The Selection Bottleneck Versus Attention Diffusion: Quantifying Coverage–Latency Tradeoffs in Guideline-Grounded Clinical QA

Ahmed Issaoui[1,2], Jennifer Dawson[3], Roger Zemek[3,4] and Arya Rahgozar[1,2,*]

[1]University of Ottawa, Ottawa, ON, Canada
[2]Ottawa Hospital Research Institute (OHRI), Ottawa, ON, Canada
[3]Children's Hospital of Eastern Ontario (CHEO), Ottawa, ON, Canada
[4]Department of Pediatrics and Emergency Medicine, University of Ottawa, Ottawa, ON, Canada

Abstract

Clinical practice guidelines (CPGs) are difficult to operationalize at the point of care due to time pressure and information overload [1]. We report a pilot evaluation on the Living Guideline for Pediatric Concussion Care [2], treating recommendations as ID-bearing structured knowledge units for grounding and citation checks. Using a retrospective development audit, we reconstruct audit-derived reference sets and compare matched systems: (S1) full-document long-context and (S2) top-k retrieval-augmented generation (RAG; $k \in \{3, 5\}$), both using GPT-5.2 via the OpenAI Responses API (reasoning effort: low). On $N = 52$ questions, S1 achieved macro-averaged recall/precision of 1.000/1.000 against audit-derived reference sets with no invalid citations observed, but higher latency (median ≈ 45 s). RAG had lower median latency (≈ 25 s) and no invalid citations observed, but lower recall: $k = 3$ yielded 0.360/1.000 and $k = 5$ yielded 0.473/0.789. These results are a one-guideline controlled pilot, not evidence that full-document prompting is generally error-free or that fixed top-k RAG is inherently low-latency.

Keywords

Clinical guidelines, Guideline QA, RAG, Long-context LLMs

1. Introduction

Clinical practice guidelines (CPGs) distill evidence but remain hard to apply under time constraints [1]. Question answering (QA) with large language models (LLMs) could help operationalize guidelines, yet clinical deployment requires grounding due to citation fabrication and related unsupported outputs [3]. For knowledge graph (KG) and LLM integration in health, guideline recommendations provide lightweight structured knowledge units: each has an identifier, belongs to a section, and can be cited or omitted by a model answer. We do not claim ontology construction or symbolic graph reasoning.

We compare two grounding strategies. In top-k retrieval-augmented generation (RAG) [4], a *selection bottleneck* occurs when retrieval excludes a recommendation needed for the answer. In full-document long-context prompting, retrieval is removed by providing the whole guideline, but latency may increase and relevant information may be harder to use as context grows. We call this risk *attention diffusion*; "lost in the middle" behavior is one documented manifestation [5]. Unlike answer-level similarity metrics, we evaluate *recommendation-indexed* correctness on a real clinical guideline [2].

Joint Proceedings of the AIME 2026 Workshops: 1st International Workshop on Multicentric and Privacy-preserving Learning in Healthcare, Foundation Models for Public Health and Epidemiology: From Promise to Practice, and First International Workshop on Knowledge Graphs for Health, July 10, 2026, Ottawa, Canada.
*Corresponding author (arahgoza@uottawa.ca).
0009-0000-9691-4875 (A. Issaoui); 0000-0003-1141-9619 (J. Dawson); 0000-0001-7807-2459 (R. Zemek); 0000-0002-2127-2449 (A. Rahgozar)

2. Methods

2.1. Guideline corpus

We evaluated guideline-grounded clinical QA using the *Living Guideline for Pediatric Concussion Care* [2]. The guideline is organized into thematic sections (e.g., diagnosis, acute management, return to activity) and comprises discrete recommendations. Each recommendation contains a short title and an explanatory paragraph. For structured grounding and evaluation, we treated each recommendation as an atomic unit and assigned it a unique identifier (ID). This preserves section–recommendation structure for citation checking, but is not a formally engineered clinical knowledge graph.

2.2. Retrospective internal audit dataset

This study leverages a retrospective audit archive collected during internal development of a guideline-grounded chatbot. The pilot evaluation set consists of $N = 52$ questions. Each audit record contains: (i) a clinical question q, (ii) a generated answer from a development-stage system variant, and (iii) qualitative feedback from an expert clinician reviewer. Two clinicians involved in developing the guideline served as reviewers; each reviewed a disjoint subset of questions. Review feedback included assessment of clinical accuracy and clarity, and also explicitly discussed *coverage* when a response was incomplete by listing missing recommendation IDs. Reviewers also occasionally flagged recommendation IDs that were included but not applicable to the question. Because the review subsets were disjoint, the archive supports expert-informed reconstruction of reference sets but not inter-rater agreement estimation.

2.3. Reconstruction of audit-derived reference sets

For each question q, we reconstruct an audit-derived reference recommendation set $G(q)$ from the archived audited response and reviewer feedback. Let $A(q)$ denote the set of *valid* recommendation IDs cited in the archived answer for q (after normalizing ID format and discarding malformed or non-existent IDs). From the reviewer comments, let $M(q)$ denote the set of recommendation IDs explicitly identified as *missing but required*, and let $W(q)$ denote the set of recommendation IDs explicitly identified as *wrongly added / not applicable*. We define:

$$G(q) = (A(q) \setminus W(q)) \cup M(q). \tag{1}$$

Because $G(q)$ is reconstructed from archived audit artifacts, it should be interpreted as an *audit-derived* reference rather than a prospectively enumerated gold standard.

2.4. Systems and controlled evaluation protocol

We compared two grounding strategies. To isolate the effect of *context provisioning*, both systems used the same base model (GPT-5.2), identical structured output requirements, and the same prompt template. The controlled evaluation compares how guideline context is supplied, not optimized models, prompts, retrieval systems, or decoding settings.

Prompting. For each question, the model input consisted of the user query q and a context block containing guideline recommendations with their IDs. The prompt instructed the model to answer the question using only the provided context and to return output in a fixed format: a short natural-language answer followed by a list of cited recommendation IDs. The only difference between S1 and S2 was how the context block was constructed (full guideline vs. retrieved top-k).

Model configuration. All generations were produced via the OpenAI Responses API with GPT-5.2, low-effort reasoning, and medium output verbosity. These settings were held constant across S1 and S2. Low-effort reasoning was used as a latency- and resource-conscious configuration for a point-of-care

QA setting; it was not tuned separately for either system. Other generation parameters were not varied between systems.

S1: Full-document long-context grounding. System S1 supplied the entire guideline document in the model context for every query, so no retrieval step was performed. Grounding therefore depended entirely on the model's ability to locate and use the relevant recommendations within a long input, rather than on a retriever's ranking. By construction, S1 cannot incur the selection bottleneck: every recommendation—including all of those required for a complete answer—is always present in the context, so a required recommendation that is nonetheless missed reflects attention diffusion (a failure to use information that is present) rather than a retrieval omission. This robustness to omission comes at the cost of placing the full guideline in context on every call, which increases input length and is expected to raise latency relative to retrieving a small top-k context.

S2: Retrieval-augmented generation (RAG). System S2 indexed the guideline at the recommendation level (one document per recommendation) and used a hybrid retriever combining PostgreSQL full-text search (`tsvector`) and dense embeddings (`text-embedding-3-large`, 3072 dimensions). The index was the context source for the controlled RAG runs; the evaluation did not treat the full guideline as part of each RAG prompt. For each query, the lexical and dense similarity scores were each min–max normalized to $[0, 1]$ and combined by an equally weighted sum (0.5 lexical, 0.5 dense); the top-k recommendations ($k \in \{3, 5\}$) by fused score were then provided as context to the model. We deliberately used equal fusion weights and a fixed top-k as a simple, untuned retrieval baseline, so that S2 reflects a standard RAG configuration rather than one optimized for this particular guideline.

2.5. Evaluation metrics and latency measurement

For a system S and question q, let $C_S(q)$ be the set of recommendation IDs cited in the system output. Cited IDs that do not exist in the guideline define the invalid set $C_S^{\text{invalid}}(q)$ [3]. Using the audit-derived reference set $G(q)$, we compute the following recommendation-level metrics:

$$\text{Recall}_S(q) = \begin{cases} \frac{|C_S(q) \cap G(q)|}{|G(q)|}, & |G(q)| > 0, \\ 1, & |G(q)| = 0. \end{cases} \tag{2}$$

$$\text{Precision}_S(q) = \begin{cases} \frac{|C_S(q) \cap G(q)|}{|C_S(q)|}, & |C_S(q)| > 0, \\ 1, & |C_S(q)| = 0. \end{cases} \tag{3}$$

$$\text{InvalidRate}_S(q) = \begin{cases} \frac{|C_S^{\text{invalid}}(q)|}{|C_S(q)|}, & |C_S(q)| > 0, \\ 0, & |C_S(q)| = 0. \end{cases} \tag{4}$$

The conventions for empty reference or citation sets avoid undefined ratios in the pilot metric. They should not be read as evidence of clinical correctness for questions without audited required IDs or for outputs without citations. Per-question scores were aggregated by macro-averaging—the unweighted mean over all $N = 52$ questions—so that each question contributes equally regardless of the size of its reference set $G(q)$ or the number of citations it elicits.

Inference and latency measurement. Latency was measured as wall-clock time from query submission to receipt of the complete response. These measurements are end-to-end observations for this implementation and hosted API setting; they were not decomposed into retrieval, network, and generation components.

3. Results and Discussion (Pilot, $N = 52$)

Table 1 summarizes aggregate recommendation-level performance and latency. In this pilot, full-document grounding reached macro-averaged recall/precision of 1.000/1.000 against audit-derived reference sets, with no invalid citations observed but higher median latency. RAG had lower median latency with no invalid citations observed, but exhibited a recall–precision tradeoff between $k = 3$ and $k = 5$.

Table 1
Pilot results ($N = 52$). Macro-averaged metrics; median latency shown.

System	Recall / Precision	Invalid	Latency
Full document	1.000 / 1.000	0.000	45 s
RAG ($k = 3$)	0.360 / 1.000	0.000	25 s
RAG ($k = 5$)	0.473 / 0.789	0.000	25 s

These results support a practical design interpretation rather than a universal ranking of grounding methods. When a guideline fits the context budget and coverage is primary, full-document grounding can avoid the retrieval-time selection bottleneck observed with fixed top-k RAG. Fixed top-k RAG limits supplied context, but can miss recommendations required for a complete answer. The reported RAG latency is implementation- and API-dependent end-to-end latency, not an inherent lower bound for RAG systems.

Recall–precision tradeoff across k. The RAG results make the coverage cost of retrieval concrete. Increasing k from 3 to 5 raised recall (0.360 to 0.473), because supplying more recommendations increases the chance that a required one is included; but precision fell (1.000 to 0.789), because the additional retrieved recommendations were not always part of the audit-derived reference set yet were sometimes cited. At $k = 3$ every cited recommendation fell within the reference set (precision 1.000), while only about a third of the required recommendations were recovered (recall 0.360)—the signature of a selection bottleneck, in which recommendations needed for a complete answer are excluded before generation. Even at $k = 5$ recall remained below 0.5, indicating that for many questions the required recommendations lay outside the top five and no amount of re-reading the retrieved context could recover them. Full-document grounding removed this ceiling by making every recommendation available in context, reaching complete recall without a precision penalty in this pilot; we interpret this as a consequence of removing the retrieval filter rather than of superior model reasoning, since the same model and prompt were used throughout.

Reproducibility and limitations. The prompt structure, context construction, ID extraction target, and scoring rules are described above, but the retrospective audit archive was internal development material and is not released with this camera-ready paper; exact question-level reproduction is therefore not possible from the paper alone. Reference sets were reconstructed from audit artifacts rather than prospectively defined gold standards and may omit relevant recommendations or partially reflect the audited answer. Review was disjoint, precluding inter-rater reliability estimates. Generalizability is limited by one guideline, $N = 52$, one model configuration, one hosted API setting, and no comparison across alternative retrievers or models. The pilot also did not separately analyze exact guideline-length sensitivity, token consumption, monetary cost, or latency breakdowns.

Future work will expand to multiple guidelines and larger, prospectively annotated datasets with overlapping expert review. We will evaluate adaptive retrieval strategies beyond fixed top-k, assess robustness to longer and more complex documents, and examine how model configuration and deployment constraints influence coverage, citation validity, latency, and resource use.

Acknowledgments

We acknowledge the CHEO Research Institute. Funding was provided by the Ontario Brain Institute and the Brain-Heart Interconnectome.

Conflicts of interest

The authors have no competing interests to declare.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-5.2 in order to: Drafting content. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. Lugtenberg, J. S. Burgers, C. F. Besters, D. Han, G. P. Westert, Perceived barriers to guideline adherence: a survey among general practitioners, BMC Family Practice 12 (2011) 98. doi:10.1186/1471-2296-12-98.

[2] PedsConcussion, Ontario Neurotrauma Foundation, Living guideline for pediatric concussion care (PedsConcussion), 2019. URL: https://pedsconcussion.com/. doi:10.17605/OSF.IO/3VWN9.

[3] A. McGowan, Y. Gui, M. Dobbs, S. Shuster, M. Cotter, A. Selloni, M. Goodman, A. Srivastava, G. A. Cecchi, C. M. Corcoran, ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search, Psychiatry Research 326 (2023) 115334. doi:10.1016/j.psychres.2023.115334.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020, pp. 9459–9474.

[5] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts, Transactions of the Association for Computational Linguistics 12 (2024) 157–173. doi:10.1162/tacl_a_00638.