

ABSTRAK

Sektor peternakan menghadapi tantangan signifikan dari penyakit ternak yang penanganannya sering kali terhambat oleh keterbatasan sumber daya manusia. Penelitian ini mengusulkan pengembangan *chatbot* berbasis *Large Language Models* (LLM) dengan pendekatan *Retrieval Augmented Generation* (RAG) sebagai solusi. Namun, implementasi model RAG-Fusion sering kali menghadapi masalah redundansi dokumen yang dapat menurunkan efisiensi dan akurasi. Tujuan dari penelitian ini adalah untuk mengoptimalkan RAG-Fusion dengan mengurangi redundansi dokumen melalui seleksi dan pemeringkatan yang lebih ketat, sehingga mampu menyediakan solusi konsultasi kesehatan ternak yang lebih cepat dan akurat.

Penelitian ini menggunakan metode pengembangan sistem RAG-Fusion yang berfokus pada penanganan redundansi dokumen hasil pengambilan dengan teknik *Semantic Deduplication* (SemDeDup). Data diperoleh dari Dinas Ketahanan Pangan, Pertanian, dan Perikanan (DKPPP) Kabupaten Temanggung serta LLM generatif Deepseek V3, yang kemudian dijadikan basis pengetahuan eksternal. Proses pengembangan mencakup pembuatan kueri alternatif, pencarian *chunks* relevan, *reranking* menggunakan *Reciprocal Rank Fusion*, serta penerapan SemDeDup dengan klusterisasi k-means dan penghitungan kesamaan kosinus untuk menghilangkan duplikasi semantik. Respons akhir dihasilkan oleh model GPT-4 berdasarkan konteks yang telah dideduplikasi.

Implementasi SemDeDup dalam sistem RAG-Fusion terbukti efektif dalam meningkatkan efisiensi konteks dan mengurangi beban komputasi pada model LLM, meskipun dengan sedikit peningkatan waktu respons total. Sistem yang dikembangkan menunjukkan keunggulan dibandingkan ChatGPT dengan memberikan jawaban yang lebih lokal dan langsung pada inti permasalahan untuk pertanyaan berbasis pengetahuan spesifik dengan sedikit trade-off pada evaluasi oleh pakar. Kontribusi penelitian ini adalah memberikan wawasan tentang integrasi RAG dan LLM untuk kesehatan ternak serta mengoptimalkan RAG melalui pengurangan duplikasi dokumen yang diambil.

Kata Kunci: *Chatbot*, GPT-4, *Large Language Model*, *Retrieval Augmented Generation*, SemDedup

ABSTRACT

Livestock sector faces significant challenges from animal diseases, the handling of which is often hampered by limited human resources. This research proposes the development of a chatbot based on Large Language Models (LLM) with a Retrieval Augmented Generation (RAG) approach as a solution. However, the implementation of the RAG-Fusion model often encounters document redundancy issues that can reduce efficiency and accuracy. The objective of this study is to optimize RAG-Fusion by reducing document redundancy through more stringent selection and ranking, thereby providing a faster and more accurate consultation solution for animal health.

This study uses a RAG-Fusion system development method focused on handling retrieval document redundancy with the Semantic Deduplication (SemDeDup) technique. Data was obtained from the Temanggung Regency Food Security, Agriculture, and Fisheries Agency (DKPPP) and the Deepseek V3 generative LLM, which were then used as an external knowledge base. The development process included generating alternative queries, searching for relevant chunks, reranking using Reciprocal Rank Fusion, and applying SemDeDup with k-means clustering and cosine similarity calculation to eliminate semantic duplication. The final response was generated by the GPT-4 model based on the deduplicated context.

The implementation of SemDeDup in the RAG-Fusion system was proven effective in improving context efficiency and reducing the computational load on the LLM model, despite a slight increase in total response time. The developed system showed superiority compared to ChatGPT by providing more localized and direct answers to specific knowledge-based questions but have small trade-off on evaluation by veterinary expert. The contribution of this research is to provide insight into the integration of RAG and LLM for animal health and to optimize RAG by reducing the duplication of retrieved documents.

Keywords: *Chatbot, GPT-4, Large Language Model, Retrieval Augmented Generation, SemDedup*