VERICITE: Evaluating Sentence-Level Citation Faithfulness in Retrieval-Augmented Medical Question Answering

Yixian Ma **Bohao Chu** **Norbert Fuhr**
University Duisburg-Essen
{yixian@stud, bohao.chu, norbert.fuhr}@uni-due.de

Abstract

Retrieval-augmented generation (RAG) reduces hallucination in large language models by grounding outputs in retrieved evidence, but it does not guarantee that the resulting citations actually support the claims they accompany. We present VERICITE, a framework for evaluating citation faithfulness in retrieval-augmented medical QA. Our system retrieves PubMed abstracts via the NCBI E-utilities API, prompts LLMs to generate answers with inline citations, and then verifies each citation at the sentence level using a DeBERTa-v3-large NLI model. We evaluate four LLMs on 500 BioASQ questions at retrieval depths $k \in \{3, 5\}$, with extended experiments up to k=15 and an oracle setting with gold-standard documents. Only 27–41% of citation pairs are supported at the sentence level at $k \in \{3, 5\}$, with rates declining further at larger k. Under the oracle condition, answer quality improves but citation faithfulness does not substantially improve, suggesting that generation-side citation behavior contributes substantially to unfaithful citations.

1 Introduction

RAG (Lewis et al., 2020) has become standard practice in medical question answering (QA) because it grounds model outputs in external evidence (Xiong et al., 2024). However, the presence of a citation does not mean the cited source actually supports the claim. A model can attach a real PubMed article to a plausible-sounding statement even when the article says something only tangentially related—what Wallat et al. (2025) call "post-rationalized" citations. In medicine this is especially dangerous: a clinician may trust a cited claim precisely because it appears backed by evidence (Sackett et al., 1996).

We ask a simple question: when an LLM cites a retrieved PubMed abstract, does that abstract actually contain a sentence supporting the claim? VERICITE retrieves abstracts via PubMed E-utilities, prompts an LLM to generate a cited answer, and then checks each citation at the sentence level using NLI. We retrieve at the abstract level to preserve clinical context but verify at the sentence level for precise grounding. Our main contributions are:

1. A biomedical RAG pipeline with sentence-level NLI faithfulness evaluation. While ALCE (Gao et al., 2023) introduced sentence-level citation evaluation in open-domain QA, we adapt this paradigm to biomedical retrieval over PubMed abstracts.

2. An oracle retrieval experiment that approximates an upper-bound retrieval condition and analyzes how citation faithfulness changes under near-perfect recall. Gold-standard documents improve answer quality but not citation faithfulness, providing suggestive evidence that generation-side factors may contribute substantially to unfaithful citations.

3. A comparison of four LLMs across four retrieval depths, showing a precision–coverage trade-off where increasing k improves recall but degrades faithfulness.

2 Related Work

RAG and citation in medical QA. The ALCE benchmark (Gao et al., 2023) introduced sentence-level citation evaluation via citation recall and precision, finding that even strong models fail to support their citations about half the time; however, ALCE targets open-domain QA. MedCite (Du et al., 2025) compared citation strategies for biomedical QA, and the TREC BioGen shared task (Gupta et al., 2024, 2026b) has established benchmarks for evaluating biomedical generative retrieval with expert assessments. Most of this work focuses on producing citations; our focus is on verifying them at fine granularity in the biomed-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 753–759

ical domain and on analyzing citation behavior under different retrieval conditions.

Citation faithfulness. Wallat et al. (2025) drew an important distinction between citation *correctness* (the source is real) and citation *faithfulness* (the source entails the claim), and found that most RAG citations are post-rationalized. Vladika et al. (2025) showed that even GPT-4o with retrieval augmentation lacks citation support for nearly half of its medical responses. BioACE (Gupta et al., 2026a) proposed automated biomedical citation evaluation and highlighted the difficulty of NLI-based assessment in this domain. Self-RAG (Asai et al., 2024) trains a single LM to critique its own retrieval via reflection tokens. We are closest to ALCE and Source-Checkup. ALCE evaluates citation quality in open-domain QA, while we adapt sentence-level evaluation to PubMed-based biomedical QA with both standard and oracle retrieval. Source-Checkup studies citation support in medical QA but lacks this retrieval comparison.

NLI for factual verification. NLI has been widely used for factual consistency checking (Maynez et al., 2020). We use DeBERTa-v3-large (He et al., 2023) fine-tuned on MNLI (Williams et al., 2018), FEVER (Thorne et al., 2018), ANLI (Nie et al., 2020), LingNLI, and WANLI.[1] This model was not trained on biomedical NLI data such as MedNLI (Romanov and Shivade, 2018); we discuss the implications in Section 6.

3 Method

Figure 1 shows the VERICITE pipeline.

3.1 Retrieval and Cited Answer Generation

Given a medical question, we retrieve the top-k PubMed abstracts using the NCBI E-utilities API (Sayers et al., 2022). The question is preprocessed by removing the leading interrogative word, preserving medical terminology for PubMed's Automatic Term Mapping. We retrieve a candidate pool of 200 results with title/abstract field restriction and keep the first k abstracts with complete metadata (Appendix A). Retrieval operates at the abstract level because isolated biomedical sentences often lack key context such as population, intervention, or outcome qualifiers. The abstracts are then passed to an LLM that generates a 4–6 sentence answer with inline document-level citations (Appendix B and E).

We intentionally use simple keyword-based retrieval (E-utilities) rather than dense retrieval, since our goal is to evaluate citation faithfulness across a range of recall levels—from low recall (standard retrieval) to near-perfect recall (oracle)—rather than to optimize retrieval itself.

Oracle retrieval. To separate retrieval from generation effects, we include an oracle condition that uses BioASQ's gold-standard PMIDs directly, bypassing search. This yields recall ≈ 1.0 (99.9%; a few gold PMIDs lack retrievable abstracts). By comparing standard and oracle conditions on the same models, we can examine whether substantially improved retrieval recall is associated with changes in citation faithfulness.

3.2 Sentence-Level Citation Faithfulness

Although citations are produced at the document level, we evaluate faithfulness at the sentence level. For each generated sentence that cites a document, we segment the cited abstract into individual source sentences and run NLI on every (generated claim, source sentence) pair. This is a sentence-to-sentence comparison, not a comparison between the generated sentence and the full abstract. Each source sentence is evaluated independently against the generated claim using the DeBERTa-v3-large NLI model described above.

We aggregate the per-sentence NLI verdicts into a single citation-level verdict for each (generated sentence, cited document) pair: **Supported** if at least one source sentence entails the claim; **Contradicted** if none entails but at least one contradicts; **Irrelevant** otherwise (all source sentences are neutral).

4 Experimental Setup

Data. We use 500 questions (excluding yes/no) from the BioASQ Task 13b training set (Tsatsaronis et al., 2015). BioASQ provides expert-authored ideal answers and gold-relevant PMIDs. Of these, 421 yield at least one abstract under standard retrieval—factoid (161), list (114), summary (146)—and are used for those conditions; per-type results are in Appendix D. The oracle condition uses all 500 questions.

[1] Checkpoint: `MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli` on Hugging Face.

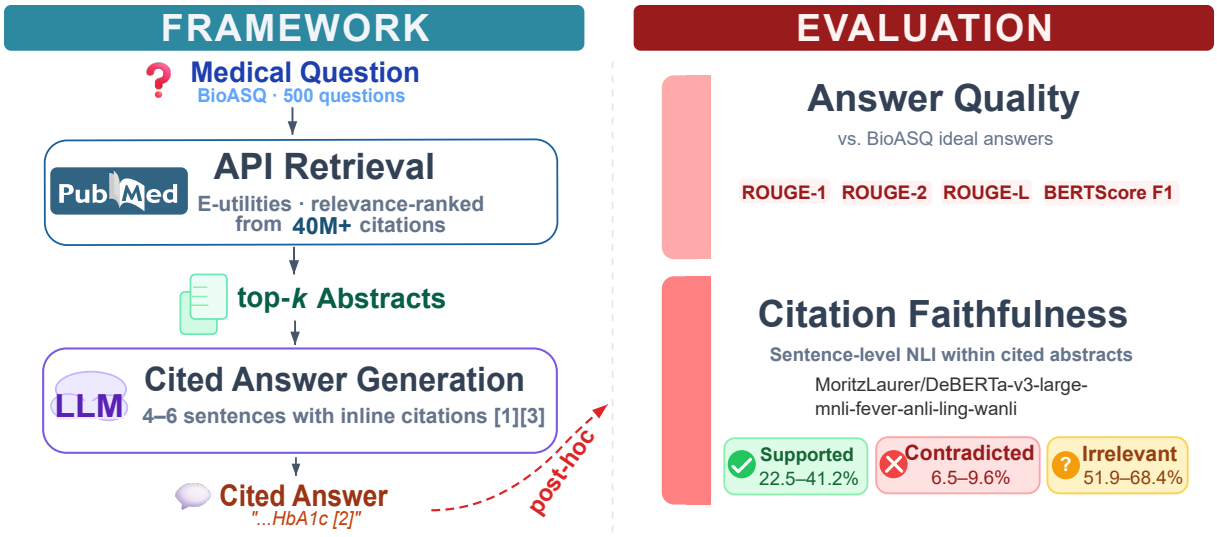

Figure 1: The VERICITE pipeline. **Left**: retrieval of top-k PubMed abstracts and cited answer generation. **Right**: post-hoc evaluation of retrieval quality, answer quality, and sentence-level citation faithfulness via NLI.

Models. We evaluate four LLMs, all given identical retrieval results and the same prompt: Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Llama-3.1-70B-Instruct (Dubey et al., 2024), GPT-4o-mini (OpenAI, 2024), and Claude-3.5-Haiku (Anthropic, 2024), accessed via API (Mistral through DeepInfra; others through OpenRouter). We include both a small open-weight model (Mistral-7B) and larger models to examine whether model scale affects citation behavior. All four run at $k \in \{3, 5\}$; Mistral-7B and Llama-3.1-70B also run at $k \in \{10, 15\}$. Temperature is 0.4 with nucleus sampling p=0.9.

Metrics. We report three groups: (1) *Retrieval*: Recall@k against BioASQ gold PMIDs. (2) *Answer quality*: ROUGE-1/2/L (Lin, 2004) and BERTScore F1 (Zhang et al., 2020) (`roberta-large`) against ideal answers. (3) *Citation faithfulness*: the fraction of (generated sentence, cited document) pairs classified as supported, contradicted, or irrelevant. Note that the ROUGE/BERTScore comparison is not entirely fair, since BioASQ ideal answers were written with access to a broader set of gold documents than our models see.

5 Results

Table 1 shows results for all models at $k \in \{3, 5\}$.

Retrieval. Recall@3 is 0.127 and Recall@5 is 0.162 over the 421 retained questions (79 returned no abstracts and were excluded). This low recall reflects two factors: BioASQ gold references are curated through expert literature review and may not rank highly in keyword search, and our E-utilities retrieval is deliberately simple. We treat this as a lightweight retrieval baseline rather than a competitive biomedical retriever; the oracle condition provides a high-recall counterpart for analyzing citation behavior under substantially improved retrieval coverage.

Citation faithfulness. Under our sentence-level NLI criterion—which may underestimate true faithfulness when models paraphrase or merge information across source sentences (Section 6)—only 29.7–41.2% of citation pairs at k=3 are supported. The majority (51.6–61.7%) are irrelevant: the cited abstract relates to the question topic but no individual source sentence entails the specific claim. Contradictions account for 6.6–9.5%, a smaller but non-trivial fraction in a medical context. These patterns are consistent with Wallat et al. (2025), who found most RAG citations to be post-rationalized.

Among models, Llama-3.1-70B achieves the highest support rate (41.2% at k=3), followed by Claude-3.5-Haiku (38.9%), GPT-4o-mini (37.8%), and Mistral-7B (29.7%). The gap between Mistral-7B and the other models suggests that model capacity may play a role, though we cannot isolate scale from other factors (training data, instruction tuning). Across both retrieval depths, GPT-4o-mini shows the lowest contradiction rate (6.5% at k=5), suggesting relatively conservative citation behavior.

Model	k	Retr. Rec.	Citation Faithfulness Sup.↑	Con.↓	Irr.↓	Answer Quality R-1	R-2	R-L	BS
Mistral-7B	3	.127	.297	.086	.617	.251	.091	.171	.862
	5	.162	.274	.078	.648	.258	.097	.177	.864
Llama-3.1-70B	3	.127	**.412**	.069	.519	.259	.107	.185	.863
	5	.162	.391	.066	.543	**.269**	**.118**	**.196**	**.865**
GPT-4o-mini	3	.127	.378	.066	.556	.249	.090	.171	.862
	5	.162	.360	**.065**	.575	.256	.097	.175	.865
Claude-3.5-Haiku	3	.127	.389	.095	**.516**	.249	.088	.171	.860
	5	.162	.364	.080	.556	.257	.096	.178	.863

Table 1: Main results on BioASQ (421 questions). ↑/↓: higher/lower is better. Best value per column in **bold**. BS: BERTScore F1 (`roberta-large`).

Model	k	Rec.	Sup.	Irr.	R-L	BS
Mistral	3	.127	.297	.617	.171	.862
	5	.162	.274	.648	.177	.864
	10	.207	.240	.667	.178	.865
	15	.226	.227	.684	.181	.865
	orac.	*.999*	*.225*	*.678*	*.212*	*.876*
Llama-70B	3	.127	.412	.519	.185	.863
	5	.162	.391	.543	.196	.865
	10	.207	.364	.560	.213	.869
	15	.226	.334	.590	.223	.871
	orac.	*.999*	*.381*	*.533*	*.256*	*.882*

Table 2: Extended depth ($k \in \{3, 5, 10, 15\}$; 421 questions) and oracle with gold PMIDs (500 questions). Support rates decrease with k; oracle retrieval does not improve faithfulness.

Extended retrieval depth. Table 2 extends the analysis to $k \in \{3, 5, 10, 15\}$ for Mistral-7B and Llama-3.1-70B. As k increases from 3 to 15, Recall@k rises to 0.226 and answer quality improves modestly, but support rates drop monotonically: from 29.7% to 22.7% for Mistral, and from 41.2% to 33.4% for Llama. The reason is straightforward: models cite more documents per answer at larger k (2.5 to 5.5 unique docs for Mistral), but many of the additional citations point to documents only loosely related to the specific claim. This constitutes a precision–coverage trade-off: increasing k improves retrieval recall but degrades citation precision.

Oracle retrieval. With recall of 99.9%, citation faithfulness does not substantially improve under oracle retrieval: support drops to 22.5% for Mistral-7B (vs. 29.7% at k=3) and remains comparable for Llama at 38.1% (vs. 41.2%). Meanwhile, answer quality improves clearly (ROUGE-L: .171 → .212 for Mistral; .185 → .256 for Llama). This divergence—better answers but not substantially better citation faithfulness—is one of our central empirical observations, suggesting that improved retrieval evidence alone may be insufficient to ensure faithful citation attribution.

The oracle setting differs from standard retrieval in both document count and question coverage, complicating strict causal interpretation. Increased document availability may encourage models to cite more sources, thereby reducing sentence-level support rates under our evaluation protocol. Nevertheless, the comparison with k=15 partially mitigates this concern, since Llama cites a comparable number of documents yet still achieves lower support rates than under oracle retrieval.

6 Conclusion

We presented VERICITE, a pipeline for evaluating citation faithfulness in retrieval-augmented medical QA. Across four LLMs at $k \in \{3, 5\}$, only 27–41% of citations are supported under sentence-level NLI evaluation; the majority are irrelevant. These rates generally decline at larger k and do not substantially improve under oracle retrieval. The oracle experiment reveals a notable divergence: providing models with expert-selected documents improves answer quality but does not substantially improve citation faithfulness. These findings suggest that retrieval quality alone may be insufficient to ensure faithful citation behavior, and that generation-side citation behavior likely plays an important role. For medical RAG systems, this means improving retrieval alone is unlikely to produce trustworthy citations; generation-side interventions such as citation-aware training or post-hoc verification are needed.

Limitations

First, we rely on a general-domain NLI model (DeBERTa-v3-large) not trained on biomedical text. Domain-specific models trained on MedNLI (Romanov and Shivade, 2018) might yield different re-

sults, and Gupta et al. (2026a) have highlighted the difficulty of NLI in biomedical settings. Without human validation, our faithfulness estimates should be treated as approximate; annotating a sample of NLI verdicts is our most important next step.

Second, our sentence-level entailment approach has known failure modes. When an LLM synthesizes information from multiple source sentences, no single source sentence may fully entail the combined claim, leading to false negatives. A generated sentence may also contain multiple sub-claims that we do not decompose. These factors mean our evaluation likely *underestimates* true faithfulness. A non-zero decoding temperature (0.4) may further increase paraphrasing diversity, reducing literal overlap with source sentences and lowering NLI entailment rates.

Third, the oracle experiment should be interpreted cautiously. In addition to near-perfect retrieval recall, the oracle setting also differs from the standard retrieval condition in both document count and question coverage, making strict causal interpretation difficult. As a result, the observed divergence between answer quality and citation faithfulness provides suggestive rather than definitive evidence that retrieval quality alone is insufficient to guarantee faithful citation behavior. A fully controlled comparison—for example, restricting oracle retrieval to the same question subset and matched document counts—would provide a stronger test.

Fourth, we evaluate on a single dataset (BioASQ) and retrieve only abstracts; full-text articles might provide sentence-level support that abstracts lack. Finally, API-based models may change across versions; we report exact model identifiers and fix decoding parameters to mitigate this.

Future Work

Our most pressing next step is human annotation of NLI verdicts to quantify evaluation reliability and compare the general-domain DeBERTa model against biomedical alternatives. We also plan to explore multi-sentence evidence aggregation (matching claims against groups of source sentences) to reduce false negatives, and to investigate dense retrieval and query expansion to improve recall without degrading citation precision.

Ethics Statement

This work evaluates citation reliability in medical QA and is not intended as a patient-facing system. AI-generated medical content should not substitute professional clinical judgment. All data are from publicly available sources.

References

Anthropic. 2024. Claude 3.5 model card.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*.

Jiawei Du et al. 2025. MedCite: Towards faithful and traceable citation in biomedical question answering. *arXiv preprint arXiv:2501.07653*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2026a. BioACE: An automated framework for biomedical answer and citation evaluations. *arXiv preprint arXiv:2602.04982*.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2024. Overview of TREC 2024 biomedical generative retrieval (BioGen) track. In *Proceedings of the Thirty-Third Text REtrieval Conference (TREC 2024)*. ArXiv:2411.18069.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2026b. Overview of TREC 2025 biomedical generative retrieval (BioGen) track. *arXiv preprint arXiv:2603.21582*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*. ArXiv:2111.09543.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for

knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

OpenAI. 2024. GPT-4o system card.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. *BMJ*, 312(7023):71–72.

Eric W Sayers, Evan E Bolton, J Rodney Brister, Kathi Canese, Jessica Chan, Donald C Comeau, et al. 2022. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1):D206–D214.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Juraj Vladika et al. 2025. Source-checkup: A framework for evaluating source attribution in LLM-generated medical responses. *arXiv preprint arXiv:2501.09940*.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 22–32.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1112–1122.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Retrieval Details

We submit the cleaned query to `esearch` with relevance ranking, restricting matches to title and abstract fields (`[tiab]`) and requiring a non-empty abstract (`hasabstract`). We retrieve 200 candidates via `efetch` and keep the first k with complete metadata. If the strict query returns fewer than k results, we relax the field restriction to allow Automatic Term Mapping, then fall back to the raw question.

For the oracle, we fetch abstracts using gold PMIDs directly. Recall is 99.9% rather than 100% because a few gold PMIDs lack abstracts. On average, models cite 5.9 (Mistral) and 5.3 (Llama) unique documents per answer under oracle.

B Generation Details

Each model receives the question and retrieved abstracts numbered $[1], \ldots, [k]$. The prompt includes a system instruction requiring the model to answer using only the provided documents, cite every sentence, and use multiple sources; two few-shot demonstrations; and the numbered abstracts and question (Appendix E). All models use the same prompt and documents. All four are accessed via API (Mistral-7B through DeepInfra; others through OpenRouter). Temperature is 0.4, p=0.9.

C Dataset Notes

Of 500 BioASQ questions, 421 yield at least one abstract under standard retrieval; the other 79 returned no usable abstracts. Recall@k ranges from 0.127 (k=3) to 0.226 (k=15). The oracle covers

Model	Type	Rec.	Sup.	Irr.	R-L	n
Mistral	Factoid	.108	.302	.603	.163	161
	List	.068	.274	.647	.146	114
	Summary	.195	.308	.610	.200	146
Llama	Factoid	.108	.421	.507	.179	161
	List	.068	.380	.551	.156	114
	Summary	.195	.425	.510	.214	146

Table 3: Citation faithfulness by question type at k=3. List questions show the lowest faithfulness, partly due to lower retrieval recall.

all 500 questions with recall of 0.999. We use questions from the Task 13b training set (2025 release), which accumulates 5,389 questions from previous years with expert-curated gold answers and relevant PMIDs.

D Results by Question Type

Table 3 breaks down citation faithfulness by question type at k=3. Summary questions show the highest support rates (30.8% for Mistral, 42.5% for Llama), likely because summary-type evidence is broader and easier to match. List questions show the lowest faithfulness (27.4% and 38.0%), partly due to low retrieval recall (6.8%). The ranking is consistent across both models.

E Prompt Template

```
You are a medical QA assistant. Below
↪  are
numbered source documents retrieved from
PubMed literature.

RULES:
1. Answer using ONLY the source
↪  documents.
2. Write between 4 and 6 sentences.
3. EVERY sentence MUST end with
↪  citation(s)
   in the format [1], [2][3], or [2,3].
4. Numbers refer to the document
↪  numbers.
5. Cite the most relevant supporting
↪  documents.
6. Do NOT add facts not in the sources.
7. If sources are insufficient, say so.

[Two few-shot demonstrations omitted]

Source documents:
[1] <abstract of document 1>
...
[k] <abstract of document k>

Question: <question>
Answer:
```