ABSTRAK

Pertumbuhan literatur ilmiah di bidang Teknologi Informasi menghadirkan tantangan kelebihan informasi (*information overload*) yang membuat pencarian informasi teknis secara manual menjadi tidak efisien. Sistem pencarian konvensional berbasis kata kunci memiliki keterbatasan dalam memahami konteks pertanyaan, sementara penggunaan *Large Language Model* (LLM) secara langsung rentan terhadap halusinasi fakta. Arsitektur *Retrieval-Augmented Generation* (RAG) hadir sebagai solusi untuk meminimalisir halusinasi dengan mengunci (*grounding*) jawaban pada dokumen referensi. Namun, performa sistem RAG sangat bergantung pada strategi pemotongan teks dokumen akademik yang panjang. Oleh karena itu, penelitian ini bertujuan untuk mengimplementasikan *chatbot* LLM dan mencari konfigurasi optimal pada parameter *Chunk Size*, *Chunk Overlap*, *Top-K*, serta metode penarikan data (*Similarity Search* dan *Maximal Marginal Relevance*) guna mencegah hilangnya konteks informasi (*Lost in the Middle*). Sistem dibangun menggunakan basis pengetahuan berupa 100 dokumen artikel ilmiah guna menguji ketahanan sistem terhadap dokumen pengecoh (*noise*), dengan evaluasi terkontrol menggunakan 10 dokumen sampel uji sebagai acuan kunci jawaban (*ground truth*). Evaluasi kinerja model diukur secara kuantitatif menggunakan metrik *Response Time*, *Coverage*, ROUGE-1, ROUGE-L, dan BERTScore. Hasil evaluasi dari 72 skenario eksperimen menunjukkan bahwa konfigurasi RAG paling optimal dicapai pada penggunaan metode *Similarity Search*, *Chunk Size* 512 karakter, *Chunk Overlap* 20%, dan batasan Top-K 8. Konfigurasi ini menghasilkan tingkat keberhasilan penarikan dokumen (*Coverage*) sebesar 100%, skor ROUGE-1 0,611, ROUGE-L 0,567, serta akurasi semantik (BERTScore) yang sangat tinggi sebesar 0,842, dengan rata-rata waktu komputasi yang efisien dan responsif yaitu 2,06 detik.

Kata kunci : *Chatbot, Large Language Model, Retrieval-Augmented Generation, Chunk Size, Top-K, Similarity Search*, Evaluasi.