Biomedical Evidence Retrieval with Agentic RAG and Dual Text Encoders

Dhruv Goyal[1,2,*,†], Ema Seibert[3,2,†], Ryan Ding[2,4,†], Matteo Migliarini[2,5] and Kevin Zhu[2]

[1]Indian Institute of Technology Bombay
[2]Algoverse AI Research
[3]University of California, Irvine
[4]University of California, San Diego
[5]Sapienza, University of Rome

Abstract

We propose an agentic RAG framework for biomedical evidence retrieval that uses iterative query refinement across PubMed and MIMIC-IV clinical notes. Using dual domain-specific encoders and self-critique loops, our system achieves competitive results on PMC-Patients and PubMedQA benchmarks, demonstrating the value of adaptive retrieval for clinical decision support.

Keywords

Agentic AI, Retrieval-Augmented Generation, Clinical AI, Healthcare

1. Introduction

Retrieval-Augmented Generation (RAG) has emerged as a leading approach for evidence-based retrieval, combining dense retrieval with generation [1]. In medicine, this paradigm was adapted using domain-specific models like BioBERT[2] to handle specialized terminology [3], yet traditional RAG pipelines are often static, retrieving once without adapting their reasoning. A more advanced paradigm, Agentic RAG, extends this by embedding autonomous decision-making and iterative reflection into the retrieval loop [4]. These systems use agentic control flows, such as corrective feedback or query routing, to achieve more adaptive and reliable reasoning [5, 6]. To address the need for structured evaluation in this area, this work benchmarks an agentic RAG framework on established biomedical QA datasets [7, 8, 9] and the Patients-PMC benchmark [10] to assess its generalization for clinical cohort discovery.

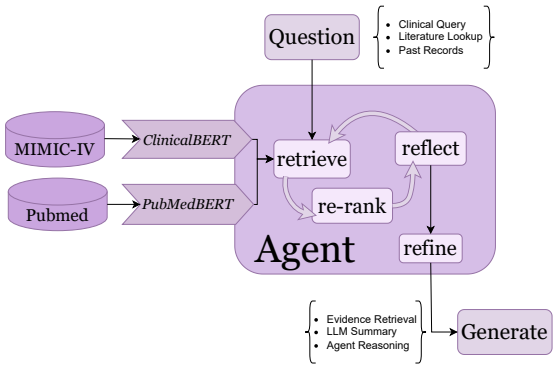

Figure 1: Hybrid biomedical RAG with iterative self-critique. Evidence from PubMed (literature) and MIMIC-IV (clinical notes) is retrieved via domain-specific encoders and re-ranked. An agent cycles between reflect and refine, yielding a final, evidence-grounded response.

NORA'25: 1st Workshop on Knowledge Graphs & Agentic Systems Interplay co-located with NeurIPS, Dec.1, 2025, Mexico City, Mexico

*Corresponding author.
†These authors contributed equally.
23b2122@iitb.ac.in (D. Goyal); seiberte@uci.edu (E. Seibert); dingryan2@gmail.com (R. Ding)
https://dhruv-portfolio-personal.netlify.app/ (D. Goyal); https://ryanding.com (R. Ding)

Table 1
Results for Patient-to-Article Retrieval (PAR) and Patient-to-Patient Retrieval (PPR) on the PMC-Patients dataset. Best results are in **bold**, second best are in *italics*.

	Patient-to-Article (PAR)				Patient-to-Patient (PPR)			
Method	**MRR@10**	**nDCG@10**	**P@10**	**R@1K**	**MRR@10**	**nDCG@10**	**P@10**	**R@1K**
Agentic (Ours)	**85.23**	**40.74**	*13.82*	*65.92*	*24.81*	**22.41**	*6.02*	*78.32*
Zhang et al. [12]	*64.44*	*28.62*	**22.12**	**69.09**	**25.35**	*22.39*	**6.65**	**83.78**
BM25	48.22	15.28	9.97	30.64	22.86	18.29	4.67	69.66
Contriever	15.03	4.62	3.41	16.74	10.50	8.01	2.24	52.64
SentBERT	10.58	3.53	2.71	13.52	5.28	3.88	1.17	37.55

2. Methodology

Our system employs an agentic RAG framework that iteratively refines queries and integrates evidence from biomedical literature (PubMed) and clinical notes (MIMIC-IV). The core is a dual-encoder retrieval pipeline orchestrated by an agentic control loop (Figure 1). We encode queries and documents using two specialized models: PubMedBERT for literature and ClinicalBERT for clinical notes, enabling parallel searches [3, 11]. Retrieved documents are then merged and refined using a cross-encoder reranker.

Instead of a single pass, an agentic loop assesses evidence quality. If deemed insufficient, the agent triggers a refinement action before re-querying, employing two main strategies: **Pseudo-Relevance Feedback (PRF)**, which refines the query embedding using top-ranked documents, and **Query Decomposition** for complex questions. The loop terminates upon result convergence or after a fixed number of iterations. Finally, a large language model (LLM) synthesizes the refined evidence into a concise, cited answer. Our full code is available at https://github.com/Dhruv-Git21/Agentic-Biomedical-Retrieval-System.

3. Results

We evaluate our agentic retrieval system on the *PMC-Patients* benchmark: covering Patient-to-Article Retrieval (PAR) and Patient-to-Patient Retrieval (PPR) [10]; and the reasoning-free setting of PubMedQA.

As shown in Table 1, our framework achieves competitive results across all tasks. On the PAR task, the system attains high performance. This strong result is expected, as PAR is a known-item retrieval task where high semantic overlap exists between the patient description and the target article. While the model also performs competitively on the more challenging PPR task, the PAR scores highlight the system's strength in precise evidence matching.

On PubMedQA, our framework attains an accuracy of 82.09%, outperforming key baselines such as BioBERT (80.80%). This demonstrates its effectiveness on standard biomedical question-answering benchmarks Table 2.

Table 2
Comparison of reasoning-free baselines on the PubMedQA dataset.

Model	**Acc**	**F1**
Agentic (Ours)	**82.09**	*62.81*
Shallow Features [7]	54.44	38.63
BiLSTM [7]	71.46	50.93
ESIM w/ BioELMo [7]	74.06	58.53
BioBERT [7]	*80.80*	**63.50**
PubMedBERT [13]	55.84	-
BioLinkBERT [14]	70.20	-
BioLinkBERT-large [14]	72.18	-
BioGPT [15]	78.20	-

4. Conclusion

In this work, we demonstrated the effectiveness of an agentic RAG framework for complex biomedical retrieval. Our system achieved competitive performance on the PMC-Patients and PubMedQA benchmarks, highlighting the advantages of agentic strategies over static pipelines. By enhancing retrieval precision and adaptability, these systems represent a promising path toward developing more reliable tools for evidence-based medicine.

5. Related Works

Biomedical RAG Foundational RAG models combined dense retrieval with generation for open-domain QA [1, 16]. In medicine, this paradigm was adapted using domain-specific pretrained models like BioBERT and PubMedBERT to handle specialized terminology [2, 3]. More recent systems like Med-PaLM 2 have integrated retrieval-based grounding with instruction tuning to achieve expert-level performance on medical benchmarks [17]. These works establish the value of domain-specific retrieval but often rely on single-pass, non-adaptive pipelines.

Agentic and Hybrid Retrieval To overcome the limitations of static retrieval, recent research has focused on more dynamic, agentic architectures. Methods like Corrective RAG (CRAG) introduce self-reflection, where the system assesses retrieval quality and triggers query reformulation if the evidence is weak [5]. Adaptive RAG classifies queries to follow different reasoning paths (e.g., simple vs. multi-hop) [6], while others integrate knowledge graphs to support complex, multi-step biomedical reasoning [18]. These approaches motivate our focus on evaluating the practical benefits of such agentic strategies. To improve retrieval quality, hybrid methods combining lexical (e.g., BM25) and semantic search are common [19, 20], and query expansion using medical ontologies like UMLS remains a critical step for bridging the vocabulary gap between user questions and scientific literature [21, 22].

Medical QA Benchmarks Our evaluation relies on established medical question-answering benchmarks that test a range of reasoning skills. These include PubMedQA (yes/no/maybe questions based on abstracts), BioASQ (idealized answer generation from multiple documents), and MedMCQA (large-scale multiple-choice questions from medical exams) [7, 8, 9]. These datasets provide a standardized foundation for assessing the performance of RAG systems.

Acknowledgments

Thanks to Algoverse Research for hosting and sponsoring this research. Also thanks to Kevin Han and Ben Liu for their support and feedback early in the development of this idea.

Declaration on Generative AI

During the preparation of this work, the authors used Claude and other Large Language Models in order to: rephrase paragraphs, shorten text passages, and assist in code writing. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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 (NeurIPS), 2020. URL: https://arxiv.org/abs/2005.11401.

[2] J. Lee, W. Yoon, S. Kim, et al., Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240. URL: https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

[3] Y. Gu, R. Tinn, H. Cheng, et al., Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (2021).

[4] A. Singh, A. Ehtesham, S. Kumar, T. Talaei Khoei, Agentic retrieval-augmented generation: A survey on agentic rag, arXiv preprint arXiv:2501.09136 (2025). URL: https://arxiv.org/abs/2501.09136. doi:10.48550/arXiv.2501.09136.

[5] S.-Q. Yan, J.-C. Gu, Y. Zhu, Z.-H. Ling, Corrective retrieval augmented generation, arXiv preprint arXiv:2401.15884 (2024). URL: https://arxiv.org/abs/2401.15884. doi:10.48550/arXiv.2401.15884.

[6] S. Jeong, J. Baek, S. Cho, S. J. Hwang, J. Park, Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 7036–7050. URL: https://aclanthology.org/2024.naacl-long.389/. doi:10.18653/v1/2024.naacl-long.389.

[7] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, X. Lu, Pubmedqa: A dataset for biomedical research question answering, in: Proceedings of EMNLP-IJCNLP 2019, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2567–2577. URL: https://aclanthology.org/D19-1259/.

[8] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artières, A.-C. Ngonga Ngomo, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, BMC Bioinformatics 16 (2015) 138. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0564-6. doi:10.1186/s12859-015-0564-6.

[9] A. Pal, L. K. Umapathi, M. Sankarasubbu, Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering, in: Proceedings of the Conference on Health, Inference, and Learning, volume 174, PMLR, 2022, pp. 248–260. URL: https://proceedings.mlr.press/v174/pal22a.html.

[10] Z. Zhao, Q. Jin, F. Chen, T. Peng, S. Yu, A large-scale dataset of patient summaries for retrieval-based clinical decision support systems., Scientific data 10 1 (2023) 909. URL: https://api.semanticscholar.org/CorpusID:266360591.

[11] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. B. A. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP@NAACL), 2019, pp. 72–78.

[12] Y. Zhang, H. Cheng, Z. Shen, X. Liu, Y.-Y. Wang, J. Gao, Pre-training multi-task contrastive learning models for scientific literature understanding, in: Findings of EMNLP'23, 2023, pp. 12259–12275.

[13] Y. Gu, R. Tinn, H. Cheng, et al., Domain-specific language model pretraining for biomedical nlp, arXiv preprint arXiv:2007.15779 (2020).

[14] M. Yasunaga, J. Leskovec, P. Liang, Biolinkbert: Pre-trained biomedical language model for biomedical text mining, in: Findings of ACL 2022, 2022.

[15] R. Luo, et al., Biogpt: Generative pre-trained transformer for biomedical text generation and mining, Briefings in Bioinformatics (2022).

[16] K. Guu, K. Lee, Z. Tung, P. Pasupat, M.-W. Chang, Realm: Retrieval-augmented language model pre-training, in: Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, 2020. URL: https://proceedings.mlr.press/v119/guu20a/guu20a.pdf.

[17] K. Singhal, T. Tu, J. Gottweis, et al., Toward expert-level medical question answering with large language models, Nature Medicine (2024). URL: https://www.nature.com/articles/s41591-024-03423-7. doi:10.1038/s41591-024-03423-7.

[18] N. Matsumoto, J. Moran, H. Choi, M. E. Hernandez, M. Venkatesan, P. Wang, J. H. Moore, Kragen: a knowledge graph-enhanced rag framework for biomedical problem solving using large language models, Bioinformatics 40 (2024) btae353. URL: https://academic.oup.com/bioinformatics/article/40/6/btae353/7687047. doi:10.1093/bioinformatics/btae353.

[19] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389.

[20] M. Luo, A. Mitra, T. Gokhale, C. Baral, Improving biomedical information retrieval with neural retrievers, in: AAAI, 2022. URL: https://arxiv.org/abs/2201.07745.

[21] A. R. Aronson, Effective mapping of biomedical text to the umls metathesaurus: The metamap program, in: Proceedings of the AMIA Symposium, 2001, pp. 17–21.

[22] F. Liu, I. Vulić, A. Korhonen, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Findings of ACL, 2021. URL: https://arxiv.org/abs/2010.11784.