Position: CARE-RAG - Clinical Assessment and Reasoning in RAG

Deepthi Potluri
Department of Computer Science
University of Texas at Austin
deepthi.potluri@utexas.edu

Aby Mammen Mathew
Department of Computer Science
University of Texas at Austin
abymmathew@utexas.edu

Alexander L. Rasgon
Behavioral Science and Psychiatry
University of Texas at Austin
alexander.rasgon@ascension.org

Jeffrey B DeWitt
Department of Computer Science
University of Texas at Austin
jefdewitt.utexas.edu

Yide Hao
Department of Statistics
University of Michigan
yidehao@umich.edu

Junyuan Hong
School of Information
University of Texas at Austin
jyhong@utmail.utexas.edu

Ying Ding
School of Information
University of Texas at Austin
ying.ding@ischool.utexas.edu

Abstract

Access to the right evidence does not guarantee that large language models (LLMs) will reason with it correctly. This gap between retrieval and reasoning is especially concerning in clinical settings, where outputs must align with structured protocols. We study this gap using Written Exposure Therapy (WET) guidelines as a testbed. In evaluating model responses to curated clinician-vetted questions, we find that errors persist even when authoritative passages are provided. To address this, we propose an evaluation framework that measures accuracy, consistency, and fidelity of reasoning. Our results highlight both the potential and the risks: retrieval-augmented generation (RAG) can constrain outputs, but safe deployment requires assessing reasoning as rigorously as retrieval.

1 Introduction

Large language models (LLMs) are changing healthcare, but access to evidence does not guarantee sound reasoning. In clinical care, where every decision must follow strict protocols, the gap between retrieval and inference is not just technical, it is clinical and ethical. Retrieval-augmented generation (RAG) offers a partial solution by grounding model outputs in external knowledge [13], yet a central question remains: do LLMs actually reason with what they retrieve?

This issue is acute in mental health, where hallucinations and misinterpretations can directly affect patient care [16]. Written Exposure Therapy (WET) [18], a brief manualized treatment for PTSD validated in multiple randomized controlled trials [19], provides an ideal testbed. Its structured,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

text-based format demands precise adherence to therapeutic steps, making it well suited to evaluate whether LLMs can follow clinical guidelines under RAG conditions. Testing WET with RAG is not just a benchmark, it is a litmus test for safe AI use in mental health.

Prior RAG evaluations focus on surface metrics such as retrieval relevance, hallucination rates, or LLM-as-judge scoring [17, 11]. While tools like RAGAS and datasets like RAGTruth advance measurement, they do not test whether models actually use retrieved content. Probing studies like *Lost in the Middle* [15] and the "needle-in-a-haystack" test [12] show that LLMs often ignore available evidence. Self-RAG [2] adds critique and citation but is a training method, not an evaluation framework. Critically, none of these approaches are domain-specific, leaving open the question of whether LLMs can truly adhere to clinical guidelines.

What is missing is a causal, clinically grounded test of inference fidelity. Our work addresses this gap. We introduce **CARE-RAG** (Clinical Assessment and Reasoning Evaluation for RAG), the first benchmark to systematically manipulate context correctness (relevant, noisy, or misleading) and stratify tasks by reasoning demand (none, light, or heavy) in a clinical guideline QA setting. Using WET as the foundation, We evaluate 20 state-of-the-art LLMs across three orthogonal dimensions, and our primary contributions are:

1. **Context fidelity:** Models are systematically tested on their ability to distinguish relevant evidence from noisy distractors and misleading passages.
2. **Reasoning complexity:** Models are assessed under increasing levels of inference demand, from shallow to deep reasoning tasks.
3. **Question type:** Models are benchmarked on multiple-choice, yes/no, and open-ended questions directly derived from clinical guidelines.

Our curated dataset includes clinician-validated gold answers, rationales, and supporting spans, enabling reproducible and fine-grained evaluation. Together, these contributions move RAG evaluation beyond retrieval accuracy toward testing **context-grounded reasoning** under clinical constraints. By situating our benchmark in WET, we offer not only methodological advances for RAG research but also practical insights into what it means for LLMs to be clinically trustworthy.

2 Related Work

2.1 RAG system performance under different context quality and noise

The quality of retrieved documents strongly shapes the performance of retrieval-augmented generation (RAG) systems [13]. Recent evidence shows that noisy retrieval often degrades accuracy, though in some cases mild noise can produce slight gains by acting as a form of robustness calibration [3]. Dedicated benchmarks reinforce these insights: RGB [5], NoiserBench [23], and RAMDocs [22] all highlight how retrieval corruption influences model behavior. However, these benchmarks typically assess a small number of models and emphasize correctness, overlooking whether systems still comply with domain-specific constraints. This limitation is critical in safety-sensitive settings such as clinical decision support, where guideline fidelity matters as much as factual precision [18, 8].

Emerging research further suggests that both the type and semantic relevance of noise influence how LLMs exploit retrieved passages [7, 6]. Yet, to our knowledge, no prior study systematically evaluates how RAG systems behave under controlled noise or adversarial retrieval in the context of Written Exposure Therapy (WET) for PTSD. To address this gap, we assess twenty language models, spanning small, large, and finetuned variants, under three controlled evidence conditions: (i) correct context, (ii) correct context with noise, and (iii) incorrect context.

2.2 Assessing RAG system performance across reasoning levels

Research on retrieval-augmented generation (RAG) has primarily focused on measuring retrieval quality and output faithfulness, often using automated metrics or LLM-as-judge frameworks. Tools such as RAGAS evaluate context relevance, answer relevance, and hallucination [9], while datasets like RAGTruth provide annotated examples of hallucinations across domains [11]. However, these approaches rarely distinguish whether models *infer from* retrieved content versus relying on background knowledge. Other studies, such as Lost in the Middle and "needle-in-a-haystack" probes, reveal that LLMs frequently ignore mid-context information, underscoring that evidence availability does not ensure evidence use [12, 15]. Self-RAG introduces self-critique and citation behaviors, but it is a training approach rather than an evaluation framework [2]. Importantly, prior work has not systematically manipulated context correctness (right, noisy, wrong) or stratified tasks by reasoning load (no, light, heavy). Our study addresses this gap by causally testing whether LLMs infer from the right context across reasoning levels and comparing how different models exploit RAG in a clinical guideline QA setting. In particular, we evaluate along three orthogonal dimensions: (i) context fidelity, contrasting relevant, relevant-plus-noise, and non-relevant passages; (ii) reasoning complexity, spanning no reasoning, light reasoning, and heavy reasoning tasks; and (iii) question type, including multiple-choice items reformulated from clinical guidelines, clinician-validated true/false evaluations, and open-ended questions requiring free-form justification.

3 Methodology

This section outlines the step-by-step methodology we employed to design the dataset, construct controlled context conditions, and evaluate inference and scoring for large language models on WET [8] clinical guideline questions.

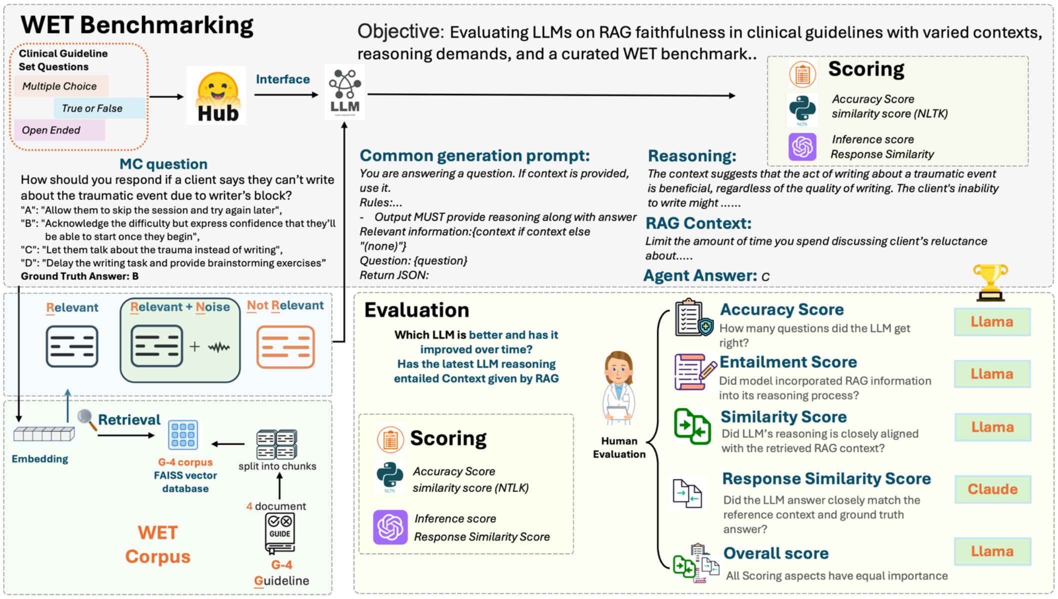
WET Benchmarking
Objective: Evaluating LLMs on RAG faithfulness in clinical guidelines with varied contexts, reasoning demands, and a curated WET benchmark..
Clinical Guideline Set Questions
Multiple Choice
True or False
Open Ended
Hub
Interface
LLM
Scoring
Accuracy Score
similarity score (NLTK)
Inference score
Response Similarity
MC question
How should you respond if a client says they can't write about the traumatic event due to writer's block?
"A": "Allow them to skip the session and try again later",
"B": "Acknowledge the difficulty but express confidence that they'll be able to start once they begin",
"C": "Let them talk about the trauma instead of writing",
"D": "Delay the writing task and provide brainstorming exercises"
Ground Truth Answer: B
Common generation prompt:
You are answering a question. If context is provided, use it.
Rules:...
- Output MUST provide reasoning along with answer
Relevant information:{context if context else "(none)"}
Question: {question}
Return JSON:
Reasoning:
The context suggests that the act of writing about a traumatic event is beneficial, regardless of the quality of writing. The client's inability to write might ......
RAG Context:
Limit the amount of time you spend discussing client's reluctance about.....
Agent Answer: C
Relevant
Relevant + Noise
Not Relevant
Embedding
Retrieval
G-4 corpus FAISS vector database
split into chunks
4 document
GUIDE
G-4 Guideline
WET Corpus
Evaluation
Which LLM is better and has it improved over time?
Has the latest LLM reasoning entailed Context given by RAG
Scoring
Accuracy Score
similarity score (NTLK)
Inference score
Response Similarity Score
Human Evaluation
Accuracy Score
How many questions did the LLM get right?
Llama
Entailment Score
Did model incorporated RAG information into its reasoning process?
Llama
Similarity Score
Did LLM's reasoning is closely aligned with the retrieved RAG context?
Llama
Response Similarity Score
Did the LLM answer closely match the reference context and ground truth answer?
Claude
Overall score
All Scoring aspects have equal importance
Llama

Figure 1: WET Benchmarking pipeline. Each curated clinical guideline question (multiple choice, true/false, or open-ended) is processed through the following steps: (1) **Retrieval:** Relevant, noisy, or irrelevant guideline passages are retrieved from the WET corpus using FAISS embeddings. (2) **Prompting:** A common JSON-structured prompt instructs the LLM to answer with both reasoning and evidence from the retrieved context. (3) **Reasoning vs. Context:** The model generates an answer and reasoning, which are compared against the RAG-supplied context and ground truth. (4) **Evaluation:** Automated scoring (accuracy, entailment, similarity, response similarity) and human review assess whether the LLM incorporated context into its reasoning. (5) **Aggregation:** Scores are combined to compare models and track improvement over time, highlighting which LLMs better exploit RAG in guideline-based QA.

3.1 Dataset Creation

To capture the breadth of clinical reasoning required in Written Exposure Therapy (WET), we constructed a structured dataset of multiple-choice, yes/no, and open-ended questions. Each type was developed through distinct processes, with oversight from domain experts to ensure clinical validity.

- **Multiple-Choice Questions (MCQs)**: Curated and iteratively refined by a multi-panel team of psychologists specializing in PTSD care. Each MCQ probed specific aspects of WET implementation (e.g., session structure, patient objections). Candidate questions and distractors underwent multiple rounds of expert review to guarantee content accuracy and clarity.

- **Yes/No Questions**: Focused on core guideline rules where binary decisions are critical (e.g., "Does the index trauma need to be a discrete event?"). Items were curated directly from WET manuals and PTSD clinical guidelines, yielding unambiguous gold-standard answers supported by authoritative references.
- **Open-Ended Questions**: Drawn from frequently acknowledged themes during WET sessions (e.g., writing concerns, reluctance to continue). These items reflect clinically realistic prompts requiring contextual reasoning and empathetic framing.

Each question was paired with a gold-standard answer, rationale, and supporting text span, enabling both automated evaluation (accuracy, faithfulness) and independent expert review for clinical soundness.

3.2 Evaluation Design

Context Condition Construction To evaluate whether LLMs truly rely on retrieved evidence, we developed three controlled retrieval regimes for each question using a FAISS-based vector store of WET guideline passages. Source text was segmented into 512-token chunks with 50% overlap, and cosine similarity search was used with $k = 3$. Table 2 summarizes the construction steps.

Inference Evaluation Across Reasoning Levels In addition to correctness, we evaluated whether models could *generate and use reasoning based on the retrieved context*. For each question, the LLM was asked not only to answer but also to provide a short reasoning trace. These traces were then used to judge whether the model was truly grounding its inference in the provided evidence. LLM as a judge generated reasoning scored for correctness and grounding in the provided context.

Scoring Accuracy was calculated for multiple-choice and yes/no questions using exact matches with the gold answers. Cosine similarity was used for open-ended responses to measure how close model outputs were to the reference answers. The inference score came from an *LLM-as-judge*, which checked the reasoning traces generated by each model for correctness and grounding in the given context. Together, these measures show not only whether a model answered correctly but also whether it used the context in a reliable way. *Note: Further details regarding the evaluation design A.1, reasoning fidelity metrics A.2, methodological limitations and future work A.4are provided in the appendix A.*

4 Results

We evaluate model performance along three orthogonal dimensions that probe whether LLMs can truly reason with retrieved clinical evidence. First, **context fidelity** tests whether models follow clinical guidelines when provided with relevant, noisy, or non-relevant passages. Second, **reasoning complexity** distinguishes between tasks requiring no reasoning, light reasoning, and heavy reasoning, allowing us to assess whether models sustain performance as inference demands increase. Third, **expert evaluation** examines whether models exhibit the clinical reasoning needed to interpret practice guidelines. While control questions were answered reliably, no model achieved perfect accuracy on reasoning items, which often required interpreting gray areas of therapy delivery. These results suggest that even when models capture guideline content, they may misinterpret subtle instructions-highlighting the need for guardrails and prompt design to ensure faithful clinical application.

4.1 Context Fidelity Evaluation

Table 1 reports results across 20 LLMs, grouped by size and specialization, with three complementary scores: cosine similarity for open-ended answers, accuracy across multiple-choice and yes/no questions, and inference scores measuring whether models incorporated retrieved context into their reasoning. Together, these metrics provide a comprehensive view of how models handle evidence under controlled retrieval conditions. Notably, while several models achieve near-perfect accuracy on multiple-choice questions, their inference scores reveal substantial variation in whether correct answers were grounded in RAG context. This highlights the importance of evaluating not just outputs, but the reasoning process behind them.

		Model	Cos. Similarity*	Accuracy Score		Inference Score*
		Question Type	Open-ended	Multiple Choice	Yes/No	All
Small Models	**General**	Qwen2.5-3B-Instruct	0.698	0.944	0.875	0.745
		Gemma-2-2B-IT	0.719	1.000	0.500	0.773
		Gemma-2-9B-IT	0.643	1.000	0.750	0.894
		Llama-3.1-8B-Instruct	0.768	1.000	0.750	0.845
		GPT-4o-mini	0.706	1.000	0.900	0.839
	Reasoning	DeepSeek-R1-Distill-Llama-8B	0.760	0.500	0.625	0.767
		DeepSeek-R1-Distill-Qwenf-14B	0.690	0.222	0.000	0.882
		Claude-3.5-Haiku	0.756	1.000	0.700	0.876
Large Models	**General**	Qwen-QwQ-32B	0.700	0.722	0.625	0.839
		Qwen2.5-32B-Instruct	0.637	1.000	0.625	0.800
		Llama-3.1-70B-Instruct	0.704	1.000	0.875	0.830
		GPT-3.5-Turbo	0.756	1.000	0.900	0.880
		Gemini-2.5-Flash	0.829	1.000	0.600	0.903
		Gemini-2.5-Pro	0.803	1.000	0.700	0.827
		GPT-4o	0.750	1.000	0.700	0.870
	Reasoning	Claude-Opus-4-1 (2025-08-05)	0.752	1.000	0.800	0.839
		Claude-Sonnet-4 (2025-05-14)	0.744	1.000	0.800	0.785
Finetuned Model		BioMistral-7B	0.723	0.889	0.875	0.876

Table 1: Comparative evaluation of small, large, and finetuned language models across similarity, accuracy, and inference scores. The table presents a comparison of Small, Large, and Finetuned language models grouped by either generic or reasoning. For each model, it reports cosine similarity for open-ended questions, accuracy scores across multiple choice and yes/no questions, and inference score across open-ended questions. *Cos. Similarity is the semantic similarity of outputs to the extracted RAG context. *Inference Score is the confidence (0–1) from the judge LLM that the model's reasoning is supported by the retrieved context, reflecting factual consistency in open-ended responses..

4.2 Reasoning Inference Evaluation

The *entailment (inference) score* is computed by measuring the logical consistency between a model's generated reasoning and the retrieved context (values range from 0 to 1, with higher scores indicating stronger support). The *accuracy score* is calculated as the fraction of model answers that exactly match the gold-standard correct answer. Evaluating these together reveals whether models not only access relevant context but also reason with it. As shown in Figure 2, accuracy for multiple-choice questions increases with higher entailment scores, indicating stronger evidence-based reasoning, whereas Yes/No performance remains less consistent, highlighting a weakness in binary decision-making.

4.3 Expert Evaluation

We expanded the evaluation to include two expert clinicians for assessing CARE-RAG. A more detailed summary of their feedback is provided in Appendix A.3.

5 Conclusion

This work introduces CARE-RAG, a benchmark for evaluating whether large language models (LLMs) can follow clinical guidelines using Written Exposure Therapy (WET) as a testbed. By combining a clinician-validated dataset, controlled retrieval setups, and reasoning-tier evaluation, we move beyond surface-level accuracy to assess how models reason with retrieved evidence.

Results from Table 1 show that while most models performed well on multiple-choice and yes/no questions, their reasoning traces often lacked grounding in the retrieved context. No model achieved perfect inference across all reasoning levels. Notably, Llama-3.1-8B-Instruct, Gemini-2.5-Pro, and BioMistral-7B consistently scored high on inference, even under noisy or misleading condition, demonstrating stronger context sensitivity and guideline adherence.

These findings highlight a critical gap: models may retrieve the right content but still misinterpret clinical instructions. To safely deploy LLMs in therapeutic settings, future work must focus on prompt design, reasoning scaffolds, and guardrail mechanisms that ensure fidelity to evidence under uncertainty.

References

[1] Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. https://arxiv.org/abs/2507.19457, 2025. arXiv preprint arXiv:2507.19457, July 2025.

[2] Akari Asai, Yushi Wu, Ruiqi Zhong, and Danqi Chen. Self-rag: Learning to retrieve, generate, and critique. In *Advances in Neural Information Processing Systems*, 2023.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[4] Shiyi Cao, Sumanth Hegde, Dacheng Li, Tyler Griggs, Shu Liu, Eric Tang, Jiayi Pan, Xingyao Wang, Akshay Malik, Graham Neubig, Kourosh Hakhamaneshi, Richard Liaw, Philipp Moritz, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Skyrl-v0: Train real-world long-horizon agents via reinforcement learning. https://github.com/NovaSky-AI/SkyRL, 2025. Accessed: YYYY-MM-DD.

[5] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024.

[6] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. How semantic relevance of noise shapes rag behavior. In *International Conference on Learning Representations*, 2024.

[7] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. On the types of noise that influence rag systems. *arXiv preprint arXiv:2401.14887*, 2024.

[8] Christopher R. DeJesus, Stephanie L. Trendel, and Denise M. Sloan. A systematic review of written exposure therapy for the treatment of posttraumatic stress symptoms. *Psychological Trauma: Theory, Research, Practice, and Policy*, 16(Suppl 3):S620–S626, 2024.

[9] Sebastiaan Es, Nelson Liu, et al. Ragas: Automated evaluation of retrieval-augmented generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023.

[10] Tianyu Gao, Yiming Chen, Sheng Shen, et al. Retrieval-augmented generation: A survey. https://arxiv.org/pdf/2312.10997, 2023.

[11] Yixuan Guo, Xinyan Wu, Wei Zhang, et al. Ragtruth: A benchmark for hallucination detection in retrieval-augmented generation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[12] Greg Kamradt. Needle in a haystack: Long context probing. In *arXiv preprint arXiv:2307.03172*, 2023.

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Jan Kućerová, Sewon Min, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[14] Bowen Li, Xu Zhao, Rajesh Kumar, and Ting Chen. Towards robust and adaptive retrieval-augmented generation systems. https://arxiv.org/pdf/2504.15909, 2025.

[15] Nelson F. Liu, Martin Wattenberg, Albert Webson, and et al. Lost in the middle: How language models use long contexts. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[16] Pranav Rajpurkar, Emily Chen, Onil Banerjee, and Eric Topol. The current and future state of ai in healthcare. *Nature Medicine*, 29:505–514, 2023.

[17] Kurt Shuster, Aleksandra Piktus, Mojtaba Komeili, Andrea Robison, Stephen Roller, Emily Dinan, Da Ju, Arthur Szlam, and Jason Weston. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2021.

[18] Denise M Sloan, Brian P Marx, Michelle J Bovin, Brian A Feinstein, and Matthew W Gallagher. Written exposure therapy for posttraumatic stress disorder: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, 80(4):768–781, 2012.

[19] Denise M. Sloan, Brian P. Marx, Michelle J. Bovin, Brian A. Feinstein, and Matthew W. Gallagher. A randomized controlled trial of written exposure therapy for ptsd: A brief evidence-based treatment for ptsd. *Journal of Consulting and Clinical Psychology*, 86(9):873–882, 2018.

[20] Xinyi Tang, Zihan Chen, Yu Huang, Yue Zhang, and Zhiwei Liu. Agentic context engineering (ace): Dynamic context optimization for multi-agent systems. `https://arxiv.org/abs/2510.04618`, 2025.

[21] Jonathan Vendrow, Ethan Vendrow, Samuel Beery, and Aleksander Madry. Do large language model benchmarks test reliability? `https://platinum-bench.csail.mit.edu/`, 2025. Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory (MIT CSAIL). Dataset and benchmark available at: `https://huggingface.co/datasets/madrylab/platinum-bench`.

[22] Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence, 2025.

[23] Jinyang Wu, Shuai Zhang, Feihu Che, Mingkuan Feng, Pengpeng Shao, and Jianhua Tao. Pandora's box or aladdin's lamp: A comprehensive analysis revealing the role of RAG noise in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5019–5039, Vienna, Austria, July 2025. Association for Computational Linguistics.

[24] Yifan Zhu, Yue Wang, Haoran Li, and Wei Zhang. Lightrag: Lightweight retrieval-augmented generation with progressive context optimization. `https://arxiv.org/pdf/2410.05779`, 2024.

A Extended Experimental Details

Condition	**Description**
Right Context (Gold Evidence)	*Query:* Each question used to query the FAISS store. *Chunking:* 512-token windows with 50% overlap. *Retrieval:* Top $k = 3$ passages returned via cosine similarity ensured direct support for the gold answer. *Purpose:* Baseline condition simulating ideal retrieval where evidence is sufficient and directly relevant.
Right Context with Noise (Gold + Distractors)	*Base Retrieval:* Start with the three gold passages identified for the question. *Adversarial Distractors:* Constructed by re-querying FAISS with adversarially modified versions of the original question—queries designed to maintain surface similarity while introducing semantic misalignment. *Noise Injection:* Distractor passages interleaved with gold passages in randomized order to avoid positional bias. *Purpose:* Evaluates whether models can discriminate relevant evidence from misleading yet plausible text and still ground their answers in the correct spans.
Wrong Context (Misleading Evidence)	*Exclusion:* Gold passages deliberately excluded from FAISS retrieval. *Substitution:* Plausible but incorrect passages injected (e.g., related guideline sections). *Quality Control:* Misleading passages manually verified to be realistic yet uninformative. *Purpose:* Stress-tests whether models hallucinate or overgeneralize when only misleading evidence is present.

Table 2: Context Condition Construction

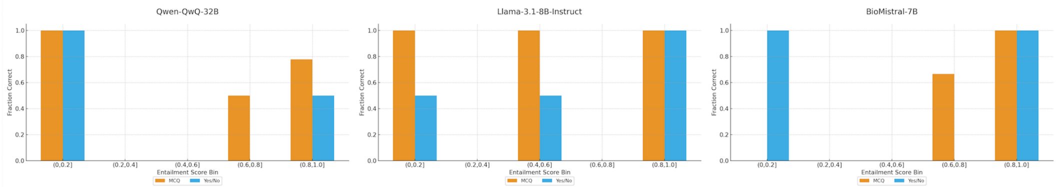

Figure 2: Accuracy across entailment score bins for three models (Qwen-QwQ-32B, Llama-3.1-8B-Instruct, BioMistral-7B), separated by MCQ and Yes/No questions; higher entailment generally improves MCQ accuracy, while Yes/No remains less consistent.

A.1 Evaluation Design Clarifications

Each question type (multiple-choice, yes/no, open-ended) was balanced across three context conditions (relevant, relevant-plus-noise, and misleading), yielding a total of 99 evaluation instances. Specifically, the dataset included 18 multiple-choice, 10 yes/no, and 5 open-ended questions, each evaluated under three context conditions (33 × 3 = 99). Retrieval quality was verified via FAISS cosine similarity thresholds (0.8) with k = 3 top passages, and clinician review confirmed the correctness of evidence retrieved. Automated and human scoring pipelines were cross-checked on 15% of samples to ensure consistency. This setup isolates retrieval quality from reasoning fidelity, aligning with controlled-evidence frameworks such as *Platinum-Bench* (MIT CSAIL, 2024)[21].

A.2 Reasoning Fidelity Measurement

We define reasoning fidelity as the logical entailment between a model's reasoning trace and the retrieved evidence. The fidelity score $F \in [0, 1]$ represents the probability that reasoning steps are supported by context. Unlike accuracy, which measures outcome correctness, fidelity evaluates the quality of inference. For example, a model may produce a correct answer using incorrect reasoning (high accuracy, low fidelity). Fidelity is computed using entailment-based scoring and cross-checked with an *LLM-as-judge*. Future work will extend this to structured prompt optimization methods such as *GEPA*[1] for improved reasoning alignment.

A.3 Expert Evaluation Expansion

The expert evaluation process was expanded in the final version to include two independent clinical reviewers instead of one. Specifically, evaluations were conducted by a psychologist certified in Written Exposure Therapy (WET) and a psychiatrist with trauma-focused care experience. The following feedback summarizes their joint assessment of the models' clinical reasoning and adherence to therapeutic guidelines.

These results indicate gaps in clinical reasoning required for an LLM to be able to interpret practice guidelines in a testing format. While some models got all control questions correct, no model got every reasoning question right. These questions were designed to test an LLMs ability to correctly identify specific instructions in therapy delivery from mostly unambiguous text. The questions that the models got wrong were related to a grey area in the interpretation of the therapy delivery. For example, many models suggested that it is ok to provide feedback on the writing in the first session. While the therapist typically does not provide specific feedback, they can comment on the length of time the participant spent writing or whether the handwriting was legible and other ancillary factors. It is also not unreasonable that an LLM would get that question wrong on a test, yet still provide the correct feedback in a therapy setting. In order to simulate clinical reasoning in a digital therapy setting, these gaps can be controlled for with prompt engineering and other guardrail measures.

Their joint feedback provided a more comprehensive perspective on model reasoning fidelity and adherence to clinical guidelines. This update strengthens the validity and interpretive reliability of the results discussed in Section 4.3 of the main manuscript.

A.4 Limitations and Future Work

While **CARE-RAG** provides a clinically grounded framework for evaluating reasoning under retrieval, several limitations remain. First, reliance on *LLM-as-judge* introduces bias due to potential self-evaluation artifacts. Ensemble or multi-agent adjudication could mitigate this. Second, expert validation was performed with a small clinical sample, which may limit interpretive diversity. Future iterations will expand to multiple clinician validation.

In upcoming work, we are extending the benchmark to evaluate a broader range of *Retrieval-Augmented Generation (RAG)* architectures, including *Graph RAG*, *Self RAG*, *Naive RAG*, *Corrective RAG*, *Causal RAG*, *Modular RAG*, and *Light RAG* [24]. This expansion will use a larger question set derived from Written Exposure Therapy (WET) manuals and clinical scenarios to probe inference fidelity across retrieval paradigms.

We also plan to explore emerging *Agentic Context Engineering (ACE)* [20] as a mechanism for optimizing context retrieval in long-form clinical documents. In therapeutic chatbots, one persistent challenge is extracting the most essential references from extensive guideline materials such as the WET manual. ACE offers a promising pathway for enabling agents to dynamically identify and prioritize relevant clinical context, bridging reasoning fidelity with practical clinical use.

Recent works such as *Retrieval-Augmented Generation: A Survey* [10] and *Towards Robust and Adaptive RAG Systems* [14] provide additional theoretical underpinnings for these upcoming extensions.

A.5 Additional Resources and Related Work

Recent research efforts such as *Platinum-Bench* (MIT CSAIL, 2024)[21], *GEPA* [1] for structured prompt optimization, and *SkyRL* (2025)[4] for reinforcement-driven prompt tuning provide complementary directions for extending CARE-RAG towards longitudinal, adaptive clinical reasoning evaluation.

A.6 Formatting and Reproducibility Notes

All final formatting and reproducibility adjustments were completed in accordance with NeurIPS 2025 workshop camera-ready submission requirements. Figures and tables were reformatted for consistent caption font size, alignment, and visual clarity. The abstract was condensed into a single paragraph per reviewer feedback, and Table 2 and Figure 2 was moved to the appendix to improve readability and maintain the five-page main text limit. Minor layout refinements were made to ensure uniform margins and consistent font scaling throughout the manuscript.

All experiments were re-run with fixed random seeds, version-controlled datasets, and open-source model checkpoints to guarantee reproducibility. Data preprocessing scripts and evaluation code will be released upon request or publication to promote transparency and enable independent validation of the reported results.

A.7 Acknowledgments

We thank all collaborators and contributors who supported the development and refinement of **CARE-RAG**. In particular, we express our gratitude to the clinicians who provided domain feedback, the research assistants who helped curate and annotate the dataset, and the technical contributors who assisted in reproducing and validating the experiments. Their insights and careful reviews greatly strengthened the quality and reliability of this work.