A Modular Architecture Proposal for Multi-Turn Conversational RAG Systems

**Guilherme C. Dutra[1], André Felipe dos S. Caraíba[1], João Pedro A. F. Matos[1]
Nádia F. F. da Silva[1], Deborah S. A. Fernandes[1], Sávio S. T. de Oliveira[1]**

[1]Instituto de informática – Universidade Federal de Goiás
Goiania, GO.

guilherme_dutra@discente.ufg.br, andrecaraiba@discente.ufg.br

joao.formiga@discente.ufg.br, nadia@inf.ufg.br

deborah.fernandes@ufg.br, savioteles@ufg.br

Abstract. *Conversational systems face growing challenges in understanding context, resolving references, and maintaining coherence across multiple user turns. The SemEval-2026 Task 8 [Katsis et al. 2026] challenges participants to build conversational Retrieval-Augmented Generation (RAG) systems capable of handling multi-turn interactions with context dependencies, coreferences, and diverse question types. We propose a modular architecture combining five complementary strategies: (1) CoT query rewriting with multi-query diversification; (2) hybrid BM25+dense search; (3) rerank relevant documents; (4) answerability detection; and (5) specialized guardrails. Our contribution demonstrates how systematic integration of classical IR techniques with advanced prompting tackles SemEval-2026 Task 8 requirements.*

1. Introduction

The SemEval-2026 Task 8 [Katsis et al. 2026] proposes evaluating conversational multi-turn RAG systems based on MTRAGEval [Katsis et al. 2025], where 85% of interactions present contextual dependencies and coreferences. Unlike single-turn RAG systems, the task requires maintaining conversational context across turns, resolving pronominal coreferences, adapting retrieval by question type (*follow-up*, *clarification*, *troubleshooting*), and detecting insufficient documents.

The data [Katsis et al. 2025] consists of multi-turn conversations with: (1) turn history, (2) current question, (3) question type, (4) document corpus, (5) gold standard documents, and (6) expected response evaluated by FANC criteria (*Faithfulness*, *Appropriateness*, *Naturalness*, *Completeness*).

We propose a modular architecture integrating: (1) *Query Rewriting* with Chain-of-Thought for contextual reformulation and *Multi-Query Diversification*, (2) *Hybrid Retrieval* (BM25 + dense embeddings [Zhuang et al. 2024], (3) reranking relevant documents), (4) *Answerability Detection*, and (5) *Specialized Guardrails*.

2. Proposed Architecture

Our architecture operates in five stages (Figure 1): (1) reformulation, (2) diversification, hybrid retrieval, (3) reranking, (4) answerability detection, and (5) generation with guardrails.

(1) Reformulation - Query Rewriting with CoT. Context-dependent questions lack information for retrieval. We use CoT prompting, where generating reasoning steps improves LLM capabilities [Wei et al. 2022]. Our CoT maps coreferences to previous turns (N=3) and reformulates into self-contained queries [Wang and Zhou 2025, Li et al. 2024]. **Example:** "and the headquarters address?" → "What is the address of Apple's headquarters?".

(2) Multi-Query Diversification. We generate five queries [Breuer 2024]: (1) original, (2) entity-focused, (3) action-focused, (4) paraphrased, (5) relation-focused [Lee et al. 2024, Zhang et al. 2024]. Each retrieves top-10 documents (50 candidates max), then reranking [Adeyemi et al. 2024, Reddy et al. 2024] selects top-5.

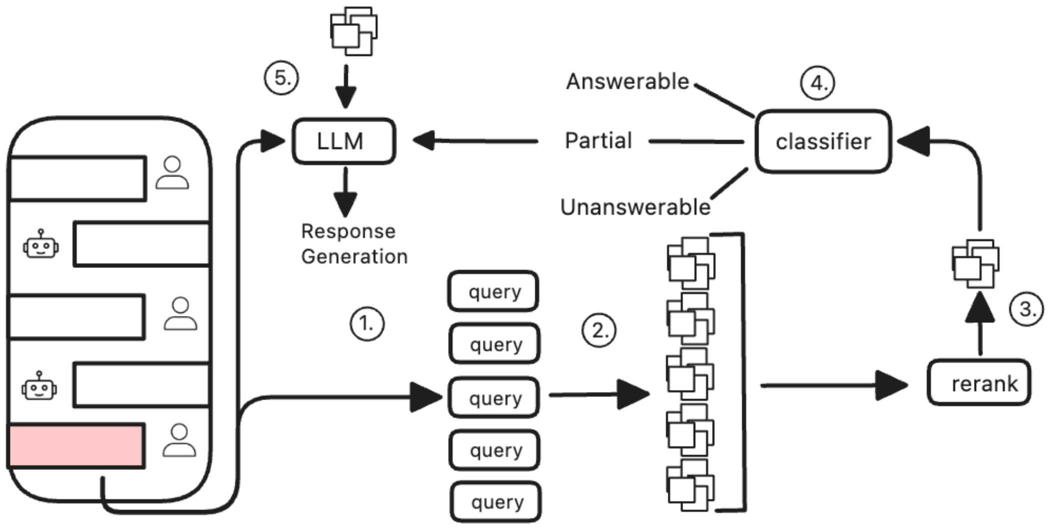

Figure 1. Pipeline: (1) CoT rewriting generates 5 queries, (2) hybrid BM25+Dense retrieval (top-10/query), (3) reranking (top-5), (4) answerability classification, (5) generation with guardrails.

(3) Hybrid Retrieval. Qdrant [Qdrant Team 2024] with: (1) BM25 for lexical retrieval, (2) dense embeddings for semantic retrieval [Zhuang et al. 2024].

(4) Answerability Detection. GPT-4o-mini [OpenAI 2024] categorizes queries as **Answerable**, **Partial**, or **Unanswerable** [Chen and Mueller 2024, Xia et al. 2025], performing conditional actions [Shi et al. 2025, Song et al. 2024].

(5) Guardrails. Six modules ensure FANC [Katsis et al. 2025, Ye et al. 2024]: answerability, ambiguity detection, coreference resolution, faithfulness, completeness, premise correction [Bassani and Sanchez 2024, Rebedea et al. 2025].

3. Conclusion

This work proposed a modular architecture for multi-turn conversational RAG addressing SemEval-2026 Task 8 [Katsis et al. 2026]. The contribution demonstrates how integrating query rewriting, multi-query diversification, hybrid retrieval, answerability detection, and guardrails addresses conversational challenges effectively. Future work includes: (1) ablation analysis, (2) fine-tuning for query rewriting, and (3) evaluation on RADBench [Kuo et al. 2025] and iKAT [Aliannejadi et al. 2024].

References

Adeyemi, M., Oladipo, A., Pradeep, R., and Lin, J. (2024). Zero-shot cross-lingual reranking with large language models for low-resource languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–656.

Aliannejadi, M., Abbasiantaeb, Z., Chatterjee, S., Dalton, J., and Azzopardi, L. (2024). TREC iKAT 2023: A test collection for evaluating conversational and interactive knowledge assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 819–829.

Bassani, E. and Sanchez, I. (2024). GuardBench: A large-scale benchmark for guardrail models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18393–18409.

Breuer, T. (2024). Data fusion of synthetic query variants with generative large language models. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 274–279.

Chen, J. and Mueller, J. (2024). Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200.

Katsis, Y., Rosenthal, S., Fadnis, K., Gunasekara, C., Lee, Y.-S., Popa, L., Shah, V., Zhu, H., Contractor, D., and Danilevsky, M. (2025). mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Transactions of the Association for Computational Linguistics*, 13:784–808.

Katsis, Y., Rosenthal, S., Fadnis, K., Gunasekara, C., Lee, Y.-S., Popa, L., Shah, V., Zhu, H., Contractor, D., and Danilevsky, M. (2026). SemEval-2026 Task 8: Multi-Turn Conversational RAG Evaluation. `https://ibm.github.io/mt-rag-benchmark/MTRAGEval/`. Accessed: 2025-01-15.

Kuo, T.-L., Liao, F., Hsieh, M.-W., Chang, F.-C., Hsu, P.-C., and Shiu, D.-s. (2025). RAD-Bench: Evaluating large language models' capabilities in retrieval augmented dialogues. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3: Industry Track*, pages 868–902.

Lee, Y., Kim, M., and Hwang, S.-w. (2024). Disentangling questions from query generation for task-adaptive retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4775–4785.

Li, Z., Liu, H., Zhou, D., and Ma, T. (2024). Chain of thought empowers transformers to solve inherently serial problems. In *The Twelfth International Conference on Learning Representations*.

OpenAI (2024). Gpt-4o mini: advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`. Accessed: 08 Oct. 2025.

Qdrant Team (2024). Qdrant - high-performance, massive-scale vector database and vector search engine. `https://github.com/qdrant/qdrant`. Accessed: 08 Oct. 2025.

Rebedea, T., Derczynski, L., Ghosh, S., Sreedhar, M. N., Brahman, F., Jiang, L., Li, B., Tsvetkov, Y., Parisien, C., and Choi, Y. (2025). Guardrails and security for LLMs: Safe, secure and controllable steering of LLM applications. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 13–15.

Reddy, R. G., Doo, J., Xu, Y., Sultan, M. A., Swain, D., Sil, A., and Ji, H. (2024). FIRST: Faster improved listwise reranking with single token decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8642–8652.

Shi, Z., Castellucci, G., Filice, S., Kuzi, S., Kravi, E., Agichtein, E., Rokhlenko, O., and Malmasi, S. (2025). Ambiguity detection and uncertainty calibration for question answering with large language models. In *Proceedings of the 5th Workshop on Trustworthy NLP*, pages 41–55.

Song, J., Wang, X., Zhu, J., Wu, Y., Cheng, X., Zhong, R., and Niu, C. (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.

Wang, X. and Zhou, D. (2025). Chain-of-thought reasoning without prompting. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Xia, Z., Xu, J., Zhang, Y., and Liu, H. (2025). A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21381–21396.

Ye, Q., Ahmed, M., Pryzant, R., and Khani, F. (2024). Prompt engineering a prompt engineer. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 355–385.

Zhang, L., Wu, Y., Yang, Q., and Nie, J.-Y. (2024). Exploring the best practices of query expansion with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1872–1883.

Zhuang, S., Ma, X., Koopman, B., Lin, J., and Zuccon, G. (2024). PromptReps: Prompting large language models to generate dense and sparse representations for zero-shot document retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4375–4391.