BAB V
KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan hasil perancangan, implementasi, dan pengujian fungsional sistem serta pengujian luaran model LLM, kesimpulan penelitian dapat dirumuskan sebagai berikut:

1. Penelitian ini berhasil mengembangkan sistem pembuatan soal otomatis dengan integrasi model LLM. Proses pembangunan sistem dimulai dari tahap analisis kebutuhan, desain sistem, dilanjutkan implementasi sistem berbasis *website* yang diintegrasikan pada LLM dengan teknik *semantic chunking* dan *semantic reranking*, kemudian diakhiri dengan pengujian sistem untuk menguji fungsionalitas serta *output* model LLM. Sisi antarmuka pengguna berhasil dikembangkan menggunakan React.js versi 19 dan Tailwind CSS versi 4. Sementara *backend* dengan intergrasi model LLM dan layanan kecerdasan buatan (Gemini API) sebagai model generatif berhasil dikembangkan menggunakan Python (FastAPI).
2. Implementasi teknik *semantic chunking* dilakukan pada tahap prapemrosesan dengan memecah teks berdasarkan kesamaan makna untuk menjaga keutuhan konteks. Sementara itu, teknik *semantic reranking* diintegrasikan pada proses *retrieval* untuk melakukan penyaringan ulang terhadap potongan teks yang paling relevan sebelum diproses oleh LLM. Implementasi kedua teknik ini terbukti mampu menghasilkan luaran soal yang memiliki kualitas kontekstual lebih tinggi dan relevan secara semantik.
3. Pengujian fungsional sistem dilakukan menggunakan metode *blackbox testing*, sedangkan pengujian perbandingan *output* RAG dan model LLM dilakukan menggunakan *framework* Deepeval. Evaluasi difokuskan pada dua aspek, yaitu pada proses *retrieval* dengan mengukur *contextual precision* dan *contextual recall* serta kualitas hasil soal dengan mengukur *answer relevancy*, *faithfulness*, dan *answer*

correctness.

4. Optimalisasi menggunakan teknik *semantic chunking* dan *semantic reranking* terbukti mampu meningkatkan kualitas *output* RAG dan model LLM. Hasil pengujian luaran model tanpa optimalisasi menggunakan kerangka kerja Deepeval menunjukkan skor rata-rata 0,604 untuk metrik *contextual precision*, 0,837 untuk *contextual recall*, 0,883 untuk *answer relevancy*, 0,716 untuk *faithfulness*, dan 0,777 untuk *answer correctness*. Skor tersebut berhasil ditingkatkan dengan implementasi teknik optimalisasi *semantic chunking* dan *semantic reranking* yang menunjukkan skor rata-rata 0,787 untuk metrik *contextual precision*, 0,850 untuk *contextual recall*, 0,961 untuk *answer relevancy*, 0,855 untuk *faithfulness*, dan 0,944 untuk *answer correctness*. Secara fungsional, pengujian *blackbox testing* menunjukkan bahwa keseluruhan fungsi dan fitur pada sistem AQG berjalan sesuai kebutuhan tanpa ditemukan kegagalan fungsi, sehingga sistem dapat dinyatakan memenuhi kebutuhan dari sisi fungsionalitas.

5.2. Saran

Terdapat beberapa saran yang dapat dijadikan acuan untuk pengembangan sistem maupun penelitian lanjutan:

1. Perlu dilakukan eksplorasi lebih lanjut pada strategi *reranking*. Disarankan untuk menguji atau membandingkan model *cross-encoder* lain untuk proses *reranking* dengan tujuan mendapatkan model yang paling sesuai untuk sistem AQG.
2. Penelitian selanjutnya disarankan untuk mengeksplorasi ambang batas *(threshold)* pemecahan kalimat yang lebih variatif untuk menemukan keseimbangan optimal antara keutuhan konteks dan efisiensi pemrosesan pada materi dengan karakteristik teks yang berbeda untuk menguji konsistensi performa *semantic chunking*.
3. Penelitian selanjutnya disarankan untuk memperluas cakupan sistem agar mampu memproses materi multimodal seperti gambar, diagram, atau tabel, sehingga sistem AQG dapat menghasilkan soal yang lebih variatif.