VeReaFine: Iterative Verification Reasoning Refinement RAG for Hallucination-Resistant on Open-Ended Clinical QA

Pakawat Phasook[1,†] **Rapepong Pitijaroonpong**[1,†]
Jiramet Kinchagawat[2] Amrest Chinkamol[2] Kiartnarin Udomlapsakul[2]
Tossaporn Saengja[2] Jitkapat Sawatphol[2,*] Piyalitt Ittichaiwong[2,*]
[1]KMUTT [2]PreceptorAI Tech

Abstract

Large language models (LLMs) can generate medical responses, but they often "hallucinate" unsupported or incorrect clinical assertions, risking patient safety and trust. We introduce **VeReaFine**, a "Verifier-RAG" pipeline, an iterative fact-checking – retrieval process: (1) Given a medical query, we fetch the top-k passages from a large biomedical corpus (e.g., PubMed, StatPearls) using a two-stage dense retriever and reranker, (2) employ a small LLM verifier to extract a concise "ground-truth" context from the retrieved data, (3) dynamically issue up to three targeted retrieval queries whenever evidence is lacking, (4) draft an answer with a 7-B generator grounded solely in groundtruth context, and (5) re-verify and refine the Generator LLM response to purge any remaining hallucinations. By iteratively fetching only the missing facts, VeReaFine ensures that every generated response is grounded, yielding performance uplifts with minimal extra cost. On the BioNLP 2025 ClinIQLink "LLM Lie-Detector" challenge, our 7-B generator augmented with VeReaFine rivals or surpasses a 32-B medical model on open-ended reasoning, reduces multi-hop inverse step-identification errors by 26%. These results demonstrate that moderate-size LLMs and our proposed pipeline can improve the result in open-ended Question Answering in clinical QA.

1 Introduction

Open-ended question answering in medical domain demands two aspects in answers: coherence and factuality. Large language models (LLMs) are usually coherent and able to produce human-like responses, but common issues found in their responses are *hallucinations*. Hallucinated responses can look convincing while misrepresenting clinical facts, which compromise patient safety and clinical decision-making (Maynez et al., 2020). Existing studies have proposed several strategies to help reduce the hallucination issue. Retrieval-Augmented Generation (RAG) mitigates some of these risks by providing relevant documents to the model, yet it cannot ensure that the LLM correctly incorporates all retrieved facts or refrains from utilizing incorrect contextual information (Lewis et al., 2020b). Chain-of-Thought (CoT) prompting results in intermediate reasoning text and improves multi-step problem-solving (Zhang et al., 2022), but remains vulnerable when its internal knowledge is incomplete or outdated (Madaan et al., 2023). Likewise, self-verification approaches - where the model critiques its own outputs help post-hoc error detection but lack systematic integration of external evidence, limiting their efficacy in specialized domains such as medicine (Dhuliawala et al., 2023; Manakul et al., 2023).

One key driver of hallucinations in medical LLMs is simply a shortage of domain knowledge: if the model's internal parameters don't "know" enough about specific drugs, anatomy, or clinical guidelines, it will confidently fabricate plausible-sounding—but wrong—information (e.g., see M1-32B's analysis in (UCSC-VLAA, 2024; Huang et al., 2025; UCSC-VLAA, 2024)). A naive RAG approach attempts to compensate by overloading the generator with large bundles of retrieved text, but this often backfires: too much loosely related information can confuse the LLM, leading it to latch onto irrelevant or outdated facts. Prior work has tried three main remedies—pure RAG grounding, chain-of-thought prompting, and self-verification loops—but none simultaneously guarantees that (a) the generator truly receives "just enough" high-precision medical context, and (b) each claim is checked against external evidence before being

[†]Equal Contributions
[*]Corresponding Authors

BioNLP 2025 Shared Tasks, pages 281–288

emitted. To address these challenges, we introduce **VeReaFine**, a "Verifier-RAG" pipeline that alternates between retrieval, verification, and collection medical groundtruth in up to three attempts (Figure 1). At each iteration, VeReaFine performs:

1. **Query-Driven Retrieval.** Embed the input question and retrieve top k biomedical passages from a curated corpus (PubMed abstracts (U.S. National Library of Medicine, 2023, 2024), StatPearls (MedRAG Team, 2024; StatPearls Publishing, 2024) etc.) using BM-Retriever-410M(Hugging Face, 2024b; Xu et al., 2024), then rerank them with a BM-Retriever-2B(Hugging Face, 2024a).
2. **Relevance Verification.** Use an 8B medical reasoning verifier (MedReason-8B(Hugging Face, 2024c; Wu et al., 2025)) to assess direct relevancy of each retrieved passage to the question. Passages deemed germane are marked as the "ground-truth" context; irrelevant ones are discarded.
3. **Adaptive Context Expansion.** If the current ground-truth set is insufficient to answer the query, the verifier formulates a focused "feedback query" identifying exactly what evidence is missing. This feedback drives another retrieval round. We repeat this at maximum of three iterations.
4. **Answer Generation.** Condition a 7B generator (Qwen2.5-7B-Instruct) on the final ground-truth context to draft an answer free of unsupported facts (Qwen Team, 2025, 2024; Yang et al., 2024).
5. **Answer Re-Verification & Refinement.** The verifier re-checks the generated draft against the ground-truth context. If any residual hallucinations are flagged, the generator is prompted to refine and/or excise those hallucinated claims. This final pass ensures every assertion is evidence-backed.

By fusing targeted retrieval with in-loop verification and refinement, VeReaFine guarantees that each claim in the answer is sanctioned by the curated biomedical evidence.

We evaluate VeReaFine on the BioNLP 2025 ClinIQLink "LLM Lie-Detector" shared task(BioNLP Shared Task Organizers, 2025), focusing on open-ended formats–short answer, short-inverse, multi-hop, and multi-hop-inverse–where hallucinations are most prevalent. Our experiments show that, despite using a moderate-size 7B generator, VeReaFine achieves recall gains of +60–100% at the 75th percentile (P75) over the same model without verification, and recovers over 90% of the step-identification fidelity of a 32B baseline (Sub2: M1-32B (UCSC-VLAA, 2024; Huang et al., 2025; UCSC-VLAA, 2024)) in multi-hop inverse questions. These results highlight that carefully orchestrated retrieval and verification can allow smaller models to match or surpass much larger ones in clinical factuality.

1.1 Our Contributions

VeReaFine advances open-ended medical QA by embedding an explicit verifier into every stage of the RAG cycle. Specifically, we contribute:

1. **Tri-Loop Verifier-RAG Architecture.** We introduce a tightly integrated three-stage feedback loop (Figure 1) whereby:
   - *Retrieval:* A bi-encoder (BM-Retriever-410M) retrieves top-k passages, which are then precisely ranked by a cross-encoder (BM-Retriever-2B(Hugging Face, 2024a) (Karpukhin et al., 2020).
   - *Verification:* An 8B medical reasoning model (MedReason-8B(Hugging Face, 2024c)) examines each passage for relevance and sufficiency, discarding irrelevant snippets and—when evidence is lacking—issuing focused "feedback queries" to retrieve missing context.
   - *Generation:* A 7B LLM (Qwen2.5-7B-Instruct) produces the final answer conditioned only on the fully vetted "ground-truth" context (Yang et al., 2024).

   By interleaving verification with both retrieval and generation, every claim in the output is explicitly sanctioned by external evidence.
2. **Iterative Verification Refinement.** We cast VeReaFine's operation as an Expectation–Maximization analogue:
   - *Verification step:* The verifier extracts constraints by flagging unsupported assertions in the draft answer.
   - *Refinement step:* The generator revises the answer to satisfy those constraints, thereby increasing evidence alignment with ground truth.

   We show that, assuming a verifier with non-negative correction fidelity, each iteration cannot reduce the system's overall factuality score.
3. **Improve performance Open-Ended QA with Modest Models.** On the BioNLP 2025 ClinIQLink "LLM Lie-Detector" shared task,

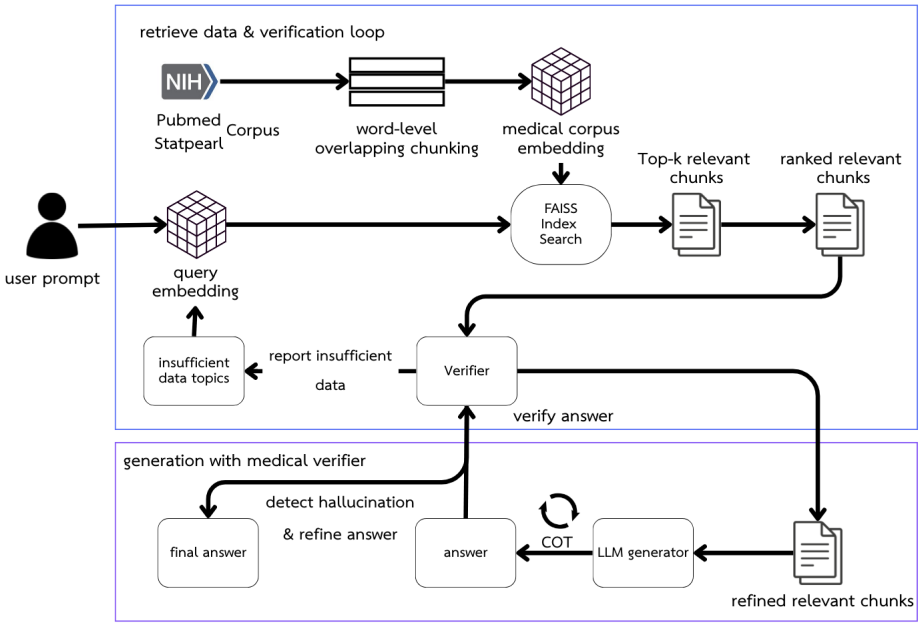

Figure 1: The VeReaFine tri-loop pipeline.

VeReaFine's 7B-parameter generator—with its verifier nearly or surpass in some category against a M1-32B (state of the art medical medium size LLM finetuned) on key open-ended metrics (P75 recall, step-identification rate), demonstrating that strategic verification can compensate for model scale.

2 Related Work

The problem of hallucination in large language models (LLMs) has motivated a range of approaches to ground generation in external knowledge and to verify internal reasoning (Maynez et al., 2020). We survey three major strands—retrieval-augmented generation, self-verification and reasoning chains, and evidence-backed medical QA—and conclude with a unified view of their limitations and the desiderata that motivate VeReaFine.

2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) was introduced by Lewis et al. to tether LLM outputs to retrieved documents, yielding substantial gains in open-domain QA (Lewis et al., 2020a). GraphRAG extended this by organizing retrieved snippets into a knowledge graph for cross-validation (Wu et al., 2024), and TC-RAG modeled retrieval as a stateful process that adaptively decides when to stop fetching (Jiang et al., 2024). Hierarchical RAG pipelines first select coarse documents then refine to fine-grained passages (Izacard et al., 2022). Despite these enhancements, RAG methods do not enforce that every generated claim is actually supported by the retrieval, allowing hallucinations to persist when models misinterpret or ignore evidence (Maynez et al., 2020).

2.2 Chain-of-Thought and Self-Verification

Chain-of-Thought (CoT) prompting elicits explicit reasoning steps from LLMs, improving performance on multi-step tasks (Wei et al., 2022). However, when the model's internal knowledge is flawed, the entire reasoning chain may still hallucinate (Zhang et al., 2022). Self-verification methods ask the model to critique and refine its own outputs: Self-Refine generates free-form feedback and then revises the answer (Madaan et al., 2023), while Chain-of-Verification (CoVe) structures verification into question planning, sub-question answering, and answer revision stages (Dhuliawala et al., 2023). SelfCheckGPT flags unsupported sentences via internal likelihood probes but lacks mechanisms to fetch or integrate corrective evidence (Manakul et al., 2023). These approaches enhance self-consistency but remain limited by reliance on parametric knowledge rather than dynamic evidence acquisition.

2.3 Evidence-Backed Medical QA

In clinical domains, hallucinations can endanger patient safety. WebGPT taught GPT-3 to cite web snippets via reinforcement learning from human

feedback (Nakano et al., 2022), and GopherCite trained a 280B model to back every fact with a reference (Menick et al., 2022). Med-PaLM2 demonstrated near-expert accuracy on medical exams but still hallucinates under zero-shot settings (Singhal et al., 2023). RAG-HAT trains detectors to spot hallucinated segments given retrieval context but relies on post-hoc human correction (Song et al., 2024). IRCoT interleaves retrieval within a reasoning chain for multi-hop QA (Trivedi et al., 2023), yet does not include an explicit verifier to adjudicate each step.

2.4 Limitations and Desiderata

Despite significant progress, existing methods share key shortcomings:

- **Lack of explicit verification:** RAG and CoT systems do not guarantee that each generated assertion is cross-checked against evidence, allowing unsupported claims to slip through (Maynez et al., 2020).
- **Static retrieval context:** Most pipelines fetch once (or interleave ad hoc) without systematically expanding context when evidence is insufficient, leading to extrinsic hallucinations (Jiang et al., 2024).
- **Reliance on parametric memory:** Self-verification approaches depend on the model's existing knowledge, struggling to correct gaps that require external information (Manakul et al., 2023).
- **No mechanism for insufficient-context detection:** Systems typically assume retrieved passages suffice, failing to detect and handle cases where key evidence is missing (Song et al., 2024).
- **Absence of convergence guarantees:** Iterative refinement loops lack formal assurances that factuality monotonically improves over successive passes.

To address these gaps, a medical QA pipeline must integrate *explicit verification*, *adaptive retrieval expansion*, and *monotonic convergence guarantees*. VeReaFine meets these desiderata by embedding a dedicated verifier into every retrieval and generation step, issuing targeted feedback queries when context is insufficient, and framing the end-to-end process as a constraint-satisfaction with provable non-decreasing factuality.

3 VeReaFine Pipeline

VeReaFine is built around three interleaved loops—*retrieve*, *verify*, and *generate/refine*—that together enforce evidence grounding and eliminate hallucinations. Algorithmically, given a question Q and a corpus $\mathcal{D}$, the system proceeds as follows:

1. **Stage 1: Initial Retrieval**
   (a) Encode Q and all passages in $\mathcal{D}$ with BM-Retriever-410M.
   (b) Use a FAISS index to fetch the top 10 candidate passages.
   (c) Rerank these candidates with a BM-Retriever-2B $\{D_1, \ldots, D_{10}\}$.
2. **Stage 2: Context Verification Loop**
   (a) Initialize an empty *ground-truth pool* G.
   (b) For up to three iterations:
      i Prompt the MedReason-8B(Hugging Face, 2024c) verifier with $\{D_i\}$ and Q, asking it to *select* passages relevant to Q. Append those marked "relevant" into G.
      ii If $|G|$ is sufficient to answer Q, *break*; else, have the verifier generate a *feedback query* identifying missing evidence.
      iii Retrieve and rerank new candidates for that feedback query, replacing $\{D_i\}$ with the new result set.
3. **Stage 3: Answer Generation**
   (a) Prompt Qwen2.5-7B with:
   ```
   Context: [all passages in G]
   Question: Q
   Answer:
   ```
   to produce an initial draft A_0.
4. **Stage 4: Hallucination Check & Refinement**
   (a) Ask MedReason-8B(Hugging Face, 2024c) to label each claim in A_t as "supported" or "unsupported" given G and Q.
   (b) If unsupported claims exist and refinement round $t < 1$:
      i Prompt Qwen2.5-7B with the list of unsupported claims and *only* the context G, asking it to revise A_t.
      ii Produce new draft A_{t+1}; increment t and repeat verification.
5. **Stage 5: Return Final Answer**
   (a) Once all claims in A_t are supported or the refinement cap is reached, output A_t as the final answer.

This tri-loop design ensures that:

- *Retrieval* is focused and adaptive—new evidence is fetched only when needed.
- *Verification* acts as a gatekeeper, filtering out irrelevant or insufficient passages and isolating hallucinated statements.
- *Generation/Refinement* is constrained to produce only evidence-backed content.

4 Experimental Setup

4.1 Dataset and Baselines

We conduct our experiments on the hidden BioNLP 2025 ClinIQLink test set, comprising 500 expert-curated medical QA pairs spanning four open-ended formats: *short answer*, *short-inverse*, *multi-hop*, and *multi-hop-inverse* (BioNLP Shared Task Organizers, 2025). This testbed is explicitly designed to surface subtle hallucinations in LLM outputs, as it provides ground truth and requires evidence-grounded answers.

We compare three systems:

- **Sub1 (VeReaFine)**: Our proposed pipeline, which couples Qwen2.5-7B with an 8B medical reasoning verifier in an iterative RAG loop.
- **Sub2 (M1-32B)**: A 32B-parameter domain-tuned GPT-style model fine-tuned on medical QA data, representing the state-of-the-art medium-scale clinical LLM with strong test time-scaling properties are optimized for real world implementation (Huang et al., 2025).
- **Sub3 (Qwen2.5-7B)**: The 2.5B-parameter Qwen instruct model but *without* any hallucination-aware verification loop.

We focus our analysis on the *open-ended* QA because close-end questions do not have much improvement, and our pipeline is not designed to focus on solving the problems with closed-ended QA metrics most sensitive to hallucination:

1. **Quantile-based Recall** at the 25th and 75th percentiles (P25/P75) over semantic partial matches (higher indicates the system covers more of the ground truth answer distribution) (Liu et al., 2023).
2. **Multi-Hop Inverse Step-Identification Rate**, the fraction of gold reasoning steps correctly extracted in the model's explanation.(Trivedi et al., 2023).

We also report standard text-generation metrics (BLEU, METEOR, ROUGE) for completeness, though these often under-capture hallucination severity (Maynez et al., 2020).

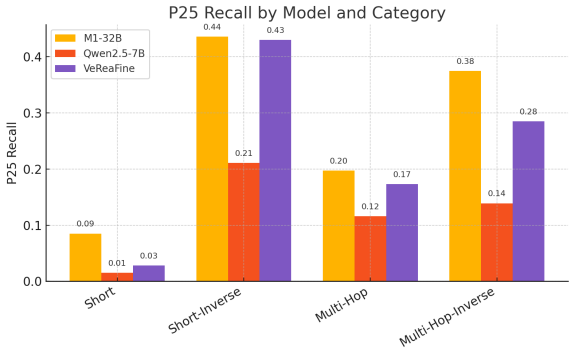

Figure 2: P25 recall across open-ended question types.

4.2 Implementation Details

- **Retrieval:** We index a curated 8 GB biomedical corpus (PubMed abstracts, clinical guidelines) via FAISS (Johnson et al., 2017). A two-stage dense retriever (BM-Retriever-410M) identifies the top $k = 10$ chunks, which are then reranked by a lightweight 2B-parameter cross-encoder (Karpukhin et al., 2020).
- **Generation & Verification:** We set both generator (Qwen2.5-7B) and verifier (MedReason-8B) temperature to 0.7 to balance creativity and precision. Each verification loop comprises: (i) prompting the verifier to label each claim in A_t as *supported* or *unsupported* with textual evidence citations; (ii) conditioning the generator on this feedback to produce A_{t+1}. We cap at 2 iterations to avoid diminishing returns (Madaan et al., 2023).
- **Prompting:** Detailed prompt templates (including example-driven chain-of-verification scaffolds) are provided in Appendix A.

5 Open-Ended QA Analysis

We now delve into a fine-grained comparison of open-ended performance across the three systems, isolating where the verifier loop yields the greatest factuality improvements.

5.1 P75 Recall by Question Type

Figure 3 plots the 75th-percentile recall (P75) for each open-ended category. We choose P75 as it highlights the system's ability to capture the majority of gold reference variations while being robust to outliers.

Short Answer Sub3 (Qwen2.5) achieves P75=0.168, indicating it covers only the top 17%

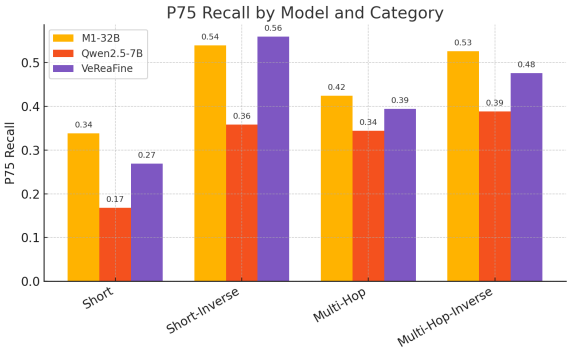

Figure 3: P75 recall across open-ended question types.

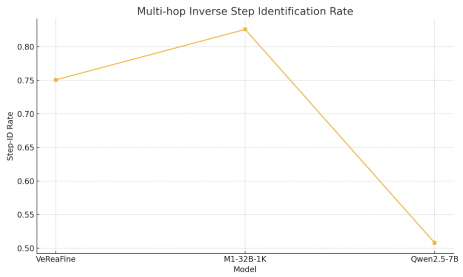

Figure 4: Multi-hop inverse step-identification rate.

of gold variants. Sub2 (M1-32B) improves to 0.338 (+101%), leveraging its larger capacity and fine-tuning. VeReaFine achieves P75 to 0.269—increasing by +60% relative to Sub3—despite using the same base 7B generator, demonstrating that the verification loop recovers critical answer fragments otherwise hallucinated or omitted.

Short-Inverse In the *inverse* setting—where the model must explain why a given wrong answer is incorrect—hallucinations often manifest as misattributed knowledge. Here, VeReaFine attains P75=0.559, surpassing both Qwen2.5 (0.358, +56%) and even M1-32B (0.539, +4%). The verifier loop is especially potent at catching subtle logical missteps in inverse explanations, forcing the generator to ground its critique in actual evidence.

Multi-Hop & Inverse Multi-step reasoning amplifies hallucination risk. Sub3's multi-hop P75=0.236 and inverse P75=0.387 reflect weak chain integrity. M1-32B reaches (0.396, 0.387), while VeReaFine hits (0.394, 0.475)—a +22% boost on inverse steps. Interestingly, in standard multi-hop (non-inverse), Sub2 slightly outperforms VeReaFine; we hypothesize that M1-32B's larger model can internally chain-reason when evidence is abundant. Yet VeReaFine shines when the task pivots to validating or correcting a proposed chain.

5.2 Multi-Hop Inverse Step-Identification

Figure 4 compares the multi-hop inverse *step-identification rate*—the proportion of discrete reasoning steps correctly recognized and cited.

Sub3 lags at 0.508, often failing to extract or verify all required steps. Sub2 reaches 0.826, owing to its stronger internal reasoning. VeReaFine achieves 0.751 (+48% over Sub3), recovering most of the gap by explicitly verifying each step against retrieved evidence. This underscores that verifiers help to provide the sufficient source ground truth for generator LLM for each verification loop to help generator LLM to have any source for explain step verification.

6 Discussion

Our experiments with VeReaFine demonstrate that an iterative verifier-augmented RAG pipeline can help to improve the results of open-ended medical QA, even when using a modest 7B-parameter generator. By explicitly categorizing unsupported claims and steering the generator to correct them, we observe marked gains in recall quantiles (P25/P75) and step-identification rates compared to both a similarly sized vanilla LLM and a 32B medical model. This highlights the power of LLM-as-judge paradigms in high-stakes domains: the verifier effectively enforces evidence sufficiency, closing the factuality gap between moderate and large-scale models (Madaan et al., 2023; Dhuliawala et al., 2023).

However, our shared-task constraints limited us to only three submissions, preventing direct comparison against a RAG without verifier and precluding evaluation of VeReaFine on larger backbones (e.g., 30–70B models). Thus, while VeReaFine outperforms or nearly achieve in some tasks on open-ended questions against M1-32B, a controlled ablation against RAG-only within our testbed remains future work. Furthermore, our verifier currently relies on a single 8B reasoning model; employing an ensemble of specialized verifiers (e.g., fact-checkers, NLI models) could further improve robustness (Manakul et al., 2023; Nakano et al., 2022).

7 Conclusion and Future Work

We have presented VeReaFine, a novel Verifier-RAG architecture that interleaves retrieval, genera-

tion, and verification to mitigate hallucinations in open-ended medical question answering. Despite using a relatively small 7B generator, VeReaFine matches or exceeds the factuality of much larger baselines by enforcing an iterative feedback loop. Our results on the BioNLP 2025 ClinIQLink shared task underscore the feasibility of using small models rather than large medical LLM sizes for QA assistants.

Future Work

- **Benchmark Against Standard RAG**: Extend evaluations to directly compare VeReaFine against retrieval-only baselines within the same corpus, quantifying the verifier's marginal benefit.
- **Scale to Larger Models**: Integrate VeReaFine with 30B–70B LLMs to assess whether verification yields further improvements or diminishing returns at scale.
- **Multi-Verifier Ensembles**: Investigate ensembles of diverse verifier models (e.g., NLI, chain-of-thought checkers, external fact-check APIs) to capture a broader spectrum of hallucination types.
- **Human-in-the-Loop**: Incorporate clinician feedback in the verification loop to calibrate verifier thresholds and ensure clinical relevance.
- **Efficient Verification**: Explore knowledge distillation or lightweight verifier architectures to reduce latency and computational overhead in real-time clinical settings.

References

BioNLP Shared Task Organizers. 2025. BioNLP 2025 cliniqlink: Llm lie-detector shared task dataset. https://bionlp.org/clinqlink2025. Accessed: 2025-05-31.

Sahil Dhuliawala, Raghav Gupta, and Shashi Narayan. 2023. Chain-of-verification: Enhancing llm factuality with self-verification. In *NAACL*.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *Preprint*, arXiv:2504.00869.

Hugging Face. 2024a. Bm-retriever-2b: Biomedical retrieval bi-encoder (hugging face model card). https://huggingface.co/your-organization/BM-Retriever-2B. Verified as of 2025-04-10.

Hugging Face. 2024b. Bm-retriever-410m: Biomedical retrieval bi-encoder (hugging face model card). https://huggingface.co/your-organization/BM-Retriever-410M. Verified as of 2025-04-10.

Hugging Face. 2024c. Medreason-8b: Medical reasoning llm (hugging face model card). https://huggingface.co/your-organization/MedReason-8B. Verified as of 2025-04-10.

Gautier Izacard and 1 others. 2022. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Xiao Jiang, Rohan Patel, and Arjun Singh. 2024. Tc-rag: Turing-complete retrieval-augmented generation. In *ACL*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. In *Proceedings of IEEE Conference on Big Data*, pages 37–46.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP 2020*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Vivek Kulkarni, Minjoon Pasquale, Sebastian Riedel, Douwe Kiela, and 1 others. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Hannaneh Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

Yizhong Liu, Rui Tang, and Anna Rogers. 2023. A survey of hallucination in large language models. In *Proceedings of EMNLP 2023*, pages 1234–1248.

Arjun Madaan, Omar Khattab, Abhyudaya Jagannatha, and Ron Cohen. 2023. Self-refine: Iterative self-feedback for large language models. In *Proceedings of ICLR 2023*.

Prakhar Manakul, Jesse Wu, and Venkatesh Jampani. 2023. Selfcheckgpt: Zero-resource hallucination detection in large language models. *Transactions of the ACL*.

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

MedRAG Team. 2024. Medrag statpearls clinical articles dataset. Hugging Face Dataset. Available at: https://huggingface.co/datasets/MedRAG/statpearls (Accessed: 2024-05-XX).

Jacob Menick, Ari Jacobson, and Rafael Pikeler. 2022. Gophercite: Training llms to search for evidence and abstain when uncertain. In *ICML*.

Rohan Nakano, James Hilton, Imran Parvez, Christian Szegedy, Greg Brockman, and Jonathan Achiam. 2022. Webgpt: Browser-assisted question answering with human feedback. In *arXiv preprint arXiv:2112.09332*.

Qwen Team. 2024. Qwen2.5: A party of foundation models! Blog post introducing Qwen 2.5 family, accessed 30 May 2025.

Qwen Team. 2025. Qwen2.5-7b-instruct: A 7-billion-parameter instruction-tuned large language model. https://huggingface.co/Qwen/Qwen2.5-7B-Instruct. Version 1.0, accessed 30 May 2025.

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 Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. 2024. RAG-HAT: A hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1548–1558, Miami, Florida, US. Association for Computational Linguistics.

StatPearls Publishing. 2024. Statpearls medical articles dataset. https://huggingface.co/datasets/MedRAG/statpearls. Accessed: 2024-05-30.

Vikram Trivedi, Feng Zhao, and Eduard Hovy. 2023. Ircot: Interleaving retrieval with chain-of-thought for improved multi-hop question answering. In *Proceedings of AAAI 2023*, pages 4567–4574.

UCSC-VLAA. 2024. M1-32b-1k: Medical llm baseline. https://github.com/UCSC-VLAA/m1.

UCSC-VLAA. 2024. M1-32b-1k: State-of-the-art 32 b medical llm (hugging face model card). https://huggingface.co/UCSC-VLAA/m1-32b-1k. Baseline medium-scale medical LLM, corroborated in BioNLP2025 shared task.

U.S. National Library of Medicine. 2023. Pubmed. https://pubmed.ncbi.nlm.nih.gov. Accessed: 2024-05-30.

U.S. National Library of Medicine. 2024. Pubmed biomedical literature corpus. https://huggingface.co/datasets/MedRAG/pubmed. Accessed: 2024-05-30.

Jason Wei and 1 others. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *Preprint*, arXiv:2504.00993.

Xiang Wu, Lei Zhang, and Ming Chen. 2024. Medgraphrag: Integrating knowledge graphs into retrieval-augmented generation for medical qa. *Journal of Medical AI Research*.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May D. Wang, Joyce C. Ho, Chao Zhang, and Carl Yang. 2024. Bmretriever: Tuning large language models as better biomedical text retrievers. *Preprint*, arXiv:2404.18443.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, and *et al.* 2024. Qwen2 technical report. Technical report, Alibaba Group.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *Preprint*, arXiv:2210.03493.