ABSTRACT

The increasing demand for efficient IT service management in higher education has exposed the limitations of traditional helpdesk models, which rely heavily on manual processes and human operators. These legacy systems often suffer from slow response times, growing ticket backlogs, and staff fatigue. To address these challenges, this study introduces an intelligent, LLM-based chatbot that integrates the DeepSeek-v3 API with a Retrieval-Augmented Generation (RAG) strategy, specifically adapted for an Indonesian university's IT service desk environment. The proposed solution is built on a modular architecture that automates document ingestion using multi-format preprocessing, dynamic chunking, and embedding generation. These embeddings are stored in Supabase to enable efficient top-k contextual retrieval. When a user submits a query, relevant knowledge segments are retrieved in real time and synthesized into coherent, context-aware responses by the DeepSeek-v3 language model. The system was evaluated through functional testing, performance benchmarking, and the standardized User Experience Questionnaire (UEQ). Results showed stable 24/7 operation with over 98% retrieval accuracy and an average response time of 2.1 ± 0.2 seconds—representing a 35% improvement over a legacy helpdesk baseline. A user study involving 26 IT staff and students yielded a high UEQ score of 4.5 ± 0.3 (out of 5), indicating strong satisfaction in both pragmatic (efficiency, dependability) and hedonic (stimulation, novelty) dimensions. Additionally, nine prompting strategies were compared using statistical analysis to optimize response quality. This study demonstrates the viability and scalability of integrating LLMs into institutional IT infrastructures and provides practical guidance for deploying RAG-based chatbots in resource-constrained academic settings.

Keywords: IT helpdesk, large language models, retrieval-augmented generation, chatbot