ABSTRAK

Perkembangan *Large Language Models (LLM)* telah memberikan kemajuan signifikan pada sistem *chatbot*, namun permasalahan *hallucination* jawaban yang terdengar meyakinkan namun tidak faktual masih menjadi tantangan utama. Penelitian ini bertujuan mengoptimalkan chatbot berbasis LLM melalui integrasi *Retrieval-Augmented Generation (RAG)* dan *web search* untuk mengurangi *hallucination* serta meningkatkan relevansi dan akurasi jawaban. Model LLaMA 3 dilakukan *fine-tuning* menggunakan dataset berbahasa Indonesia pada domain web programming. Sumber pengetahuan eksternal diperoleh melalui proses web *scraping* menggunakan Firecrawl, yang kemudian diindeks pada *vector store* untuk mendukung pencarian berbasis *semantic search*. Sistem dirancang dengan alur *end-to-end* yang memadukan tahap *retrieval*, *grading* dokumen, dan *generation*. Evaluasi dilakukan menggunakan metrik BERTScore terhadap 200 pertanyaan uji. Hasil pengujian menunjukkan peningkatan performa signifikan pada model LLM+RAG dibandingkan LLM murni, dengan nilai F1-score meningkat dari 0,5729 menjadi 0,6928. Integrasi RAG dan *web search* terbukti efektif dalam mengurangi *hallucination* dan menghasilkan jawaban yang lebih faktual serta kontekstual.

Kata Kunci *: Large Language Models*, *Chatbot*, *Retrieval-Augmented Generation*, *Hallucination*.

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG