MedCite: Can Language Models Generate Verifiable Text for Medicine?

Anonymous ACL submission

Abstract

Existing LLM-based medical question-answering systems lack citation generation and evaluation capabilities, raising concerns about their adoption in practice. In this work, we introduce MedCite, the first end-to-end framework that facilitates the design and evaluation of LLM citations for medical tasks. Meanwhile, we introduce a novel multi-pass retrieval-citation method that generates high-quality citations. Our evaluation highlights the challenges and opportunities of citation generation for medical tasks, while identifying important design choices that have a significant impact on the final citation quality. Our proposed method achieves superior citation precision and recall improvements compared to strong baseline methods, and we show that evaluation results correlate well with annotation results from professional experts.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in various natural language processing tasks, such as question answering (QA) and instruction following (Kaplan et al., 2020; Wei et al., 2022a,b). Progress in LLMs has also enabled the development of medical agents that understand language used by patients and physicians, offering rich just-in-time assistance (Singhal et al., 2022, 2023; Temsah et al., 2023; Tangadulrat et al., 2023; Maples et al., 2024).

While the early signs are positive, current LLM-powered medical QA systems still have multiple limitations. For example, medical data often contains sensitive information, such as personal health records, which cannot be used for training large language models without strict compliance with ethical standards (Gilbert et al., 2023). Furthermore, trustworthiness is particularly important in the medical field. Issues such as hallucination, where the model generates information that is incorrect

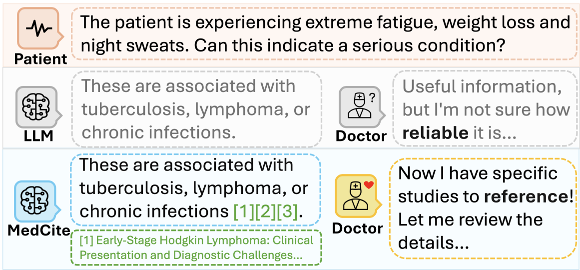


Figure 1: Medical QA system comparison. State-of-the-art systems generate answers without citations. MedCite not only generates answers but also associates each answer with citations, improving the verifiability and trustworthiness of the medical system.

or misleading, pose significant challenges to the reliability of LLM-based medical systems (Pal et al., 2023; Ahmad et al., 2023; Huang et al., 2024). To overcome the issue, researchers and practitioners have studied retrieval-augmented generation (RAG) (Xiong et al., 2024a; Yang et al., 2024), which combines LLMs with information retrieval from external trustworthy data source (Canese and Weis, 2013). By providing the model with accurate and relevant medical knowledge, these systems allow LLMs to maintain relevance in responses.

Despite promising results, existing methods lack *verifiability* (Liu et al., 2023), meaning that the answers provided are not backed by reliable sources or evidence. This can lead to misinformation and potentially harmful consequences if incorrect medical advice is followed. For instance, as shown in Fig. 1, when providing a diagnosis based on a list of symptoms without any references, the accuracy of prognosis or treatment recommendations cannot be assured, which creates a sense of uncertainty, leading to suboptimal or even harmful decisions.

One promising approach to mitigate the verifiability concern is through *attribution* (Bohnet et al., 2022; Huang and Chang, 2024), i.e., associating statements with *citations*, which offers the system more credibility and accountability while providing users a way to explore the source in greater

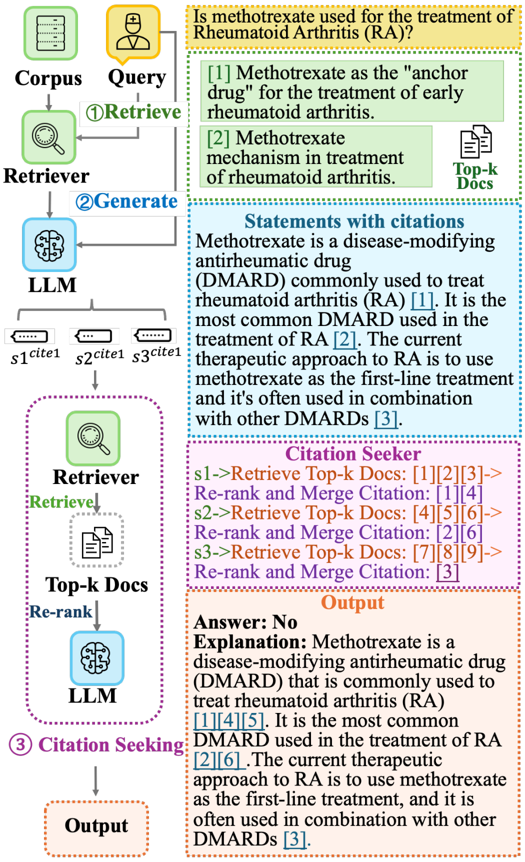


Figure 2: The overview figure of MedCite.

depth and verify the information source. However, although there are prior efforts that analyze citation capabilities through LLMs on general domain QA tasks (Liu et al., 2023; Gao et al., 2023c; Djeddal et al., 2024), citing sentences for medicine is especially challenging and not widely adopted in practice due to the following reasons.

First, existing works on medical QA often leverage multi-choice accuracy to benchmark and evaluate their performance (Xiong et al., 2024a; Yang et al., 2024; Yu et al., 2024), which focuses on evaluating the ability to select the correct answer from a set of given options. However, citation generation is more challenging due to its open-ended nature. For example, when prescribing medication for a rare genetic disorder or planning surgery for a patient, both physicians and patients have to rely on richer information. Therefore, a query can have multiple answers, supported by multiple possible sources. These aspects are important to consider for the evaluation of citation generation methods, but existing medicine QA frameworks do not inherently account for them.

Second, there is a huge design space for citation generation with complex interactions among retrievers (Asai et al., 2024; Izacard et al., 2022), backbone LLMs (MetaAI, 2024; Zhang et al., 2024; OpenAI, 2023), and citation generation algorithms (Gao et al., 2023c). Therefore, it can be challenging to determine which aspects contribute the most. However, the analysis is crucial for developing strategies to improve the verifiability of medical systems.

Third, while there has been a continuous rise in the number of contributions in this field (Gao et al., 2023c; Xiong et al., 2024a; Yang et al., 2024; Yu et al., 2024), there is a noticeable deficiency in open-source frameworks useful for designing, developing, and evaluating citation generation quality for medical tasks. The existing citation evaluation frameworks are constructed from generic questions, where the selection of metrics and evaluators for medical tasks remains an open question. Moreover, it is quite costly to obtain high-quality medical expert annotations, which demands a high-quality classifier to judge whether a citation attributes to a statement.

In this paper, we tackle these challenges with the hope of fostering research in improving verifiability for medicine systems. In particular, our contributions are as follows:

- An in-depth study of different citation methods and key design components for medical tasks using LLMs, ranging from text generation methods, information retrieval methods, and citation attribution methods. Our study disentangles the importance of different factors from the backbone LLM.
- We present MedCite, the first end-to-end system for enabling LLMs to generate verifiable texts for medical QA systems with automatic evaluation. Meanwhile, we introduce a novel multi-pass retrieval-citation method that conciliates retrieval-augmented generation and post-generation citation.
- A comprehensive evaluation of MedCite across different LLMs, which shows that MedCite outperforms existing methods in both text generation and citation generation quality by up to 47.39% recall and 31.61% precision respectively. We conduct human evaluation by having medical doctors verify the attribution results. The results show that our automatic evaluation pipeline correlates well with domain expert judgments, demonstrating the effectiveness of efficient and automatic citation evaluation for medicine.

2 Problem Setup

In this section, we first formulate the citation generation task for biomedical QA and then give an overview of the approaches that we will examine experimentally in the following section. Due to space limitations, we have included the related work on biomedical QA and citation methods for LLM generation in Appendix A.

2.1 Problem Objective

The objective is to develop a system that automatically add relevant and accurate citations to text statements generated by a large language model. In particular, the inputs to the system include a user query q, an LLM Φ, an external database D, which contains ground truth documents. The outputs of the system include a generated text passage, which contains a list of statements $S = \{s_0, s_1, ..., s_n\}$ by Φ. For each statement s_i, a set of in-line citations $C_i = \{c_i^0, c_i^1, ..., \}$, where $c_i^j \in$D, is assigned to it.

2.2 Dataset

Following prior work (Bolton et al., 2024; Yasunaga et al., 2022; Xiong et al., 2024a), We use the BioASQ-Y/N dataset (Nentidis et al., 2024), which is a commonly used dataset for benchmarking biomedical question answering systems. The dataset consists of questions, human-annotated answers, and relevant contexts that provide the necessary information to answer the questions. The BioASQ-Y/N dataset has three characteristics that motivate us to use it for the study: 1) Unlike other datasets used for medical QA (Jin et al., 2020; Hendrycks et al., 2021; Pal et al., 2022), which are primarily multi-choice QA tasks, BioASQ-Y/N not only provides option choices (Yes/No) but also a gold set of answers w.r.t the informativeness of answer statements. 2) BioASQ-Y/N provides ground truth labels of the supporting documents for each question. Meanwhile, it can be easily modified to answer questions without the ground-truth documents provided, which represents a more realistic medical setting. 3) It has not so far been used by existing generic citation methods. Apart from BioASQ used for the analysis in Section 3, we also include PubMedQA (Jin et al., 2019) in Section 5. We include the details of the datasets and hyperparameters in Appendix D. On the external database side, we primarily consider PubMed database (Canese and Weis, 2013), which contains 24.6 million biomedical documents vetted by medical professionals. This vast database provides access to a wealthy source of precise and legitimate documents LLM-generated text can attribute to.

2.3 Evaluation Metrics

For medicine QA, evaluating both text and citation generation quality is crucial to ensure that the outputs of LLMs are not only coherent and relevant but also well-supported by accurate citations. As such, we consider the following aspects.

Answer correctness. Different from multi-choice QA, real medical systems often generate long and open-ended answers. Therefore, we use ROUGE-L (Lin, 2004) and MAUVE (Pillutla et al., 2021) to evaluate the correctness and relevance of the answer based on the ground truth answer. We still let the model generate a Yes/No answer in addition to the long answer, such that we can make comparisons with existing non-citation methods.

Citation quality. We consider an attribution judge *Attr* : $\mathcal{X}, \mathcal{Y} \rightarrow \{0, 0.5, 1\}$ that outputs 1 if the statements $\mathcal{X}$ can be fully attributed to the statements $\mathcal{Y}$, i.e., $\mathcal{Y}$ is the source of $\mathcal{X}$, 0.5 if $\mathcal{X}$ can be partially attributed to $\mathcal{Y}$, and 0 otherwise. To justify the introduction of partial support, we refer to findings from recent studies, such as Wührl et al. (2024), which showed that in medical fact-checking tasks, 62.4% of claims were partially supported by evidence. This highlights the importance of capturing partial attribution, as it is a frequent occurrence in real-world medical statements.

For the use of citations in medical QA, an answer can have multiple verifiable statements, and multiple citations may be attached to support one statement. With the attribution judge, we measure citation qualities with two metrics: *citation recall*, and *citation precision*. Both citation recall and precision heavily affect the usability of medical QA, as a high recall means that the generated responses are well supported by evidence, and a high precision indicates that the assigned citations have high quality that can be used to verify the truthfulness of the generated texts. For simplicity, let us consider a single statement s with n citations $c_1, c_2, \ldots, c_n$, where each of them is a set of axioms.

Citation recall. We define recall as a statement-level metric, which measures whether all the information in the statement is fully supported by the citations. Such a metric can be formally defined as

$$\text{Recall}(s, c_1, \cdots, c_n) = \begin{cases} 1 & \text{if } s \subseteq \bigcup_{i=1}^{n} c_i, \\ 0 & \text{if } s \not\subseteq \bigcup_{i=1}^{n} c_i. \end{cases} \quad (1)$$

In our experiments, we use the concatenation of the citation documents to represent the union of citations and make a judgment on whether the statement can or cannot be fully supported by the concatenated citations. Using the attribution judge defined above, we have $\text{Recall}(s, c_1, \cdots, c_n) = 1$ if and only if $Attr(s, \bigcup_{i=1}^{n} c_i) = 1$. We then average over all statements to get the final recall score of an answer passage.

Citation precision. Following previous research (Liu et al., 2023), we define the precision metric as a citation-level measurement, which assesses if each individual citation contributes to supporting the statement. Precision is 1 for a citation if it either fully or partially supports the statement by containing at least some of the necessary axioms from the statement. Formally, the precision of a citation c_i for the given statement s is defined as

$$\text{Precision}(s, c_i) = \begin{cases} 1 & \text{if } s \cap c_i \neq \emptyset, \\ 0 & \text{if } s \cap c_i = \emptyset. \end{cases} \quad (2)$$

The precision of c_i for s is computed as 1 if and only if $Attr(s, c_i) > 0$. When having multiple statements, we compute its citation precision by averaging the precision scores of all citations in it.

Citation $\mathbf{F}_1$. We use citation F_1 (Liu et al., 2023) to measure the combined citation precision and recall via: $F_1 = 2 \times \frac{citation\ precision \times citation\ recall}{citation\ precision + citation\ recall}$.

3 Citation Procedure Analysis

This section explores and quantifies which choices are important for successfully citing sentences for medical tasks. Given that each component can be varied, we investigate how each of these components impacts the citation generation quality while isolating the other components.

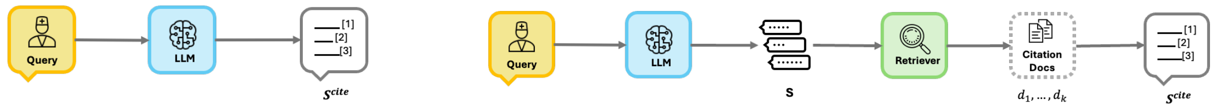
3.1 Parametric vs. Non-Parametric Citation

Recent LLMs can be prompted to include citations in the text they generate by relying on its parametric contents, i.e., information internalized from the training data. Given this advancement, one question naturally arises: *can we rely on LLMs to self-cite their generated sentences?* We compare this strategy with non-parametric citation where we generate citations by purely relying on non-parametric information-retrieval (IR) contents, e.g., PubMed. In particular, for parametric citation, we generate a prompt that includes the user question, and a directive instruction for LLM to generate answers while adding in-line citations in formatted output for each statement. In this case, the model solely depends on its pre-training data to generate citations. For the non-parametric citation, we let LLM to directly generate an answer without citations. Then we use a dense retriever MedCPT (Jin et al., 2023) to retrieve a list of relevant document (e.g., top-3) from D for each answer statement and those documents as in-line citations. The prompts used can be found in Appendix B.

Citation Method	Model	Accuracy (EM)	Text Quality		Citation Quality	
			MAUVE	ROUGE-L	Rec.	Prec.
Parametric (LLM)	Llama-3-8B-I.	74.76	61.94	17.72	/	/
	UltraMedical	69.09	67.70	13.96	/	/
	GPT-4o	**88.51**	**74.82**	**20.03**	/	/
Non-parametric (IR)	Llama-3-8B-I.	73.95	65.31	19.05	60.89	53.90
	UltraMedical	68.12	51.18	12.69	52.48	62.32
	GPT-4o	**87.70**	**70.15**	**20.20**	**79.72**	**80.95**

Table 1: Comparison of parametric (LLM) vs. non-parametric (IR) citation methods across different LLMs.

Table 1 compares the parametric vs. non-parametric citation results across different LLMs. We find that while LLMs have made significant strides in understanding and following human instructions, they do have limitations when it comes to generate citations in medical settings. In particular, both Llama-3-8B-I. and UltraMedical cannot follow those instructions accurately. As a result, the generated citations are either incorrect, fabricated, or ill-formatted, and even though a small proportion of them do exist, such citations might not be freely accessible (e.g., some scientific articles are behind a paywall). As such, without API access to the content of any scientific articles generated as citations by parametric methods, it is challenging to automatically evaluate their quality. This is unsurprising because these models are still trained on next-token prediction, and LLM needs to extrapolate the citation information with its pretraining knowledge or hallucination. Interestingly, GPT-4o not only achieves the highest accuracy on the BioASQ task, but it can also consistently follow the instructions to generate well-formatted citations. However, the references GPT-4o generated are outdated (all before year 2018), making it hard to include new studies. We include several examples of generated citations in Appendix C. This observation highlights a critical limitation of the parametric-only approach when applied to citation generation, particularly in the medicine domain with public LLMs. Given these limitations, we focus on non-parametric citation methods using trusted datasets like PubMed in the remainder of the experiments.



(1) Parametric citation generation

(2) Non-parametric citation generation

Figure 3: Comparison of parametric (LLM) and non-parametric (IR) citation generation pipelines.

3.2 RAG Makes Better Citations

While non-parametric citation improves the citation quality, the answer statement generation still relies on the pre-training data itself. Therefore, the answers can be based on outdated or incomplete medical data. Despite multiple recent papers observing that adding retrieval-augmented generation (RAG) helps improve LLM to better understand biomedical tasks, producing higher accuracy than non-RAG based approaches (Xiong et al., 2024a; Yang et al., 2024; Yu et al., 2024), experiments on how RAG affects both text and citation quality are rarely reported. We dig deeper into the role of RAG in citation generation by comparing several different methods. For all methods, we use the same dense MedCPT based retriever to assign top-k (e.g., top-3) relevant documents as citations.

Non-RAG (CoT): We perform chain-of-thought (CoT) prompting (Wei et al., 2022b) to leverage the reasoning capability of LLMs to provide an answer (e.g., a polar Yes/No answer) and text explanations S to the question q. This is similar as the method in (Xiong et al., 2024a), but no supporting context is retrieved from the external database.

RAG: We first retrieve a shortlist of top-k supporting documents $\{d_1, ..., d_k\}$ to the query q from D. We then feed the concatenate shortlist documents together with q to the LLM, and instruct the LLM to generate the answer and text explanations S.

RAG w. Oracle: Similar to the above configuration, but using the ground truth supporting documents (i.e., assuming a perfect retriever) in BioASQ for each question.

Retrieval Method	Accuracy (EM)	Text Quality		Citation Quality	
		MAUVE	ROUGE-L	Recall	Precision
Non-RAG	71.36	53.24	18.07	**59.05**	**52.93**
RAG	82.85	52.22	14.79	49.01	42.77
RAG w. Oracle	**94.34**	**63.45**	**20.63**	57.46	43.20

Table 2: Comparison of RAG and Non-RAG methods for citation generation.

Table 2 shows a comparison of non-RAG and RAG methods for medicine. Interestingly, we observe that without RAG, the correctness of generated answers tend to be low (71.36). However, the citation recall and precision are relatively high. Conversely, integrating RAG leads to a significant increase in answer correctness while resulting a decrease in citation recall and precision. This is because the citation quality metrics only assess whether the LLMs' generated statements are *supported* by verifiable sources, rather than directly assessing the *correctness* of each statement. Therefore, it is possible that the identified citations can still support a hallucinated statement, even though that statement is irrelevant to the user's question. This finding suggests that we need to holistically assess LLM's capability for both text and citation generation. Specifically, we treat the correctness of the answers (e.g., accuracy and text quality) as a prerequisite of the evaluation of citations, and enabling citation capabilities should not compromise the quality of answer generation.

Table 2 also shows that by using the ground truth documents (oracle), the best obtainable results for accuracy using RAG can go up to 94.34% and the citation recall and precision can go up to 57.46% and 43.20% respectively, leaving room to investigate better retrieval augmentation methods. Nevertheless, these results indicate that RAG is crucial to generating context relevant texts and is a critical step for getting high-quality citations. Therefore, we use RAG in the remainder of our experiments.

3.3 The Efficacy of Citation Seeker

Till now we have fixed the choice of the citation seeker, i.e., how to find relevant documents and assign them as citations to a statement. However, one may wonder how the choice of citation-seeking methods affects citation quality. To investigate this, we consider the following strategies:

Pre-generation shortlist + LLM rerank. For each generated statement, we instruct LLM to assign a document retrieved as one of the supporting documents from the pre-generation retrieval. No additional retrieval is needed in this case.

Retriever-only re-retrieval. For each generated statement, we relaunch the retriever to retrieve top-k documents relevant to the statement from D and append those as citations for each statement.

Re-retrieval + NLI rerank. For each generated statement, we relaunch the retriever to retrieve top-k documents relevant to the statement from D, and use a light-weighted medicine NLI model to assign those retrieved documents as citations.

Re-retrieval + LLM rerank. Similar as above config, except that we instruct LLMs to assign retrieved documents as citations.

Attribution Strategy	Accuracy (EM)	Text Quality		Citation Quality	
		MAUVE	ROUGE-L	Recall	Precision
Pre-Gen. shortlist + LLM rerank	83.33	59.22	16.78	54.66	41.40
Retriever-only re-retrieval	83.33	59.22	16.78	65.69	47.69
Re-retrieval + NLI rerank	83.33	59.22	16.78	65.38	55.12
Re-retrieval + LLM rerank	83.33	59.22	16.78	**65.78**	**60.95**

Table 3: Comparison of citation seeking methods. We use the Hybrid configuration described in Section 4 because it has overall better citation quality.

Table 3 shows the comparison results of different citation seeking strategies. We find that *re-retrieval + LLM reranking* has the best overall performances, which confirms the benefits of (1) re-retrieval leads to improved citation precision and recall, and (2) citation reranking as an effective approach for seeking high quality citations. It is worth noting that while *re-retrieval + NLI reranking* achieves similar citation recall but with 5.8% lower recall precision, the NLI model is overall much more lightweight than an LLM. Therefore, if cost is a major constraint, a medicine-specialized NLI classifier can also be considered for citation seeking.

4 MedCite: A Citation Generation System for LLM-Powered Medical QA

In the previous section we investigate several important design choices for citation generation of medical tasks. We now aggregate these improvements and evaluate their combined impact and provide it as an open-source framework MedCite (Fig. 2). We call our final approach *MedCite-Hybrid*. Specifically, MedCite-Hybrid is built with non-parametric citation (§ 3.1), RAG (§ 3.2), and the retrieval + LLM reranking citation seeking method (§ 3.3). Additionally, we investigate another two important factors that have been underemphasized in previous work: (1) what if we combine parametric and non-parametric citations through multi-pass approaches; and (2) the impact of retriever choices to the citation seeking.

Multi-pass citation generation. Intuitively, it seems possible to leverage both LLM's internal parametric knowledge to provide initial answer and citations while employing post-generation non-parametric method to validate and refine the citations, utilizing the externally retrieved content. To verify our hypothesis, we consider a new multi-pass method: Similar to the approach in § 3.2, we employ RAG to generate answers. Different from that approach, we instruct LLM to assign citations to statements based on the retrieved documents while answering the question. Then we retrieve top-k relevant documents to each statement. We deduplicate any redundant citations from these two stages and combine the remaining ones to form the final citations. Table 4 presents the comparison results of the double-pass method against non-parametric citation. The results indicate that the double-pass approach consistently outperforms the non-parametric method in citation precision and recall while maintaining comparable and slightly better answer correctness. By combining the strengths of both generative and retrieval systems, the double-pass method mitigates the limitations inherent in each individual approach.

Configuration	Accuracy (EM)	Text Quality		Citation Quality	
		MAUVE	ROUGE-L	Rec.	Prec.
Non-parametric RAG + Citation Seeker	82.85	52.22	14.79	49.01	42.77
Hybrid Double-pass	**83.33**	**59.22**	**16.78**	**65.69**	**47.69**

Table 4: Comparison of non-parametric and MedCite's double-pass method for citation generation.

Hierarchical two-stage ranking based citation retrieval. Another factor is the choice of retriever for the citation seeking. The recently proposed MedRAG (Xiong et al., 2024a) uses a Reciprocal Rank Fusion (RRF) based hybrid method to combine results from BM25 (Robertson and Zaragoza, 2009) and MedCPT (Jin et al., 2023) to find supporting documents in the pre-generation phase. However, while it is possible to find a broad range of relevant documents to enhance the context of LLM-generated answers, citation retrieval must be more fact-focused to ensure precise and accurate referencing. In the ablation studies, we show that a hierarchical two-stage ranker that first retrieve documents based on key word matching through BM25 (Robertson and Zaragoza, 2009) and then semantic retriever based on MedCPT (Jin et al., 2023) brings further improvements in performance in citation quality, validating the importance of the choice of retriever for citation.

5 Evaluation

5.1 Main Results

We compare MedCite with three baseline methods: the medical domain RAG method and two

Model	Method	Acc. (EM)		Text Gen. Quality				Citation Quality					
				MAUVE		ROUGE-L		Recall		Precision		F1-Score	
		BioASQ	PubMedQA	BioASQ	PubMedQA	BioASQ	PubMedQA	BioASQ	PubMedQA	BioASQ	PubMedQA	BioASQ	PubMedQA
Llama-3-8B-I.	MedRAG	82.85	**70.80**	53.74	42.39	14.78	14.22	/	/	/	/	/	/
	PRG	**84.95**	69.40	**72.53**	**47.79**	**17.97**	**20.99**	35.44	30.08	38.71	35.00	32.50	36.73
	PGC*	72.10	55.80	61.90	44.53	18.06	19.11	64.75	62.18	69.32	**71.75**	66.96	66.62
	MedCite	**84.95**	69.40	**72.53**	**47.79**	**17.97**	**20.99**	**74.86**	**69.50**	**69.47**	67.73	**71.74**	**68.60**
UltraMedical	MedRAG	**74.92**	**65.00**	57.24	**58.82**	**17.33**	**20.54**	/	/	/	/	/	/
	PRG	63.43	53.60	**63.87**	48.02	13.27	14.89	27.54	28.51	30.80	31.17	28.01	30.94
	PGC	68.12	44.80	50.71	41.04	12.69	13.33	49.91	54.28	**62.18**	**72.82**	55.37	**62.21**
	MedCite	63.43	53.60	**63.87**	48.02	13.27	14.89	**74.93**	**60.12**	45.42	64.19	**66.71**	53.14
GPT-4o	MedRAG	92.39	73.80	51.29	38.00	15.77	24.11	/	/	/	/	/	/
	PRG	**92.56**	**75.60**	60.74	52.32	19.97	**27.18**	53.86	51.33	57.27	55.27	52.45	56.26
	PGC	87.70	50.60	**67.01**	**61.72**	**20.80**	21.37	79.59	75.94	81.01	82.40	80.29	79.04
	MedCite	**92.56**	**75.60**	60.74	52.32	19.97	**27.18**	**84.86**	**84.54**	**83.85**	**89.43**	**84.36**	**86.48**

Table 5: Comparison results of MedCite and alternative methods on BioASQ and PubMedQA datasets. * The generation phase for PGC utilizes CoT, which is non-RAG. Consequently, the Accuracy (EM) score for PGC is the same as that of the CoT (non-RAG) method.

general-domain citation methods from recent work, including the post-retrieval generation and post-generation citation method across different backbone LLMs: (1) **MedRAG**: The method described in (Xiong et al., 2024a). (2) **Post-retrieval generation (PRG)**: Following the method in (Gao et al., 2023c), we prompt LLMs with a query, a list of retrieved documents and instruct the LLMs to include citations in their generated answer. (3) **Post-generation citation (PGC)**: Following RARR (Gao et al., 2023a), We perform chain-of-thought (CoT) prompting (Wei et al., 2022b) to let LLM generate an answer, followed by the re-retrieval + LLM reranking to assign citations to each statement. We evaluate three models: Llama-3-8B-I. (Llama-3-8B-Instruct) (MetaAI, 2024), UltraMedical (Zhang et al., 2024), and commercial LLM GPT-4o (gpt-4o-0806) (OpenAI, 2024).

We present the main results in Table 5. The main takeaways from the experiments are as follow.

Generated responses remain correct with enabled citations. State-of-the-art medical QA systems such as MedRAG do not have citations in their generated answers. We show that it is possible to enable citations in medical systems while maintaining the correctness of generated answers. In particular, both MedCite and PRG are able to achieve comparable accuracy, MAUVE, and ROUGE scores to MedRAG on Llama-3-8B-I. and GPT-4o while providing citations to support generated answers. On the other hand, UltraMedical obtains the highest accuracy with MedRAG despite with an absolute accuracy (74.92%) much lower than Llama-3-8B-I. (82.85%) and GPT-4o (92.39%). By examining the generated output from UltraMedical, we find that adding additional instructions seems to confuse the model, leading to incorrect responses. This can be because UltraMedical was trained with a context length of 2048, making it harder for the model to focus on the most relevant parts of the prompt as additional instructions are provided.

MedCite outperforms PRG and PGC in citation quality. While both PRG and PGC enable citation for medicine, MedCite outperforms the two methods by a large margin (e.g., 71.74% vs. 66.96% and 32.50% on BioASQ). MedCite outperforms PRG because MedCite's second pass of citation seeking leverages post-generation non-parametric retrieval to refine the citations, which allows LLMs to mitigate citation hallucinations. MedCite obtains better performance than PGC, because it exploits pre-generation retrieval and LLM's internal parametric knowledge to obtain an initial set of citations, which turns out to be useful for obtaining high-quality final citations. These results have demonstrated MedCite-Hybrid's effectiveness in combining the strengths of both generative and retrieval systems for citation generation.

MedCite consistently brings citation quality improvements over different LLMs. We see a universal trend that MedCite improves citation recall and F_1 score across LLMs. Using GPT-4o as the backbone LLM leads to the highest-performing citation quality (e.g., GPT-4o 86.48 vs. Llama-3-I. 68.60 in F_1 on PubMedQA), mainly driven by its advanced reasoning and instruction following capabilities. In contrast, citation quality is the lowest when the system is evaluated on UltraMedical (e.g., 66.71 on BioASQ). These results underscore that incorporating MedCite bolsters LLM's capacity to generate verifiable texts.

5.2 Ablation Studies

Citation retrieval analysis. We evaluate how different citation retrievers affect the quality of MedCite. In particular, we compare semantic-only (Jin et al., 2023), lexical-only (Robertson and Zaragoza, 2009), and retrieval-fusion via RRF-2 (Xiong et al., 2024a), and hierarchical two-stage retriever. Different from prior findings that the RRF-2 based hybrid retriever leads to the best performance results, we find that lexical-only (e.g., BM25) retriever leads to the higher citation quality. Unlike the retriever used in RAG, which aims to provide supporting documents for LLM generation, citation retrieval requires examination of precise medical terminology and quoting verbatim from the source. For example, in our experiment, given the LLM claim "peptides are short chains of amino acids, and chlorotoxin is a specific type of peptide," the semantic retriever retrieves a document discussing the features of calitoxin. Although both calitoxin and chlorotoxin are toxins, the document does not help support the claim. Therefore, it cannot serve as a valid citation for this statement. Because of this, a lexical retriever based on exact match provides more precise citations. In contrast, semantic-only and retrieval-fusion based retrievers negatively affect the citation quality. Finally, the hierarchical two-stage retriever fist performs lexical retrieval to obtain a long list of citation candidates followed by a semantic retriever to rank the long list by the similarity score between the query and the citation candidates. As a result, it offers the best-performing results among our tested configurations by achieving a good trade-off between citing comprehensively and precisely.

Retriever Type	Method	Accuracy (EM)	Citation Quality	
			Rec.	Prec.
Lexical-only	BM25	94.34	77.53	79.89
Semantic-only	MedCPT	94.34	65.93	66.78
Combination	RRF-2	94.34	75.74	76.46
Hierarchical	BM25 then MedCPT	94.34	**77.84**	**80.02**

Table 6: Effectiveness of different retrievers on MedCite quality with Llama-3-8B-I. using Oracle relevant documents as the supporting documents in the pre-generation retrieval stage and re-retrieve top-3 documents per statement with LLM reranking.

Attribution judge analysis and human annotations. We dig deeper by evaluating the performance of various models in making attribution judgments for medical tasks and comparing their results with the judgments of the professional medical doctor. We include several SoTA medicine-specialized NLI models such as SciFive-Pubmed+PMC Large on MedNLI(Phan et al., 2021), public top-performing LLMs such as Llama-3.1-I. (Dubey et al., 2024) and Mistral-Instruct (Jiang et al., 2023), and commercial LLMs such as GPT-4o. To assess the correlation between the model judgment and the human annotation of attributability, we ask medical doctors to manually annotate 100 pairs of statements and citations, using the same judgment criteria described in § 2.3 with the guidelines in Appendix E. We calculate Cohen's Kappa score (McHugh, 2012), providing a measure of agreement. Surprisingly, Table 7 indicates that existing medicine-specialized NLI models exhibit poor correlation with professional medical doctor judgments (e.g., <22.3% score in precision judge). Also interestingly, GPT-4o/GPT-3.5 are not the top performing models in this context. Instead, public models such as Llama-3.1 and Mistral achieve the best correlation with expert judgments, demonstrating a higher level of agreement with medical professionals. We include some hypothesis on why this happens in Appendix E.3. Nevertheless, given the high correlation between recent top-performing LLMs and expert judgments, we consider using LLMs as attribution judgments to be more promising for medicine, and we see this as an opportunity for future work.

Model	Source	Domain	Cohen's Kappa Score	
			Rec. Judge	Prec. Judge
SciFive-MedNLI	Open	Medical	0.2593	0.1945
JSL-MedPhi2-2.7B	Open	Medical	0.1845	0.2218
UltraMedical	Open	Medical	0.4518	0.2162
Llama-3.1-8B-Instruct	Open	General	0.5862	**0.5422**
mistral-7B-Instruct	Open	General	**0.6211**	0.4241
GPT-3.5-Turbo	Close	General	0.3834	0.4075
GPT-4o	Close	General	0.4146	0.4075
GPT-4o-mini	Close	General	0.3834	0.3894

Table 7: Correlation of different models' attribution judge with human annotations.

6 Conclusion

We introduce MedCite, the first end-to-end framework fostering research that targets improving the verifiability and trustworthy of medical systems with citations. Our in-depth examination of important design choices for LLM-based medical systems inspires us to propose MedCite-Hybrid, a novel method for generating high quality citations for medical systems. Extensive evaluation across LLMs show that our approach leads to consistent improvements to citation generation over alternative methods.

7 Ethical Considerations and Limitations

The primary goal of this work is to assess and improve the verifiability in LLM-based medical systems via citations. In addition to gaining trust from physicians and patients, there is also urgency from regulation and audition consideration, where the US Food and Drug Administration (FDA) has called for regulation methods for using LLMs in the medical industry (Baumann, 2024). However, incorrect citations can have serious consequences in the medical field, as they can affect patient and physician's treatment decisions. As such, the deployment of LLM-based systems in medical contexts requires careful design while adhering to ethical considerations, e.g., the system should augment human decision-making rather than replace it, and human oversight remains critical to validate generated citations.

While conducting this research, we have also identified several critical yet unexplored challenges in generating citations for medicine. For instance, manual human verification by professional medical doctors remains a costly and time-intensive process, making it difficult to scale. Additionally, whether a document supports a statement can be subject to interpretation, even among medical experts, who may disagree on the extent to which a document partially supports a statement. Therefore, it is crucial to assess whether a high level of consensus among doctors can be achieved. Another challenge is the limited availability of medical datasets that include both ground truth answers and supporting documents, aside from BioASQ and PubMedQA. The absence of certain information, such as ground truth references, in medical datasets complicates the overall verifiability evaluation in medicine. Future work should focus on developing high-quality citation datasets for medicine, which would significantly enhance the trustworthiness and effectiveness of medical QA systems, ultimately benefiting healthcare professionals and patients.

While MedCite is specifically tailored to the medical domain, its generalization to other fields presents notable challenges. Key components, such as multi-pass citation generation and curated database reliance, may not translate directly to general-domain applications without significant modifications. For example, the availability of well-curated corpora like PubMed is unique to the medical field. General domains often lack centralized resources, requiring extensive dataset preparation or integration of diverse sources. Similarly, retriever selection, such as the use of MedCPT in this study, may need to be adapted to align with the characteristics and retrieval objectives of different fields. The effectiveness of retrieval configurations and strategies could vary significantly depending on corpus diversity and domain-specific needs. Moreover, citation evaluation strategies may need to accommodate varying requirements across domains. In medicine, most claims necessitate citations due to high stakes and reliance on specialized knowledge, whereas general domains may involve claims rooted in common sense or widely accepted facts. Evaluating citations in such contexts might require adjustments to account for optional citations or more loosely defined relevance criteria. Automatic evaluation approaches, while valuable, would also need adaptation to handle the simpler or binary relationships typical of claims and citations in general fields. These limitations suggest that while MedCite's core framework offers a strong foundation, further work is needed to ensure its components are broadly applicable to non-medical domains.

References

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? *CoRR*, abs/2305.18248.

Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. 2023. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv preprint arXiv:2311.01463*.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and

Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net.

Jeannie Baumann. 2024. ChatGPT Poses New Regulatory Questions for FDA, Medical Industry. https://news.bloomberglaw.com/us-law-week/chatgpt-poses-new-regulatory-questions-for-fda-medical-industry. Accessed: 14-October-2024.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *CoRR*, abs/2212.08037.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7b parameter language model trained on biomedical text. *Preprint*, arXiv:2403.18421.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

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 *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*.

Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1).

Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont D. Antieau, Andrew S. Bennett, James J. Cimino, John W. Ely, and Hong Yu. 2011. Askhermes: An online question answering system for complex clinical questions. *J. Biomed. Informatics*, 44(2):277–288.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019)*.

Hanane Djeddal, Pierre Erbacher, Raouf Toukal, Laure Soulier, Karen Pinel-Sauvagnat, Sophia Katrenko, and Lynda Tamine. 2024. An evaluation framework for attributed information retrieval using large language models. *CoRR*, abs/2409.08014.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. Rarr: Researching and revising what language models say, using language models. *Preprint*, arXiv:2210.08726.

Luyu Gao et al. 2023b. Rarr: Researching and revising what language models say, using language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen.

2023c. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6465–6488. Association for Computational Linguistics.

Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt, and Paul Wicks. 2023. Large language model ai chatbots require approval as medical devices. *Nature Medicine*, 29(10):2396–2398.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

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*. OpenReview.net.

Jie Huang and Kevin Chang. 2024. Citation: A key to building responsible and accountable large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 464–473. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John C. Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024. Position: Trustllm: Trustworthiness in large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Siqing Huo, Negar Arabzadeh, and Charles L. A. Clarke. 2023. Retrieving supporting evidence for generative question answering. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2023*, pages 11–20. ACM.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2567–2577. Association for Computational Linguistics.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinform.*, 39(10).

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.

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.

Minsuk Lee, James J. Cimino, Hai Ran Zhu, Carl L. Sable, Vijay Shanker, John W. Ely, and Hong Yu. 2006. Beyond information retrieval - medical question answering. In *AMIA 2006, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 11-15, 2006*. AMIA.

Xinze Li, Yixin Cao, Liangming Pan, Yubo Ma, and Aixin Sun. 2024. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 493–516. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2356–2362, New York, NY, USA. Association for Computing Machinery.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7001–7025. Association for Computational Linguistics.

Alejandro Lozano, Scott L Fleming, Chia-Chun Chiang, and Nigam Shah. 2023. Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, pages 8–23. World Scientific.

Bethanie Maples, Merve Cerit, Aditya Vishwanath, and Roy Pea. 2024. Loneliness and suicide mitigation for students using gpt3-enabled chatbots. *npj mental health research*, 3(1):4.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

MetaAI. 2024. Introducing Meta LLaMA-3. https://ai.meta.com/blog/meta-llama-3/.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Salvador Lima-López, Eulàlia Farré-Maduell, Martin Krallinger, Natalia V. Loukachevitch, Vera Davydova, Elena Tutubalina, and Georgios Paliouras. 2024. Overview of bioasq 2024: The twelfth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II*, volume 14959 of *Lecture Notes in Computer Science*, pages 3–27. Springer.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024. OpenAI GPT-4o API. https://platform.openai.com/docs/models/gpt-4o.

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.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *Preprint*, arXiv:2106.03598.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Comput. Linguistics*, 49(4):777–840.

Revanth G. Reddy, Yi R. Fung, Qi Zeng, et al. 2023. Smartbook: Ai-assisted situation report generation. *CoRR*, abs/2303.14337.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean-Baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Benjamin Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Andrew Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale R. Webster, Joelle K. Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of gemini models in medicine. *CoRR*, abs/2404.18416.

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.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net.

Pasin Tangadulrat, Supinya Sono, Boonsin Tangtrakulwanich, et al. 2023. Using chatgpt for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. *JMIR Medical Education*, 9(1):e50658.

Mohamad-Hani Temsah, Fadi Aljamaan, Khalid H Malki, Khalid Alhasan, Ibraheem Altamimi, Razan Aljarbou, Faisal Bazuhair, Abdulmajeed Alsubaihin, Naif Abdulmajeed, Fatimah S Alshahrani, et al. 2023. Chatgpt and the future of digital health: a study on healthcare workers' perceptions and expectations. In *Healthcare*, volume 11, page 1812. MDPI.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera y Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in 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*.

Orion Weller, Marc Marone, Nathaniel Weir, Dawn J. Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. "according to . . . ": Prompting language models improves quoting from pre-training data. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers*, pages 2288–2301. Association for Computational Linguistics.

Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E. Ho, and James Zou. 2024. How well do llms cite relevant medical references? an evaluation framework and analyses. *Preprint*, arXiv:2402.02008.

Amelie Wührl, Yarik Menchaca Resendiz, Lara Grimminger, and Roman Klinger. 2024. What makes medical claims (un)verifiable? analyzing entity and relation properties for fact verification. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. *CoRR*, abs/2408.00727.

Hua Yang, Shilong Li, and Teresa Gonçalves. 2024. Enhancing biomedical question answering with large language models. *Inf.*, 15(8):494.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. *Preprint*, arXiv:2210.09338.

Xi Ye, Ruoxi Sun, Sercan Ö. Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation generation. *Preprint*, arXiv:2311.09533.

Haoran Yu, Chang Yu, Zihan Wang, Dongxian Zou, and Hao Qin. 2024. Enhancing healthcare through large language models: A study on medical question answering. *CoRR*, abs/2408.04138.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Jinfang Hu, Zhiyuan Liu, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine. *CoRR*, abs/2406.03949.

Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. Chatgpt hallucinates when attributing answers. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2023*, pages 46–51. ACM.

A Related Work

A.1 Biomedical Question Answering

Biomedical question answering (QA) is a specialized field within natural language processing. It focuses on answering questions related to biomedical and clinical domains. Early approaches rely heavily on rule-based systems (Lee et al., 2006; Cao et al., 2011). These methods utilize structured databases and ontologies to retrieve answers to clinical questions. While pioneering, these systems were limited by their reliance on predefined rules and lack of scalability. Subsequently, ML/DL based solutions have brought significant improvements to biomedical QA. Models such as BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2019) adapt pretrained BERT (Devlin et al., 2019) to biomedical texts, resulting in improved performance on various biomedical QA tasks (Yang et al., 2024). Recently, generative models represent a newer paradigm in biomedical QA. Models such as GPT-3.5/4 (Brown et al., 2020; OpenAI, 2023) and Med-Gemini (Saab et al., 2024) generate answers directly from input text without relying on predefined answer options, which enable more flexible and contextually appropriate responses. However, generative models also pose challenges, such as the risk of generating incorrect or hallucinated answers. To tackle the issue, recent work employs retrieval-augmented generation (RAG) to retrieve relevant documents and generate answers based on the retrieved information (Lozano et al., 2023; Xiong et al., 2024a; Yang et al., 2024; Yu et al., 2024; Zakka et al., 2024; Xiong et al., 2024b). Different from these efforts, we focus on improving the verifiability of medical systems.

A.2 Citation Methods for LLM Generation

The integration of citation mechanisms in LLM based generation is a burgeoning area of research. Recent advancements in LLMs can be prompted to include citations in the text it generates (Brown et al., 2020; Thoppilan et al., 2022; Anil et al., 2023; OpenAI, 2023, 2024). However, the accuracy and relevance of these citations can be a challenge. Similar as hallucination in generated texts, the model (e.g., ChatGPT) can generate plausible-looking citations that are not actually accurate or verifiable (Zuccon et al., 2023). Multiple methods have been proposed to add citations to LLM-generated content. Direct model-driven attribution methods allow the model to self-attribute, though this often leads to unreliable results (Sun et al., 2023; Agrawal et al., 2023; Weller et al., 2024). Post-retrieval generation (PRG) involves retrieving a list of documents relevant to the user query before generating an answer and the relevant documents (Guu et al., 2020; Borgeaud et al., 2022; Reddy et al., 2023). Post-generation citation (PGC) seeks relevant documents after generating the answer (Huo et al., 2023). Both PRG and PGC offer more reliable attribution but increase system complexity (Gao et al., 2023b), and as we show in the paper, they may not achieve the optimal citation quality for medicine systems due to the nuanced nature of biomedical queries and the need for precise, verifiable citations. Our hybrid double-pass citation method aims to address these gaps by integrating RAG with post-generation refinement. Fine-tuning LLMs for citation generation represents another approach, where models are trained using curated or synthetic data to directly produce citations during text generation (Ye et al., 2024). Finally, there has been overall an absence of automated evaluation for the citation methods over LLM-based QA. Therefore, there has been efforts that aim to improve the evaluation protocols and benchmarks for LLM attributions (Rashkin et al., 2023; Gao et al., 2023c; Li et al., 2024). Different from those efforts, which measures citations for general domain subjects, our evaluation is medicine-centric and we also explore the other components, such as medical-specific retrieval and citation seeking strategies that impact LLM based medicine tasks.

A.3 Evaluation Frameworks for LLM-Generated Citations in Medical Domain

Wu et al. (2024) introduced an evaluation pipeline for assessing the validity of LLM-generated citations in medical domain, focusing on URL-based online sources. Their work highlights significant limitations in LLM citation quality, with even top-performing models like GPT-4 (RAG) failing to fully support all statements in nearly half of their responses.

While both our study and Wu et al. share the goal of improving citation reliability, our work differs in scope and methodology. Wu et al. provides a comprehensive evaluation pipeline, with primary focus on analyzing citation quality for parametric methods by prompting API-based LLMs to provide source URL in their answer, rather than proposing methods to address identified gaps. In contrast, our work not only evaluates but also introduces a modular framework combining hierarchical retrieval and multi-pass citation to improve citation quality for biomedical tasks. In Section 3.1, we explain why parametric methods are unsuitable especially for open-source LLMs due to challenges such as fabricated citations, lack of access to reliable content, and the difficulty of automatic evaluation without API-level access to online sources. By emphasizing domain-specific hierarchical retrieval from curated medical sources like PubMed, we address challenges unique to the biomedical domain, such as ensuring precision for highly specialized terms like drug names or genomic markers.

B Prompt Templates

Prompts template for CoT-generation

```
You are a helpful medical expert, and your task is to answer a
multi-choice medical question."'

general_cot = Template("'
Here is the question:
{{question}}

Here are the potential choices:
{{options}}

Please first think step-by-step and then choose the answer from
the provided options. Organize your output in a json formatted as
Dict{"step_by_step_thinking": Str(explanation),
"answer_choice": Str{A/B/C/...}}. Your responses will be used
for research purposes only, so please have a definite answer.
```

Figure 4: Prompt templates used for CoT generations.

Prompts template for Parametric-citation-generation

```
You are a helpful medical expert, and your task is to answer a
multi-choice medical question using the relevant documents.
Please first think step-by-step and then choose the answer from
the provided options. Organize your output in a json formatted
as Dict{"step_by_step_thinking": Str(explanation),
"answer_choice": Str{A/B/C/...}}. Your responses will be used
for research purposes only, so please have a definite answer.

Here is the question:
{{question}}

Here are the potential choices:
{{options}}

Please think step-by-step and generate your output in json and
include citations after each sentence. When citing several
documents, use [1][2][3]. Cite at least one document and at
most three documents in each sentence. Organize your output in
a json formatted as Dict{"step_by_step_thinking":
Str(explanation), "answer_choice": Str{A/B/C/...},
"references:Str(References)"}:
```

Figure 5: Prompt templates for parametric citations.

Prompts template for MedRAG-generation

```
You are a helpful medical expert, and your task is to answer a
multi-choice medical question using the relevant documents.
Please first think step-by-step and then choose the answer from
the provided options. Organize your output in a json formatted
as Dict{"step_by_step_thinking": Str(explanation),
"answer_choice": Str{A/B/C/...}}. Your responses will be used
for research purposes only, so please have a definite answer.

Here are the relevant documents:
{{context}}

Here is the question:
{{question}}

Here are the potential choices:
{{options}}

Please think step-by-step and generate your output in json:
```

Figure 6: Prompt templates for MedRAG generations.

Prompts template for MedRAG-cite-generation

```
You are a helpful medical expert, and your task is to answer a
multi-choice medical question using the relevant documents.
Please first think step-by-step and then choose the answer from
the provided options. Organize your output in a json formatted
as Dict{"step_by_step_thinking": Str(explanation),
"answer_choice": Str{A/B/C/...}}. Your responses will be used
for research purposes only, so please have a definite answer.

Here are the relevant documents:
{{context}}

Here is the question:
{{question}}

Here are the potential choices:
{{options}}
Please think step-by-step and generate your output in json
using only the provided documents (some of which might be
irrelevant) and cite them using [] after which your statement is
based on it. When citing several documents, use [1][2][3]. Cite
at least one document and at most three documents in each
sentence. Organize your output in a json formatted as
Dict{"step_by_step_thinking": Str(explanation),
"answer_choice": Str{A/B/C/...}}:
```

Figure 7: Prompt templates for MedRAG plus citation generations.

C Examples of Generated Medical Citations

Table 8 shows examples of generated medical references with parametric citation method using Llama-3-8B-I., UltraMedical, and GPT-4o. For

Prompts template for Citation-seeker

You are a helpful medical expert and your task is to try to find documents that supports the statement, given relevant documents.

Here are the relevant documents:
{{context}}

Here is the statement:
{{statement}}

Output only the document IDs with which supports the statement.
Do not output other things.

Figure 8: Prompt templates for citation seekers.

Prompts template for Attribute Judge

Recall:
You are a helpful medical expert.
Based on the document, determine whether the statement is fully supported or not.
Options:
- Fully Supported: The statement is fully supported by the document.
- Not Fully Supported: The statement is not fully supported by the document.
Provide only your chosen option.
Document: {{premise}}
Statement: {{hypothesis}}

Precision:
You are a helpful medical expert.
Based on the document, determine whether it supports the statement.
Options:
- Fully Support: The document fully supports the statement.
- Partial Support: The document supports part of the statement, but some parts are missing.
- Cannot Support: The document cannot support the statement.
Provide only the chosen option.
Document: {{premise}}
Statement: {{hypothesis}}

Figure 9: Prompt templates for attribute judge.

Llama-3-8B-I., the URL provided in Reference [1] is incorrect, and References [2] and [3] have different authors despite having the same title. Upon inspection, it was found that the article in question does not exist. UltraMedical includes poorly formatted in-line citations and fabricated references. GPT-4o, on the other hand, provides correct references, but they are challenging to evaluate due to limited API access to the sources.

D Additional Experimentation Details

D.1 Datasets

We use medical question answering datasets that have ground truth answers to evaluate MedCite. Specially, we use BioASQ (Nentidis et al., 2024) and PubMedQA (Jin et al., 2019) in the final evaluation. In both cases, we only use questions and remove all ground truth supporting contexts, which represents a more realistic setting as often no demonstrations are provided in real usage scenarios. Table 9 summarizes the details about these two datasets.

PubMedQA. PubMedQA is a dataset for biomedical question answering (QA) tasks. The questions are either the titles of existing research articles or derived from them. The context provides the abstract of the article. The answer includes a ground truth answer to the question, which is derived from the conclusion of the abstract.

BioASQ-Y/N. BioASQ-Y/N is also a biomedical QA dataset. For each instance in the dataset, it contains a question, contexts that provide the information to answer the question, and human annotated answers.

D.2 Hyperparameters

To ensure reproducibility, we use greedy decoding for all LLMs. For retrieval, we use a hierarchical two-stage ranking process: (1) BM25 implemented with Pyserini (Lin et al., 2021) using default hyperparameters for indexing, and (2) MedCPT Cross-Encoder[1] with default settings to rank the retrieved documents for a given query. We retrieve the top-32 documents for answer generation, ensuring they fit within the model's context window, and discard those with lower similarity scores if necessary. We retrieve top-3 documents for a single statement when seeking citation after answer generation.

D.3 Correlation between Rouge-L and Accuracy

The following table illustrates the relationship between ROUGE scores and accuracy under different conditions:

The table clearly demonstrates a positive correlation between ROUGE scores and accuracy. Specifically, when the system prompt is introduced to the MEdRAG model, the ROUGE score increases from 17.04 to 17.98, and the accuracy also improves from 0.8414 to 0.8576. This indicates that by optimizing the prompts, we can enhance the model's output quality and accuracy to some extent. Moreover, when a new prompt is introduced, although the ROUGE score slightly decreases, the accuracy drops more notably, suggesting that the new prompt may have affected the model's performance in certain aspects. Most notably, when using the oracle relevant documents, both the ROUGE score and accuracy reach their peak values, further confirming the positive correlation between ROUGE scores and the accuracy of the model's output. These results suggest that ROUGE scores can serve as an effective metric to assess and opti-

[1] https://huggingface.co/ncbi/MedCPT-Cross-Encoder

Question: Is obesity related to cognitive decline?

Llama-3-8B-I. answer:

Obesity is a complex condition that is associated with various health risks, including cardiovascular disease, type 2 diabetes, and certain types of cancer [1]. Research has also suggested that obesity may be linked to cognitive decline, as it can lead to chronic inflammation, oxidative stress, and changes in brain structure and function [2]. Studies have shown that obese individuals are at a higher risk of developing dementia and Alzheimer's disease, and that weight loss may help to improve cognitive function [3]. Therefore, it is likely that obesity is related to cognitive decline.

References: [1] World Health Organization. (2018). Obesity and overweight. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>; [2] Yaffe, K., et al. (2013). The association between obesity and cognitive decline in older adults. Journal of the American Medical Association, 310(11), 1155-1162; [3] Luchsinger, J. A., et al. (2013). Obesity and cognitive decline in older adults. Journal of Alzheimer's Disease, 35(2), 247-255.

UltraMedical answer:

Obesity is a condition characterized by excessive body fat that poses a risk to health. It is associated with various health issues, including cardiovascular diseases, diabetes, and certain types of cancer. To determine if obesity is related to cognitive decline, we need to consider the evidence linking obesity to brain health. 1. Research has shown that obesity can lead to changes in the brain's structure and function. For instance, studies have indicated that obesity is associated with reduced brain volume and alterations in brain activity, which could potentially affect cognitive functions. 2. Additionally, obesity is known to increase the risk of metabolic syndrome, which includes conditions such as hypertension, high blood sugar, and high cholesterol. These conditions have been linked to an increased risk of cognitive decline and dementia. 3. Inflammation is another factor to consider. Obesity is a state of chronic low-grade inflammation, and inflammation is a known risk factor for cognitive decline. Adipose tissue, particularly visceral fat, secretes pro-inflammatory cytokines that can affect brain health. 4. Furthermore, obesity can lead to sleep apnea and other sleep disturbances, which are associated with cognitive deficits. Poor sleep quality has been linked to a higher risk of cognitive decline and Alzheimer's disease. 5. Based on the evidence, it is reasonable to conclude that obesity is related to cognitive decline. The mechanisms include direct effects on brain structure and function, metabolic disturbances, inflammation, and sleep disturbances. Therefore, the answer is A. Yes, obesity is related to cognitive decline.

References: [1] Hall, K. D., & Guo, J. (2017). Obesity: portable operating instructions. JAMA, 318(17), 1685-1686; [2] Yaffe, K., et al. (2013). The metabolic syndrome, inflammation, and risk of cognitive decline and dementia: the Health, Aging, and Body Composition Study. JAMA, 310(3), 197-208; [3] Spiegel, K., et al. (2009). Sleep, cognition, and health: emerging research directions. Dialogues in Clinical Neuroscience, 11(4), 667-679.

GPT-4o answer:

Obesity has been linked to cognitive decline through various mechanisms. Excess body fat can lead to inflammation and insulin resistance, both of which are associated with negative effects on brain health [1][2]. Additionally, obesity is a risk factor for cardiovascular diseases, which can impair blood flow to the brain and contribute to cognitive decline [2][3]. Studies have shown that individuals with obesity may have a higher risk of developing conditions like Alzheimer's disease and other forms of dementia [1][3]. Therefore, there is substantial evidence supporting the relationship between obesity and cognitive decline.

References: [1] Smith, E., Hay, P., Campbell, L., & Trollor, J. N. (2011). A review of the association between obesity and cognitive function across the lifespan: implications for novel approaches to prevention and treatment. Obesity Reviews, 12(9), 740-755; [2] Kivipelto, M., Ngandu, T., Fratiglioni, L., Viitanen, M., Kreholt, I., Winblad, B., & Nissinen, A. (2005). Obesity and vascular risk factors at midlife and the risk of dementia and Alzheimer disease. Archives of Neurology, 62(10), 1556-1560; [3] Whitmer, R. A., Gunderson, E. P., Quesenberry, C. P., Zhou, J., & Yaffe, K. (2007). Obesity in middle age and future risk of dementia: a 27 year longitudinal population based study. BMJ, 335(7630), 705.

Table 8: Examples of generated medical references using parametric citation methods. The citations are often incorrect and ill-formatted. Most importantly, they are hard to evaluate due to lack of API access with sources.

Dataset	Size	Question Type	Example Question	GT Answer
PubMedQA*	500	Yes/No/Maybe	Is anorectal endosonography valuable ... ?	yes
BioASQ-Y/N	618	Yes/No	Is medical hydrology the same as Spa ... ?	yes

Table 9: The two datasets used in MedCite experiments.

Conditions	ROUGE-L Score	Accuracy
medrag + medcpt	17.04	0.8414
medrag + MedCPT + system prompt	17.98	0.8576
medrag + medcpt + new prompt	17.34	0.8269
oracle relevant docs	22.00	0.9401

Table 10: Analysis of ROUGE-L Scores and Accuracy under Different Conditions

mize the output quality of Large Language Models (LLMs).

E Annotation Guidelines and Analysis

Below we provide the guidelines we used for the human annotation in Section E.3. We ask annotators to follow these guidelines to make an attribution judge.

E.1 Annotation Guidelines

Citation Recall measures how well the combination of all citations supports the statement.

- For each statement, review all the provided citations (e.g., PubMed articles) as a group.
- Determine if the combined information from these citations fully supports, or cannot fully support the statement.

Citation Precision measures how well each individual citation supports the statement.

- For each citation, evaluate whether it alone fully supports, partially supports, or does not support the statement.
- Repeat this evaluation for each citation independently.

Note: Please only use the abstract of the PubMed article as a citation, not the whole body (only review the abstract instead of the whole article).

Clarification on "Fully Supported": The determination depends on the relationship between the statement and the content in the citation(s).

- **Words Not Mentioned in Articles:** If the word(s) in the statement represent something entirely different from what the article describes (e.g., distinct medical terms with no overlap), the statement cannot be considered "fully supported." In such cases, the support would likely be "not supported" or "partially supported", depending on how closely related the information is.

  If the word(s) describe a subclass or specific instance of a broader concept mentioned in the article (e.g., the article discusses a class of treatments, and the statement mentions one treatment within that class), the citation may qualify as "partially supported".
- **Fully Supported Criteria:** A statement can only be considered "fully supported" if all key terms and concepts in the statement are directly addressed and explicitly supported by the information in the citation(s).

E.2 Examples

- **Statement:** "Fruits like apples are a rich source of Vitamin C."
- **Citation 1:** The article mentions that "fruits like oranges, strawberries, and kiwis are excellent sources of Vitamin C".
- **Citation 2:** The article discusses "apples being nutritious but focuses on their fiber content", without mentioning Vitamin C.

Recall (Combination of Citations): If you look at both citations together, they do not fully support the statement. Although Citation 1 mentions fruits rich in Vitamin C, it does not explicitly include apples, and Citation 2 does not provide relevant information about Vitamin C in apples.

Recall Score: Not supported.

Precision (Each Citation Individually):

- **Citation 1:** Provides "partial support" because it mentions fruits rich in Vitamin C but does not specify apples.
- **Citation 2:** Provides "no support" because it does not mention Vitamin C at all.

E.3 Attribution Judge Analysis and Human Annotations in Details

While prior studies often assume that Natural Language Inference (NLI) models correlate well with human judgements in making attribution evaluation (Gao et al., 2023c; Bohnet et al., 2022), those studies focus on general domain questions. To our knowledge, no study has evaluated the effectiveness of different models in attribution judgment for medical tasks. We evaluate the performance of various models in making attribution judgments for medical tasks and comparing their results with professional medical doctor judgments as discussed in **??**. Surprisingly, Table 7 indicates that existing medicine-specialized NLI models exhibit poor correlation with professional medical doctor judgments (e.g., <22.3% score in precision judge). Also interestingly, GPT-4o/GPT-3.5 are not the top performing models in this context. Instead, public models such as Llama-3.1 and Mistral achieve the best correlation with expert judgments, demonstrating a higher level of agreement with medical professionals. We hypothesize that this could be because public LLMs might have been trained on datasets that include more medical literature, although it is hard to verify because the details of the datasets used for training these models are not publicly disclosed. At the same time, we recognize that reasoning capabilities play a central role in the attribution judgment task, as described in Appendix B, the prompt requires models to evaluate the connection between a premise and a hypothesis based on self-contained excerpts. However, domain knowledge remains essential for interpreting specialized claims(Wadden et al., 2020). For instance, verifying claims such as "Cardiac injury is common in critical cases of COVID-19" requires medical expertise to connect elevated troponin levels with cardiac injury. Thus, better-performing LLMs likely benefit from extensive pretraining on medical datasets, which enhances both reasoning and domain-specific understanding. Nevertheless, given the high correlation between recent top-performing LLMs and expert judgments, we consider using LLMs as attribution judgements to be more promising for medicine, and we see this as an opportunity for future work.

We acknowledge that expert annotations can vary significantly, especially in knowledge-intensive domains such as medicine. For instance, in the SciFact dataset for scientific claim verification, the inter-annotator agreement measured by Cohen's kappa is approximately 0.75 (Phan et al., 2021). Similarly, in our annotation process, we observed Cohen's kappa scores of 0.83 for statement-level recall and 0.66 for citation-level precision. These results indicate a comparable level of agreement to prior studies, despite the inherent challenges of maintaining consistency in complex annotation tasks.