BAB 5. PENUTUP

5.1. Kesimpulan

Berdasarkan rangkaian kegiatan perancangan, implementasi, pengujian, dan evaluasi HerbaAI sebagai *chatbot* konsultasi serta rekomendasi produk herbal HNI berbasis *Large Language Model* (LLM) dengan pendekatan *Retrieval-Augmented Generation* (RAG), kesimpulan penelitian ini adalah sebagai berikut.

1. HerbaAI telah berhasil dirancang dan dikembangkan sebagai *web chatbot* herbal yang mengintegrasikan model LLaMA 3.1 8B Instant dengan pendekatan RAG. *Pipeline* RAG yang diterapkan terdiri atas tahapan *indexing*, *query classification*, *retrieval*, *generation*, dan *session memory*. ChromaDB digunakan sebagai basis data vektor, SQLite digunakan untuk menyimpan riwayat percakapan, sedangkan *localStorage* digunakan untuk mempertahankan identitas sesi dan tampilan percakapan pada peramban. Berdasarkan hasil pengujian *black box* secara manual dan otomatis, fungsi utama sistem, seperti konsultasi, penyajian sumber informasi, pengelolaan sesi, *streaming* jawaban, responsivitas antarmuka, penanganan kondisi kesalahan, serta pembatasan pertanyaan di luar domain, telah berjalan sesuai dengan rancangan yang ditetapkan.
2. Sistem yang menggunakan pendekatan RAG menghasilkan kesesuaian tekstual terhadap jawaban referensi yang lebih tinggi dibandingkan sistem LLM tanpa RAG. Evaluasi terhadap 20 pertanyaan dalam *golden dataset* menunjukkan bahwa sistem non-RAG memperoleh nilai ROUGE-1 sebesar 0,2065, ROUGE-2 sebesar 0,0527, dan ROUGE-L sebesar 0,1655. Sementara itu, sistem RAG memperoleh nilai ROUGE-1 sebesar 0,4998, ROUGE-2 sebesar 0,3533, dan ROUGE-L sebesar 0,4437. Dengan demikian, penerapan RAG memberikan peningkatan absolut sebesar 0,2933 pada ROUGE-1, 0,3006 pada ROUGE-2, dan 0,2782 pada ROUGE-L. Temuan tersebut memperlihatkan bahwa penggunaan konteks yang diperoleh dari basis pengetahuan membantu sistem menghasilkan jawaban yang lebih mendekati jawaban referensi. Meskipun demikian, nilai ROUGE hanya menggambarkan tingkat kesamaan tekstual dan tidak dapat digunakan secara langsung untuk menilai kebenaran faktual jawaban.
3. Hasil evaluasi menggunakan RAGAS menunjukkan bahwa HerbaAI dapat menghasilkan jawaban yang relevan dan sebagian besar telah didukung oleh konteks hasil *retrieval*. Sistem memperoleh skor *faithfulness* sebesar 0,8440, *answer relevancy* sebesar 0,9518, *context precision* sebesar 0,9211, dan *context recall* sebesar 0,7504. Nilai tersebut menunjukkan bahwa jawaban yang dihasilkan umumnya sesuai dengan pertanyaan dan konteks yang diambil memiliki tingkat ketepatan yang tinggi. Namun, kelengkapan konteks masih perlu ditingkatkan, khususnya untuk pertanyaan yang jawabannya tersebar pada beberapa bagian atau dokumen yang berbeda.

5.2. Saran

Berdasarkan hasil penelitian dan evaluasi yang telah dilakukan, beberapa saran untuk pengembangan HerbaAI pada penelitian selanjutnya adalah sebagai berikut.

1. Mekanisme *retrieval* perlu disempurnakan untuk meningkatkan kelengkapan konteks karena nilai *context recall* masih lebih rendah dibandingkan metrik RAGAS lainnya. Pengembangan dapat dilakukan dengan menerapkan *structure-aware chunking*, menambahkan metadata yang lebih terperinci, menggunakan *reranker*, serta menentukan jumlah konteks secara dinamis sesuai dengan karakteristik pertanyaan. Setiap pendekatan tersebut perlu diuji menggunakan metrik *context precision* dan *context recall* agar peningkatan cakupan konteks tetap mempertahankan ketepatan hasil *retrieval*.

2. Basis pengetahuan perlu diperbarui dan diverifikasi secara berkala agar informasi mengenai produk, aturan penggunaan, kontraindikasi, dan resep herba tetap sesuai dengan sumber HNI terbaru. Dokumen tambahan sebaiknya diperoleh dari sumber resmi atau telah melalui validasi oleh pihak yang memiliki kompetensi dalam bidang herbal maupun kesehatan. Selain itu, pengelolaan versi dokumen dan pencatatan sumber perlu diterapkan agar setiap perubahan informasi dapat dilacak secara jelas.

3. Evaluasi teknis pada penelitian berikutnya perlu menggunakan *golden dataset* dengan jumlah pertanyaan yang lebih besar dan variasi yang lebih beragam. Dataset dapat mencakup pertanyaan dengan bahasa sehari-hari, kesalahan penulisan, makna ambigu, lebih dari satu maksud atau *multi-intent*, percakapan lanjutan, serta pertanyaan yang membutuhkan penggabungan informasi dari beberapa dokumen.

4. Evaluasi pengguna perlu melibatkan jumlah partisipan dan kelompok pengguna yang lebih luas, seperti mentor HNI, konsumen, calon konsumen, serta pengguna yang belum terbiasa berinteraksi dengan *chatbot*. Aspek yang dapat dinilai meliputi kemudahan penggunaan, keterbacaan jawaban, kejelasan sumber informasi, tingkat kepercayaan pengguna, pemahaman terhadap batasan medis, dan manfaat HerbaAI dalam membantu proses pencarian informasi mengenai herbal.

5. Pengujian terhadap keamanan dan pembatasan respons perlu diperluas dengan menambahkan variasi pertanyaan di luar domain serta serangan *prompt injection*. Skenario pengujian dapat mencakup instruksi bertingkat, percakapan *multi-turn* yang berupaya mengubah peran sistem, permintaan untuk menampilkan instruksi internal, dan pertanyaan medis dengan tingkat risiko tinggi. Hasil pengujian tersebut dapat menjadi dasar untuk memperkuat rancangan *prompt*, mekanisme klasifikasi pertanyaan, dan respons *fallback* agar sistem tetap memberikan jawaban sesuai dengan ruang lingkup layanan yang telah ditetapkan.