ABSTRACT

The growth of scientific literature in the Information Technology field presents an information overload challenge that makes manual technical information retrieval inefficient. Conventional keyword-based search systems have limitations in understanding the context of questions, while the direct use of Large Language Models (LLMs) is highly prone to factual hallucinations. The Retrieval-Augmented Generation (RAG) architecture provides a solution to minimize hallucinations by grounding answers in reference documents. However, the performance of the RAG system relies heavily on the text-chunking strategy for lengthy academic documents. Therefore, this study aims to implement an LLM chatbot and find the optimal configuration for the Chunk Size, Chunk Overlap, Top-K parameters, and retrieval methods (Similarity Search and Maximal Marginal Relevance) to prevent the loss of information context (Lost in the Middle). The system is built using a knowledge base consisting of 100 scientific article documents to test the system's robustness against distractor documents (noise), with a controlled evaluation using 10 sample documents as the ground truth reference. Model performance is evaluated quantitatively using Response Time, Coverage, ROUGE-1, ROUGE-L, and BERTScore. The evaluation results from 72 experimental scenarios indicate that the most optimal RAG configuration is achieved using the Similarity Search method, a Chunk Size of 512 characters, a Chunk Overlap of 20%, and a Top-K of 8. This configuration yielded a document retrieval success rate (Coverage) of 100%, a ROUGE-1 score of 0.611, a ROUGE-L score of 0.567, and a very high semantic accuracy (BERTScore) of 0.842, with an efficient and responsive average computation time of 2.06 seconds.

Keywords : Chatbot, Large Language Model, Retrieval-Augmented Generation, Chunk Size, Top-K, Similarity Search, Evaluation