ISSN 2093-0542

KOREAN SOCIETY FOR INTERNET INFORMATION

The 20th Asia Pacific International Conference on Information Science and Technology

(APIC-IST 2025)

July 06-09, 2025, SAii Laguna Resorts, Phuket, Thailand

http://www.apicist.org

Proceedings of APIC-IST 2025

| Organized by |

Korean Society for Internet Information (KSII)

Contents

Developing a RAG-based Intelligent Chatbot using Dify and Ollama: Focusing on educating Developers on the LMS Environment

Jaeho Kim, Ji Hoon Kong, Ki Du Kim, and R. Young Chul Kim*
Department of Software Communication, Hongik University
Sejong, South Korea
[e-mail: jaehokim1005@g.hongik.ac.kr, go400s@gmail.com, kdkim@tta.or.kr, *bob@hongik.ac.kr]
*Corresponding author: R. Young Chul Kim

Abstract

The conventional LMS solutions don't support knowledge-based requirements within domain-specific educational contexts. The near future LMS platforms should help with the knowledge-driven demands of learners. To solve this, we propose a Learning Management System (LMS) with a Retrieval-Augmented Generation (RAG) system, as outlined in the official NestJS documentation. We developed the Retrieval-Augmented Generation (RAG) system in collaboration with Ollama and Dify, which facilitates the rapid and easy construction and operation of RAG systems. The system provides trainees and developers with answers to development-related queries, based on official documentation. This allows the RAG within the LMS to assist developers in resolving knowledge-based issues encountered during their projects. Ultimately, the RAG system enhances the quality of education in the software education domain, enabling personalized learning for learners.

Keywords: Learning management system, retrieval-augmented generation, chatbot, Dify, Ollama

1. Introduction

As AI technology advances, the educational domain is undergoing a profound paradigm shift, fundamentally reshaping both learner experiences and pedagogical strategies [1, 2, 3, 4]. In this context, Learning Management Systems (LMS) serve as a pivotal infrastructure for integrating AI technologies, playing an indispensable role in modern Artificial Intelligence in Education (AIED) [4, 5]. Before the advent of AI, while LMS platforms offered high efficiency in areas such as learning content delivery and personnel management [6, 7], they inherently faced limitations in terms of real-time responsiveness and personalized learning support [7]. For instance, instructors struggled to provide uniform support to all learners, and the varied expression of error messages or coding issues from students made it challenging for instructors to offer swift and precise answers. Furthermore, there were limitations in providing immediate assistance during project execution when real-time questions arose, often resulting in learners being confined to unidirectional references of learning materials, such as videos or files.

Recently, the emergence of Large Language Models (LLMs) has demonstrated exceptional performance in natural language-based question answering, summarization, explanation, and coding, thereby proving their considerable

This research was conducted with the support of the Korea Creative Content Agency (Project Name: Artificial Intelligence-Based Interactive Multimodal Interactive Storytelling 3D Scene Authoring Technology Development, Project Number: RS-2023-00227917, Contribution Rate: 100%) and the Korea Research Foundation's four, Brain Korea 21 (Project Name: Ultra-Distributed Autonomous Computing Service Technology Research Team, Project Number: 202003520005).

potential for application within the educational domain [8, 9]. While general LLMs exhibit limitations in incorporating up-to-date knowledge and often require extensive training data for domain-specific expertise [10], Retrieval-Augmented Generation (RAG) has emerged as a prominent approach to address these challenges [10, 11]. RAG is a powerful AI technique that simultaneously ensures information accuracy, domain-specificity, and knowledge currency by enabling LLMs to retrieve relevant external documents before text generation [10].

This study proposes a novel knowledge support system that integrates the RAG technique into Learning Management Systems (LMS), thereby mitigating the limitations of existing systems, bolstering real-time problem-solving capabilities, and ultimately maximizing learner potential and enhancing productivity. For system construction, a RAG-based LMS was effectively implemented by utilizing Dify, an open-source tool supporting Backend as a Service, and Ollama, an LLM execution environment. This chatbot-based AI-powered knowledge support system provides learners with real-time, query-response-based feedback, delivered as personalized and domain-specific answers. Furthermore, it offers tangible benefits to all stakeholders by reducing repetitive inquiries for instructors and fostering the potential for customized guidance for learners. This paper elucidates the theoretical background and technical implementation process of the proposed system, providing meticulous detailing of its application in a real-world setting within the Hongik University Metaverse Academy.

2. Related Work

LMS has become an indispensable element in education [6]. The accelerated integration of AI into Learning Management Systems (LMS) primarily occurred after the COVID-19 pandemic [12]. COVID-19 necessitated online education, shifting the Learning Management System (LMS) from a supplementary role to a central position in online learning [12].
Furthermore, the rapid advancement of artificial intelligence technology is bringing about transformative changes in the education sector, which can be observed across various aspects of Learning Management System (LMS) platforms, including the personalization of learning experiences and enhanced instructional support [5, 13]. As this study focuses on the implementation case of a developer education LMS chatbot utilizing Dify and Ollama, the following specific cases will be primarily examined.

2.1 An AI Application in LMS

Lee *et al.* [14] implemented a RAG system for a student Q&A chatbot within an LMS, leveraging past chat history data and FAQ data provided by the LMS as retrieval sources. While this study provided Q&A functionality to students, it primarily focused on enabling inquiries related to LMS content and feature utilization rather than providing knowledge-based learning support.

2.2 An AI Technology Application to Fulfill Knowledge-Based Learner Needs

Many developers and problem solvers are currently utilizing large language models (LLMs) for various tasks. Furthermore, recent research has actively explored RAG systems that provide answers by referencing libraries such as PyTorch and NumPy within the Python language [15]. However, these studies are limited to the scope of Python libraries, and research on retrieval-augmented generation through framework-level documentation reference remains insufficient.

3. Our RAG-based LMS Chatbot System

This section describes the overall design and implementation process of the RAG-based LMS chatbot system, utilizing the Dify platform and Ollama.

3.1 System Architecture

The system primarily consists of an Angular frontend application, using Nginx as a proxy server; a NestJS backend application with Express as its web server; and a Flask-based backend application with Gunicorn as its web server. Here, the Flask-based backend application leverages the open-source Dify platform. Dify is an open-source platform that maximizes development convenience by providing Backend as a Service for RAG pipeline, Agent construction, and management. It enhances scalability for AI function integration by modularizing each layer. For ease of management, Ollama is designated as the LLM Provider, and Docker Containers are configured to share the same Docker Network. Ollama is a tool that serves various LLMs and embedding models, facilitating their easy use in a local environment.

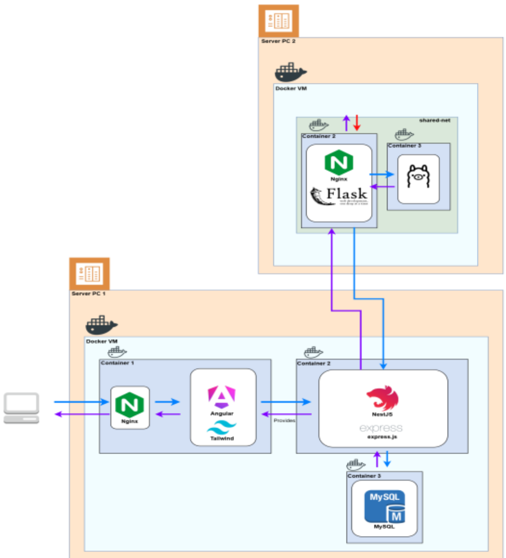

Fig. 1. System Architecture

3.2 Data Source

The knowledge base for the retriever constructed in this study comprises 134 Markdown files from the official NestJS documentation [16]. NestJS is a framework used for modern web application development, specifically for TypeScript-based Backend development. Since the official NestJS documentation is extensive and highly reliable, a chatbot referencing it for developer education can provide junior developers facing real development challenges with answers that thoroughly refer to rich resources, offering high accuracy and diverse code implementation possibilities.

3.3 Chunking

To prevent the 134 original NestJS documentation files from failing to reflect the overall context and to avoid weakened semantic connectivity among individual chunks, a Parent-Child Chunking strategy is employed.

- **Parent Chunk:** Each of the entire Markdown documents is defined as a single Parent Chunk to preserve the overall context and structure of the document.
- **Child Chunk:** Child Chunks are generated by splitting Parent Chunks into segments with a maximum length of 512 characters. User queries first undergo similarity calculation with these Child Chunks to select initial similar Chunks, after which the original Parent Chunk corresponding to the selected Child Chunk is passed to the LLM.

3.4 Embedding

The 134-chunked NestJS official documents undergo vector embedding and indexing processes. The BGE M3 model, obtained from Ollama, is used as the embedding model. This model was selected because the bge-m3 model's excellent multilingual support is expected to process queries from Korean users within Hongik University effectively. All embeddings were conveniently performed via the Dify platform's web interface.

3.5 Generation

Qwen3 and Exaone3.5 are selected as generation models. This decision is based on the same reasoning as the embedding model selection: the need for models that excel in Korean language processing.

3.6 Feedback Loop and System Refinement

For continuous improvement, enhanced answer quality, and personalized learning support, this system collects learner queries and stores their inputs and outputs in a database to identify learner needs. MySQL is used as the database, and access is implemented via TypeORM within NestJS.

4. Conclusions

Until now, there have been inherent limitations of conventional Learning Management Systems (LMS). To overcome these challenges, we propose integrating a RAG-based chatbot into an LMS. Specifically, by leveraging domain-specific knowledge from the official NestJS documentation as a data source and utilizing LLMs and Embedding Models through Ollama within the Dify platform, we successfully establish a personalized Q&A service for junior developers.

References

[1] E. Molina, C. Cobo, J. Pineda, and H. Rovner, "AI revolution in education: What you need to know," *Digital Innovations in Education*, World Bank, 2024.

[2] U.S. Department of Education, Office of Educational Technology, Artificial Intelligence and Future of Teaching and Learning: Insights and Recommendations, Washington, DC, 2023.

[3] UNESCO, International Forum on AI and Education: Ensuring AI as a Common Good to Transform Education, Synthesis Report, UNESCO, 2022, (ED-2022/WS/14)

[4] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol.8, pp. 75110-75122, 2020. doi:10.1109/ACCESS.2020.2988510.

[5] R. Luckin, W. Holmes, M. Griffiths, and L. B. Forcier, Intelligence Unleashed: An argument for AI in Education, Pearson, 2016.

[6] A. Aldiab et al., "Education System: A Case Review for Saudi Arabia," in *Proc. of the 2nd International Conference on Energy and Power (ICEP2018)*, *Energy Procedia*, vol.160, pp.731-737, 2019.

[7] N. A. Adzharuddin and L. H. Ling, "Learning Management System (LMS) among University Students: Does It Work?," *International Journal of e-Education, e-Business, e-Management and e-Learning*, vol.3, no.3, pp. 248-252, 2013.

[8] M. Neumann, M. Rauschenberger, and E.-M. Schön, "We Need To Talk About ChatGPT: The Future of AI and Higher Education," in *Proc. of the 2023 IEEE/ACM 5th International Workshop on Software Engineering Education for the Next Generation (SEENG)*, 2023. doi:10.1109/SEENG59157.2023.

[9] I. Pesovski, R. Santos, R. Henriques, and V. Trajkovik, "Generative AI for Customizable Learning Experiences," *Sustainability*, vol.16, no.7, 2024. doi:10.3390/su160703034.

[10] W. X. Zhao et al., "A Survey of Large Language Models," *arXiv preprint arXiv:2303.18223*, 2025. doi:10.48550/arXiv.2303.18223.

[11] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *arXiv preprint arXiv:2005.11401*, 2021. (Accepted at NeurIPS 2020). doi:10.48550/arXiv.2005.11401.

[12] S. Qazi et al., "AI-Driven Learning Management Systems: Modern Developments, Challenges and Future Trends during the Age of ChatGPT," *Computers, Materials and Continua*, vol.80, no.2, pp.3289-3314, 2024. doi:10.32604/cmc.2024.048893.

[13] I. Roll and R. Wylie, "Evolution and Revolution in Artificial Intelligence in Education," *International Journal of Artificial Intelligence in Education*, vol.26, no.2, pp.582-599, 2016. doi:10.1007/s40593-016-0110-3.

[14] J. S. Lee, J. M. Lee, and J. H. Yoo, "University Learning Management System Student Q&A Chatbot Using Large Language Models Based on Retrieval-Augmented Generation," *Journal of Broadcast Engineering*, vol.29, no.5, pp.581-591, 2024. doi:10.5909/JBE.2024.29.5.581.

[15] Z. Z. Wang, A. Asai, X. V. Yu, F. F. Xu, Y. Xie, G. Neubig, and D. Fried, "CodeRAG-Bench: Can Retrieval Augment Code Generation?," *arXiv preprint arXiv:2406.14497*, 2025. doi:10.48550/arXiv.2406.14497.

[16] NestJS Official Documentation, [Online]. Available: https://docs.nestjs.com/