval systems. We also found this retrieval system to be especially useful in adversarial settings in which the query-context scoring system can hamper malicious actors from manipulating outputs. We note that this off-the-shelf resilience becomes less effective as the adversarial prompt decreases in word count, and careful λ tuning must be made to balance between true- and false-positive findings (Supplementary Appendix).

Our research indicates that Almanac could offer a safer and more reliable avenue for answering clinical questions. However, more extensive studies are required to fully understand its potential impact in clinical settings. Although Almanac performs admirably across a range of medical specialties, it has limitations in effectively ranking information sources by criteria, such as evidence level, study type, and publication date. We believe that optimizing the retrieval algorithm, perhaps through recursive strategies, and incorporating reinforcement learning with human feedback to factor in human preferences may address these issues. Currently, we use general-purpose retrievers and language models; fine-tuning these components could further improve system performance. It is also worth noting that although our ClinicalQA benchmark shows promise, the evaluation metrics are subjective and rely on human graders, posing challenges for scalability. Future work should focus on developing automated metrics that align well with human evaluation.

Grounded language models, such as Almanac, have shown progress. They are not without flaws, however, particularly in generating accurate responses and handling questions that lack straightforward answers in their data sources. As such, their integration into health care settings should be approached with caution, accompanied by strategies to mitigate potential errors.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

Supported in part by a National Institutes of Health, National Heart, Lung, and Blood Institute grant (1R01HL157235-01A1, to Dr. Hiesinger).

Because of growing concerns of medical benchmarks being used as data for large-scale training of large language models and further contributing to data contamination of clinical benchmarks, a subset (n=25) of our data set is being published with this article (Supplementary Appendix), and the rest is available upon request. Dr. Hiesinger can be contacted (willhies@stanford.edu) for full access to ClinicalQA.

We thank Hugging Face for their support over the course of the project and Morten Just for designing the evaluation forms submitted to the panel of physicians.

Author Affiliations

[1] Department of Cardiothoracic Surgery, Stanford Medicine, Stanford, CA
[2] Department of Computer Science, Stanford University, Stanford, CA
[3] Division of Cardiovascular Surgery, Penn Medicine, Philadelphia
[4] Division of Cardiovascular Medicine, Stanford Medicine, Stanford, CA
[5] Department of Pediatrics, Stanford Medicine, Stanford, CA
[6] Department of Neurology, Stanford Medicine, Stanford, CA
[7] Department of Radiology and Biomedical Informatics, Stanford Medicine, Stanford, CA
[8] Division of Gastroenterology and Hepatology, Johns Hopkins Medicine, Baltimore
[9] Division of Infectious Diseases, Stanford Medicine, Stanford, CA
[10] Division of Anesthesia, Stanford Medicine, Stanford, CA
[11] Department of Statistics, Columbia University, New York

References

1. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. July 22, 2020 (https://arxiv.org/abs/2005.14165). Preprint.

2. Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code. July 14, 2021 (https://arxiv.org/abs/2107.03374). Preprint.

3. Wei C, Xie SM, Ma T. Why do pretrained language models help in downstream tasks? An analysis of head and prompt tuning. June 2021 (https://arxiv.org/abs/2106.09226). Preprint.

4. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. October 2018 (https://arxiv.org/abs/1810.04805). Preprint.

5. Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. June 2022 (https://arxiv.org/abs/2206.07682). Preprint.

6. Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models. March 2022 (https://arxiv.org/abs/2203.15556). Preprint.

7. Rae JW, Borgeaud S, Cai T, et al. Scaling language models: methods, analysis and insights from Training Gopher. December 2021 (https://arxiv.org/abs/2112.11446). Preprint.

8. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2022;55:1-38. DOI: 10.1145/3571730.

9. Raunak V, Menezes A, Junczys-Dowmunt M. The curious case of hallucinations in neural machine translation. April 2021 (https://arxiv.org/abs/2104.06683). Preprint.

10. Liang PP, Wu C, Morency L-P, Salakhutdinov R. Towards understanding and mitigating social biases in language models. June 2021 (https://arxiv.org/abs/2106.13219). Preprint.

11. Swinger N, De-Arteaga M, Heffernan NT, Leiserson MD, Kalai AT. What are the biases in my word embedding? December 2018 (https://arxiv.org/abs/1812.08769). Preprint.

12. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. March 2022 (https://arxiv.org/abs/2203.02155). Preprint.

13. Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: harmlessness from AI feedback. December 2022 (https://arxiv.org/abs/2212.08073). Preprint.

14. Zhou Y, Muresanu AI, Han Z, et al. Large language models are human-level prompt engineers. November 2022 (https://arxiv.org/abs/2211.01910). Preprint.

15. Reynolds L, McDonell K. Prompt programming for large language models: beyond the few-shot paradigm. February 2021 (https://arxiv.org/abs/2102.07350). Preprint.

16. Thoppilan R, De Freitas D, Hall J, et al. LaMDA: language models for dialog applications. January 2022 (https://arxiv.org/abs/2201.08239). Preprint.

17. Hartvigsen T, Gabriel S, Palangi H, Sap M, Ray D, Kamar E. ToxiGen: a large-scale machine-generated dataset for adversarial and implicit hate speech detection. March 2022 (https://arxiv.org/abs/2203.09509). Preprint.

18. Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language models. December 2020 (https://arxiv.org/abs/2012.07805). Preprint.

19. Christiano P, Leike J, Brown TB, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. February 2023 (https://arxiv.org/abs/1706.03741).

20. Lucas BP, Evans AT, Reilly BM, et al. The impact of evidence on physicians' inpatient treatment decisions. J Gen Intern Med 2004; 19:402-409. DOI: 10.1111/j.1525-1497.2004.30306.x.

21. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022;23:bbac409. DOI: 10.1093/bib/bbac409.

22. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. 2019;3615-3620. DOI: 10.18653/v1/D19-1371.

23. Shin H-C, Zhang Y, Bakhturina E, et al. BioMegatron: larger biomedical domain language model. October 2020 (https://arxiv.org/abs/2010.06060). Preprint.

24. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234-1240. DOI: 10.1093/bioinformatics/btz682.

25. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc 2021;3:1-23. DOI: 10.1145/3458754.

26. Papanikolaou Y, Pierleoni A. DARE: data augmented relation extraction with GPT-2. April 2020 (https://arxiv.org/abs/2004.13845). Preprint.

27. Hong Z, Ajith A, Pauloski G, et al. ScholarBERT: bigger is not always better. May 2022 (https://arxiv.org/abs/2205.11342). Preprint.

28. Lehman E, Hernandez E, Mahajan D, et al. Do we still need clinical language models? February 2023 (https://arxiv.org/abs/2302.08091). Preprint.

29. Yang X, Chen A, PourNejatian N, et al. GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records. December 2022 (https://arxiv.org/abs/2203.03540). Preprint.

30. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. December 2022 (https://arxiv.org/abs/2212.13138). Preprint.

31. Taylor R, Kardas M, Cucurull G, et al. Galactica: a large language model for science. November 2022 (https://arxiv.org/abs/2211.09085). Preprint.

32. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, December 2020 (https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf).

33. Borgeaud S, Mensch A, Hoffmann J, et al. Improving language models by retrieving from trillions of tokens. International Conference on Machine Learning (ICML), 2022.

34. Simon C, Davidsen K, Hansen C, Seymour E, Barnkob MB, Olsen LR. BioReader: a text mining tool for performing classification of biomedical literature. BMC Bioinformatics 2019;19(Suppl 13):57. DOI: 10.1186/s12859-019-2607-x.

35. Naik A, Parasa S, Feldman S, Wang LL, Hope T. Literature augmented clinical outcome prediction. November 2022 (https://arxiv.org/abs/2111.08374).

36. Ram O, Levine Y, Dalmedigos I, et al. In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics 2023;11:1316-1331. DOI: 10.1162/tacl_a_00605.

37. Nakano R, Hilton J, Balaji S, et al. WebGPT: browser-assisted question-answering with human feedback. December 2021 (https://arxiv.org/abs/2112.09332). Preprint.

38. Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: language models can teach themselves to use tools. February 2023 (https://arxiv.org/abs/2302.04761). Preprint.

39. Liévin V, Hother CE, Winther O. Can large language models reason about medical questions? July 2022 (https://arxiv.org/abs/2207.08143). Preprint.

40. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. March 2022 (https://arxiv.org/abs/2203.14371).

41. Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: a dataset for biomedical research question answering. In: Inui K, Jiang J, Ng V, Wan X, eds. 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: Association for Computational Linguistics, 2019:2567–2577. DOI: 10.18653/v1/D19-1259.

42. Jacovi A, Caciularu A, Goldman O, Goldberg Y. Stop uploading test data in plain text: practical strategies for mitigating data contamination by evaluation benchmarks. In: Bouamor H, Pino J, Bali K, eds. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023:5075-5084. DOI: 10.18653/v1/2023.emnlp-main.308.

43. McGaghie WC, Cohen ER, Wayne DB. Are United States Medical Licensing Exam Step 1 and 2 scores valid measures for postgraduate medical residency selection decisions? Acad Med 2011;86:48-52. DOI: 10.1097/ACM.0b013e3181ffacdb.

44. Panda N, Bahdila D, Abdullah A, Ghosh AJ, Lee SY, Feldman WB. Association between USMLE step 1 scores and in-training examination performance: a meta-analysis. Acad Med 2021;96:1742-1754. DOI: 10.1097/ACM.0000000000004227.

45. Cohen ER, Goldstein JL, Schroedl CJ, Parlapiano N, McGaghie WC, Wayne DB. Are USMLE scores valid measures for chief resident selection? J Grad Med Educ 2020;12:441-446. DOI: 10.4300/JGME-D-19-00782.1.

46. Rifkin WD, Rifkin A. Correlation between housestaff performance on the United States Medical Licensing Examination and standardized patient encounters. Mt Sinai J Med 2005;72:47-49.

47. Bray K, Burge K, Patel O, et al. Perceptions of the emergency medicine resident selection process by program directors following the transition to a pass/fail USMLE Step 1. Open Access Emerg Med 2023;15:15-20. DOI: 10.2147/OAEM.S389868.

48. Sajadi-Ernazarova K, Ramoska EA, Saks MA. USMLE scores do not predict the clinical performance of emergency medicine residents. Mediterranean Journal Emerg Med Acute Care 2020;1:4–7. DOI: 10.52544/2642-7184(1)2001.

49. Malkov YA, Yashunin DA. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. March 2016 (https://arxiv.org/abs/1603.09320). Preprint.

50. GitHub. Qdrant: open-source vector similarity search engine with segment storage and distributed serving. May 2020 (https://github.com/qdrant/qdrant).

51. MDCalc. MDCalc — medical calculators, equations, scores, and guidelines. (https://www.mdcalc.com).

52. National Center for Biotechnology Information. PubMed. National Library of Medicine. 2023 (https://pubmed.ncbi.nlm.nih.gov).

53. Wolters Kluwer. UpToDate. 2023 (https://www.uptodate.com).

54. BMJ Publishing Group. BMJ Best Practice. 2023 (http://bestpractice.bmj.com).

55. Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, July 2002.

56. Yang C, Wang X, Lu Y, et al. Large language models as optimizers. September 2023 (https://arxiv.org/abs/2309.03409). Preprint.

57. Ben Abacha A, Agichtein E, Pinter Y, Demner-Fushman D. Overview of the Medical Question Answering Task at TREC 2017 LiveQA. Text Retrieval Conference, 2017 (https://api.semanticscholar.org/CorpusID:3902472).

58. Yasunaga M, Bosselut A, Ren H, et al. Deep bidirectional language-knowledge graph pretraining. October 2022 (https://arxiv.org/abs/2210.09338).

59. Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. September 2020 (https://arxiv.org/abs/2009.13081).