2nd INTERNATIONAL CONFERENCE ON
Artificial Intelligence
31 May - 1 June 2026

IICAI
Second International Interdisciplinary Conference on Artificial Intelligence

Implementation of a Retrieval-Augmented Generation-Based Question Answering System Using Large Language Models

Mohammad Mahdi HajiRamezanAli, Mohammad Mehdi Homayounpour
Master's Degree in Artificial Intelligence
Ph.D. in Electrical Engineering, Faculty Member at Shahab Danesh University
hajiramezanali@shdu.ac.ir, homayoun@shdu.ac.ir

Abstract

Question-answering (QA) systems are essential in applications such as search engines, virtual assistants, customer service, and education. This is particularly important for Persian, which faces limited training data and computational resources. In this study, we propose a hybrid QA framework combining large language models (LLMs) with retrieval-augmented methods. We compared several LLMs and evaluated their ability to answer complex, multi-step questions. To improve reliability, we introduced an out-of-domain detection mechanism using model-generated data and a lightweight neural network. Experimental results show that Qwen2-7B-Instruct, combined with retrieval augmentation and ranking models, achieved the best performance. The proposed out-of-domain detection mechanism also improved accuracy by 9.51% over human performance while preventing irrelevant responses.

Research method

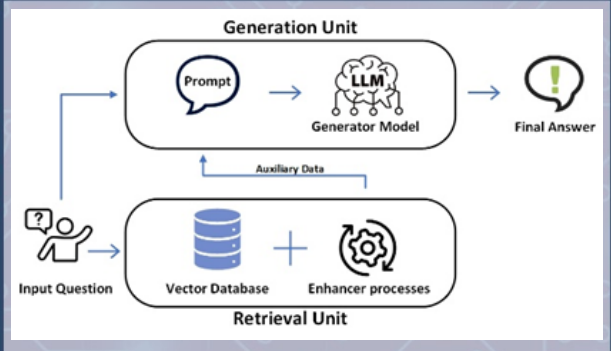

Figure 1: Overview of proposed method

Dataset and experimental setup:

- Utilizing ShahabDanesh University website as RAG data source.
- Operating "BAAI/bge-reranker-v2-m3" as reranker and "Beautiful soap", "Requests" and "HAZM" libraries for data cleaning.
- Compared No RAG (Baseline), Simple RAG and RAG with Reranking methods to find optimal solution for auxiliary data retrieval.

conclusion

Improving factors:

- Utilizing "Qwen2-7B-Instruct" generator.
- Incorporating RAG with Reranking method.
- Storing knowledge in chunk-based manner.
- Use of conversational memory for multi-turn questions.
- Deploying lightweight neural classifier to filter out out-of-scope queries.
- Retrieving up to 7 samples from vector Database.

Potential limitations:

- Incomplete information retrieval due to fixed chunk size.
- Performance degradation in multi-turn conversations (topic shifts).
- Increased computational overhead.

Negative factors:

- Implementing No RAG approach.
- Usage of "Mistral-7B-Instruct-v0.3" model.
- Excessive retrieval context.

The purpose of the research

This study presents a hybrid RAG-based question-answering framework that improves response quality and reliability by addressing intent recognition, out-of-domain detection, multi-step queries, and support for low-resource languages like Persian. To achieve this objective, a set of research questions was formulated, and a series of experiments was conducted to identify the most effective system design based on the outcomes of these evaluations. The research questions are:

- What is the optimal method for storing datasets in a vector database?
- To what extent does information retrieval improve the overall performance of the system?
- Which model demonstrates superior performance in response generation?
- How does the system behave when faced with complex, multi-step queries?
- How can the system's performance in identifying irrelevant or out-of-domain questions be improved?

Practical or simulation results

Retrieval parameters evaluated :

- Examined performance of chunk-based and sentence-based data vectors.
- Retrieved 3 to 20 vectors in simple RAG and 5 to 50 vectors in RAG with reranking.
- Stored chunk-based vectors in following configurations: 32, 64, 128, 256, 512, 1024 and 2048 words.

Models experimented:

- Llama-3.1-8B-Instruct
- Qwen2-7B-Instruct
- Mistral-7B-Instruct-v0.3

Two-step question answering:

- Acquiring One-shot (memoryless) and Full-shot (memory accompanied) dataset for two-step question answering test.

Out-Of-Domain detection:

- Employing "Gemini 2.5 Pro" and "GPT-4o" for out-of-domain dataset preparation.
- Designing a Lightweight Neural Network for Out-of-Scope Question Detection.

References

[1] R. Akkiraju *et al.*, "FACTS About Building Retrieval Augmented Generation-based Chatbots," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.07858.

[2] C. Sharma, "Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers," May 2025, [Online]. Available: http://arxiv.org/abs/2506.00054.

[3] D. Cabezas, R. Fonseca-Delgado, I. Reyes-Chacón, P. Vizcaino-Imacaña, and M. Morocho-Cayamcela, "Integrating a LLaMa-based Chatbot with Augmented Retrieval Generation as a Complementary Educational Tool for High School and College Students," in *Proceedings of the 19th International Conference on Software Technologies*, SCITEPRESS - Science and Technology Publications, 2024, pp. 395–402.

[4] H. Hosseini, M. S. Zare, A. H. Mohammadi, A. Kazemi, Z. Zojaji, and M. A. Nematbakhsh, "PersianRAG: A Retrieval-Augmented Generation System for Persian Language," in *2024 15th International Conference on Information and Knowledge Technology (IKT)*, Dec. 2024, pp. 272–278.

[5] S. B. Hosseinbeigi, S. Asghari, M. Ali, S. Kashani, M. H. Shalchian, and M. A. Abbasi, "Advancing Retrieval-Augmented Generaton for Persian Development of LanguageModels, Comprehensive Benchmarks, and best Practices for Optimization." , [Online]. Available: https://arxiv.org/abs/2501.04858.

[6] C. Qian, H. Qi, G. Wang, L. Kunc, and S. Potdar, "Distinguish Sense from Nonsense: Out-of-Scope Detection for Virtual Assistants," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, PA, USA: Association for Computational Linguistics, Dec. 2022, pp. 502–511.

[7] P. Wang *et al.*, "Beyond the Known: Investigating LLMs Performance on Out-of-Domain Intent Detection," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia: ELRA and ICCL, May 2024, pp. 2354–2364. [Online]. Available: https://aclanthology.org/2024.lrec-main.210/.

[8] S. Chang *et al.*, "RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation," in *Advances in Neural Information Processing Systems 37*, San Diego, California, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024, pp. 21999–22027.

[9] Y. Wang, R. Ren, J. Li, X. Zhao, J. Liu, and J.-R. Wen, "REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2024, pp. 5613–5626.

[10] R. Zadkamali, S. Momtazi, and H. Zeinali, "Intent detection and slot filling for Persian: Cross-lingual training for low-resource languages," *Natural Language Processing*, vol. 31, no. 2, pp. 559–574, Mar. 2025.

[11] M. M. HajiRamezanAli, H. Deldar, and M. M. Homayounpour, "Improvement in intent detection and slot filling by model enhancement and different data augmentation strategies," *1st International Conference on Artificial Intelligence*, vol. 2, no. 1, pp. 7–14, 2024.

[12] M. Akbari, A. Mohades, and M. H. Shirali-Shahreza, "A Hybrid Architecture for Out of Domain Intent Detection and Intent Discovery," in *2025 11th International Conference on Web Research (ICWR)*, IEEE, Apr. 2025, pp. 137–144.

[13] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," Feb. 2019, [Online]. Available: http://arxiv.org/abs/1803.08375.

[14] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," Jun. 2023, [Online]. Available: http://arxiv.org/abs/1606.08415.

[15] G. Cybenko and G. Cybenkot, "Approximation by superpositions of a sigmoidal function Approximation by Superpositions of a Sigmoidal Function*," vol. 2, no. 4, 1989.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 2017, [Online]. Available: http://arxiv.org/abs/1412.6980.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.