Query Reformulation and Dense-Lexical Retrieval Fusion for Multi-Turn Retrieval-Augmented Generation

LLM-Driven Query Enrichment and Cross-Encoder Reranking for the MTRAG Benchmark

Sijan Shrestha

Howard University, sijan.shrestha@bison.howard.edu

Saurav Keshari Aryal

Howard University, saurav.aryal@howard.edu

While large language models increasingly serve as chat-based assistants, grounding their responses in retrieved evidence across multi-turn conversations remains a significant challenge, particularly when questions reference earlier turns, when the system must recognize unanswerable queries rather than hallucinate, and when relevant passages shift as the conversation evolves. We address these challenges on the MTRAG benchmark across four domain-specific corpora: ClapNQ (Wikipedia), Cloud (technical documentation), FiQA (financial), and Govt (government web pages). Our system employs a hybrid retrieve-then-rerank architecture. Queries are first augmented through LLM-driven query rewriting, breaking down entities and query itself, and generating hypothetical embeddings (HyDE) for semantic matching. Results from dense vector search and lexical matching are then fused via Reciprocal Rank Fusion and reranked though cross-encoder. Llama-3.3-70B-Instruct then generates responses based strictly on the most relevant text passages. The system achieves an nDCG@5 of 0.4098 on passage retrieval, a harmonic mean of 0.7462 on reference-grounded generation, and 0.5796 on end-to-end RAG.

CCS CONCEPTS • Information systems~Information retrieval~Retrieval models and ranking • Information systems~Information retrieval~Retrieval tasks and goals~Question answering • Information systems~Information retrieval~Information retrieval query processing~Query reformulation • Information systems~Information retrieval~Retrieval models and ranking~Rank aggregation • Information systems~Information retrieval~Retrieval models and ranking~Combination, fusion and federated search

Additional Keywords and Phrases: retrieval-augmented generation, multi-turn RAG, hypothetical document embeddings, cross-encoder reranking

REFERENCES

[1] Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems. arXiv preprint arXiv:2501.03468 (2025). https://arxiv.org/abs/2501.03468

[2] Gordon V. Cormack, Charles L. A. Clarke and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09). Association for Computing Machinery, New York, NY, USA, 758–759. https://doi.org/10.1145/1571941.1572114

[3] Meta AI. 2024. Llama 3.3. Retrieved from https://ai.meta.com/llama/

[4] Ho Bae. 2025. A Study on Enhancing Zero-Shot Dense Retrieval Using Query and Hypothetical Document Embedding Combination. *The Transactions of the Korea Information Processing Society* 14, 3 (2025), 161–171.

[5] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 1762–1777.