Enhancing Large Language Models with Retrieval-Augmented Generation: Improvements and Applications in the Tourism Industry

Osaka University of Economics and Law Li Feng
Osaka University of Economics and Law Iwata Yoritaka
Osaka University of Economics and Law Fukase Kiyoshi

Keywords: Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Tourism Industry

1. Introduction

In recent years, generative AI (generative artificial intelligence) such as Claude2, and ChatGPT, which are based on large language models (LLMs), have achieved remarkable progress in natural language processing [1]. It demonstrates powerful text generation capabilities and versatility across various domains. However, despite these advances, several critical challenges remain. LLMs are prone to generate hallucinated information[2], struggle to provide up-to-date knowledge, and require substantial computational resources for training and deployment. To address these limitations, Retrieval-Augmented Generation (RAG) has been proposed as a promising approach. By integrating external knowledge retrieval with generative models, RAG has been shown to enhance the accuracy of answer, reduce hallucinations, and take into account domain-specific adaptation without retraining the entire model [3].

Meanwhile, the tourism industry is undergoing rapid digital transformation (DX), with increasing reliance on online travel agencies (OTAs), AI-based chatbots, and smart tourism services [4]. For travelers, timely information, personalized recommendations, and multilingual support are desirable. Conventional LLM applications often fail to meet these needs due to outdated knowledge or insufficient domain-specific information. Therefore, exploring the application of RAG to enhance LLMs and its potential in the tourism industry is both technically and practically significant. Using RAG, more reliable and context-aware information for travelers can be obtained. This study aims to investigate how RAG can improve LLM performance and be applied to real-world tourism scenarios.

2. Objective

The primary purpose of this study is to clarify the role of RAG in improving their overall performance of LLM. In particular, its key contribution to enhance knowledge freshness, reducing hallucination rates, lowering computational costs, and enabling greater domain adaptability is identified. By conducting a theoretical analysis and reviewing prior research, this study formulates the hypothesis that RAG-based architecture provides a more reliable and efficient approach to LLM.

Furthermore, as a secondary purpose, this study aims to explore practical applications of RAG within the tourism industry. Tourism industry provides an ideal testing ground for the implementation of advanced language technologies, as the tourism industry requires timely, accurate, and context-sensitive information. By examining potential usage of AI-powered travel consultation, automated itinerary generation, multilingual support, and crisis management, this study seeks to evaluate how RAG can contribute to DX in the tourism sector. The findings are expected to offer insights into both the technical significance of RAG for natural language processing and its practical value for industry innovation.

3. Problems of conventional LLM

LLMs have achieved strong performance in natural language processing, but three critical problems are reported [5]. First, LLMs often generate hallucinations, producing outputs that sound plausible but are not true. Second, they suffer from knowledge, because the information encoded during training cannot be easily updated. Finally, computational cost is a major issue, as retraining or fine-tuning these large models requires significant resources. These problems limit their reliability in knowledge-intensive tasks.

4. Implementation of RAG

RAG is proposed to address the above problems. The key idea is to add a retrieval module to the generation process. When a user query is given, the system retrieves relevant passages from an external knowledge source and then uses them. This design enables the model to access up-to-date and reliable information, and reduce reliance on memorized knowledge. Retrieval can be implemented through sparse methods like BM25 [6].), dense methods (embedding-based search), or hybrid approaches, supported by vector database technology for scalability.

According to the literature, RAG has been actively studied in tasks such as question answering, knowledge-intensive dialogue, and fact verification. Compared with conventional LLMs, RAG-implemented LLM has shown to improve accuracy and reduce hallucination. Furthermore, it allows for domain-specific adaptation without the need for expensive retraining. Applications in fields such as medicine and law confirm that RAG can significantly increase the reliability of generated outputs. Overall, research trends suggest that RAG is becoming a central method for enhancing the performance of LLMs. Note here that RAG has not been seriously applied to the LLM-based generative AI in the tourism industry.

5. Summary

In conclusion, several main improvements can be achieved by RAG:

1) Knowledge freshness: Since RAG retrieves information at inference time, it can provide answers that reflect the latest knowledge.

2) Hallucination mitigation: By grounding responses in retrieved documents, RAG improves factual reliability.

3) Cost efficiency: RAG reduces the need for frequent retraining of large models, lowering computational costs.

4) Domain adaptability: Retrieval queries can be tailored to specific domains, enabling flexible applications.

These four advantageous points are crucially valuable in the to the LLM-based generative AI in the tourism industry.

Despite these advantages, some issues still remain unsolved. First, the effectiveness of RAG depends heavily on the quality of retrieval, as irrelevant or noisy documents may still lead to poor responses. Second, handling unstructured or multimodal data remains difficult, as most current methods focus on text. Third, there are multilingual limitations, since high-quality retrieval resources are not equally available across all languages. These open issues highlight the need for further improvement before RAG can be fully deployed in complex real-world applications such as tourism.

References:

[1]Undetectable 「Claude vs GPT 4: Key Differences Compared - Undetectable AI」,https://undetectable.ai/blog/claude-vs-gpt-4/（参照日時:2025-8-24）

[2] Y. Zhang, Y. Li, et al., “Siren’s song in the ai ocean: A survey on hallucination in large language models,” arXiv:2309.01219, 2023.

[3] C. Mala, G. Gezici, et al. “Hybrid Retrieval for Hallucination Mitigation in Large Language Models: A Comparative Analysis” arXiv:2504.05324

[4] Gretzel, U., Sigala, M., Xiang, Z. et al. “Smart tourism: foundations and developments”, Electron Markets 25, 179–188 (2015).

[5] Y Gao, Y Xiong, et al., “Retrieval-augmented generation for large language models: A survey” arXiv:2312.10997, 2023

[6] Qiita 株式会社 「BM25 の Python 高速ライブラリ BM25-Sparse を日本語で使いたい」 https://business.ntt-east.co.jp/（参照日時:2025-8-24）