BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY
Excel@FIT 2026
http://excel.fit.vutbr.cz

Advanced RAG Architecture for Intelligent Travel Agency Assistance

David Švancer

Abstract
Traditional chatbots in the travel industry often struggle with hallucinations and loss of conversational context, especially when processing unstructured data like guest reviews. This work proposes an Advanced RAG architecture utilizing semantic routing, multi-query expansion, and cross-encoder reranking to ensure factual accuracy. Experimental results demonstrate that the proposed multi-stage pipeline significantly outperforms naive RAG systems in terms of faithfulness and context precision. The system provides a robust solution for travel agencies to leverage large-scale review datasets for personalized and reliable customer support.

*xsvancd00@vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

[Motivation] In the travel industry, information accuracy is paramount. Unlike general-purpose AI, large language models are known to produce hallucinated outputs, making it critical that a travel chatbot does not hallucinate prices, amenities, or hotel characteristics and vibes [1]. With thousands of unstructured guest reviews available, there is a massive potential to provide personalized advice that goes beyond simple SQL filtering of database attributes.

[Problem definition] Naive RAG systems often fail in two ways: first, they lose track of context in multi-turn dialogues (the "it" or "there" problem); second, they often retrieve irrelevant documents because a simple vector search might miss specific keywords or nuanced sentiments in reviews. The goal is to build a pipeline that is context-aware, factually grounded, and capable of mining deep insights from travel reviews.

[Existing solutions] Many baseline RAG implementations rely on a single vector-based retrieval step. While efficient, these solutions lack the ability to handle complex queries or distinguish between different intents (e.g., small talk vs. specific hotel search), often leading to irrelevant or generalized responses [2].

[Our solution] This work proposes a modular Advanced RAG pipeline. It incorporates a History-Aware Rewriter to maintain context, a Semantic Router to optimize execution paths, and a combination of Hybrid Search, Multi-query and a Cross-Encoder Reranker to ensure that only the most relevant review segments are presented to the LLM [3].

[Contributions] This work provides a comprehensive comparison of embedding models (BGE vs. Nomic) and validates the incremental benefits of advanced components like Multi-Query generation and Reranking through standardized Ragas metrics.

2. Advanced RAG Architecture (Figure 1)

The core of the project is the multi-stage pipeline shown in **Figure 1**. To solve the issue of context loss, the first step is **History-Aware Retrieval**, which reformulates user queries based on previous dialogue turns.

The **Query Router** then analyzes the intent; if a travel search is detected, the query is expanded via **Multi-Query Generation** to capture different semantic perspectives. We use a **Hybrid Search** approach (combining Vector DB with BM25 [4]) to ensure we don't miss specific hotel names or attributes. Finally, the top-k retrieved documents are processed by a **Reranker**, which re-orders them based on their true relevance to the user's specific need before being passed to the Large Language Model [3].

3. Experimental Evaluation

The performance of the system was measured using three distinct benchmarks, as illustrated in the graphs on the poster.

3.1 Embedding Model Comparison

As shown in **Figure 3** We benchmarked **BGE-M3** against **Nomic-Embed**. While both models performed well, **BGE-M3** provided better alignment with the travel-specific vocabulary found in our review dataset, leading to higher hit rates in initial retrieval.

3.2 Incremental Pipeline Improvements

We evaluated the "evolution" of the architecture from a Naive RAG to our final Advanced version. The results clearly show that adding **Multi-Query expansion** significantly boosts *Context Recall*, while the **Reranker** is the primary driver for high *Context Precision*.

3.3 Ragas Metrics

The final evaluation focused on **Answer Correctness** and **Faithfulness**. Our Advanced RAG architecture achieved a **16.7% improvement in Answer Correctness** compared to the semantic baseline. This enhancement demonstrates the system's superior ability to synthesize accurate responses for complex multi-step queries. While maintaining a solid grounding with a **Faithfulness score of 69.2%**, the architecture effectively ensures that the LLM's responses are derived from the retrieved review data, significantly reducing the risk of hallucinations compared to standard LLM outputs [5, 1].

4. Contextual Dialogue Example

The **Chat Example** on the poster demonstrates the system's ability to mine insights from reviews that simple database filters would miss. When a user asks for a "quiet hotel," the system does not just look for a metadata tag; it identifies reviews mentioning "excellent soundproofing" or "peaceful nights." Furthermore, it handles follow-up questions by correctly identifying that "there" refers to the previously discussed place.

5. Conclusions

The project demonstrates that for domain-specific applications like travel, a naive RAG approach is insufficient. By implementing a structured, multi-stage pipeline, we can provide users with reliable, review-based insights that are both contextually aware and factually accurate. Future work involves integrating real-time pricing APIs to supplement the static review knowledge base.

Acknowledgements

I would like to thank my supervisor, Ing. Martin Kostelník, for his valuable help and guidance throughout the work on this thesis.

References

[1] Ziwei Ji et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023.

[2] Patrick Lewis, Ethan Perez, Aleksander Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

[3] Xinyu Gao Kangxiang Jia Jinliu Pan Yuxi Bi Yi Dai Jiawei Sun Meng Wang Haofen Wang Yunfan Gao, Yun Xiong. Retrieval-augmented generation for large language models: A survey. *Frontiers of Computer Science*, 2024.

[4] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 2009.

[5] Shahul Es, Jidnya Shah, Prateek Yadav, Rishi Puri, Nils Schwenke, Robert Watson, and Maria Watson. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*, 2023.