BAB V
PENUTUP

Bab ini menyajikan kesimpulan yang diperoleh dari seluruh rangkaian penelitian serta saran untuk pengembangan lebih lanjut.

5.1. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan mengenai analisis performa *reasoning models* pada *QA system* berbasis *retrieval-augmented generation* (RAG) sebagai layanan informasi PPMB UPN "Veteran" Jawa Timur, dapat ditarik kesimpulan sebagai berikut:

1. **Perbandingan Performa Antarkategori Model**

   Reasoning models menunjukkan keunggulan performa yang signifikan dibandingkan *non-reasoning models* pada *QA system* berbasis RAG sebagai layanan informasi PPMB UPN "Veteran" Jawa Timur. Secara agregat, *reasoning models* memperoleh rata-rata skor RAGAS sebesar 0.7772, sementara *non-reasoning models* memperoleh skor 0.7289, menghasilkan peningkatan performa sebesar 6.63%. Keunggulan paling menonjol terlihat pada metrik *factual correctness* dengan selisih 0.0781 atau peningkatan performa sebesar 15.95%, yang mengindikasikan bahwa kemampuan penalaran *reasoning models* berkontribusi signifikan dalam meningkatkan akurasi faktual jawaban.

   Pada aspek performa multibahasa, *reasoning models* menunjukkan konsistensi yang lebih baik dengan keunggulan di ketiga bahasa yang diuji. Bahasa Indonesia dengan *gap* 5.65%, Bahasa Inggris dengan *gap* 6.31%, dan Bahasa Jawa(Suroboyoan) dengan *gap* 7.99%. Temuan menarik adalah keunggulan *reasoning models* semakin menonjol ketika menghadapi kompleksitas linguistik yang lebih tinggi, seperti pada Bahasa Jawa yang merupakan *low-resource language* dengan karakteristik dialek yang unik.

Analisis kualitatif mengungkapkan bahwa *reasoning models* unggul terutama pada *query* kompleks yang memerlukan sintesis data dari berbagai sumber dokumen. Pada *query* sederhana yang bersifat *simple fact retrieval*, kedua kategori model menunjukkan performa yang setara. Namun, performa kedua kategori model sama-sama terbatas ketika sistem *retrieval* gagal mengambil dokumen yang relevan, mengindikasikan bahwa kualitas *retriever* tetap menjadi faktor pembatas fundamental dalam arsitektur RAG.

2. Model dengan Performa Paling Optimal

Model gemini-2.5-flash dari kategori *reasoning models* menunjukkan performa paling optimal untuk diimplementasikan pada *QA system* berbasis RAG sebagai layanan informasi PPMB UPN "Veteran" Jawa Timur. Model ini unggul dalam beberapa aspek:

a. **Performa Tertinggi:** Memperoleh rata-rata skor RAGAS tertinggi sebesar 0.8207, dengan skor *faithfulness* 0.9195, *factual correctness* 0.6274, dan *semantic similarity* 0.9151.
b. **Konsistensi Multibahasa:** Menunjukkan performa superior secara konsisten di ketiga bahasa (Inggris: 0.8387, Indonesia: 0.8255, Jawa: 0.7978) dengan *robustness score* tertinggi yang mengombinasikan performa absolut dengan konsistensi multibahasa.
c. **Efisiensi Biaya:** Meskipun bukan model dengan biaya terendah, gemini-2.5-flash menawarkan *cost per performance point* yang rasional sebesar $0.008777 dengan proyeksi biaya operasional bulanan sekitar $2,158.50 untuk skenario *deployment* skala besar (10,000 query per hari).

Sebagai alternatif untuk *deployment* dengan prioritas efisiensi biaya, model gemini-2.0-flash (*non-reasoning models*) dapat dipertimbangkan dengan biaya operasional sangat rendah ($387.30 per bulan untuk 10,000 query per hari) namun dengan *trade-off* performa yang lebih rendah (rata-rata skor 0.7252).

5.2. Saran

Berdasarkan hasil penelitian dan keterbatasan yang ditemui selama proses penelitian, berikut adalah saran untuk pengembangan lebih lanjut:

1. **Optimasi Sistem Retrieval**

   Penelitian ini menggunakan variabel terkontrol pada parameter *retrieval* untuk memastikan perbandingan performa murni merefleksikan kapabilitas *generation model*. Penelitian selanjutnya disarankan untuk mengeksplorasi optimasi pada komponen *retrieval*, seperti eksperimen dengan berbagai *chunking strategies*, *hybrid retrieval* yang menggabungkan *sparse* dan *dense retrieval*, serta teknik *reranking* untuk meningkatkan kualitas konteks yang diberikan kepada *generator*.

2. **Pengembangan Query Classification System**

   Berdasarkan temuan analisis kualitatif yang menunjukkan bahwa *reasoning models* unggul pada *query* kompleks sementara kedua kategori model setara pada *query* sederhana, penelitian selanjutnya dapat mengembangkan sistem klasifikasi *query* otomatis yang dapat merutekan *query* kompleks ke *reasoning models* dan *query* sederhana ke *non-reasoning models* yang lebih efisien biaya, sehingga mencapai keseimbangan optimal antara performa dan efisiensi.

3. **Ekspansi Cakupan Bahasa**

   Penelitian ini menguji tiga bahasa (Indonesia, Inggris, dan Jawa/Suroboyoan). Mengingat keragaman bahasa daerah di Indonesia, penelitian selanjutnya dapat memperluas cakupan bahasa untuk menguji *robustness* model pada bahasa daerah lainnya yang relevan dengan konteks calon mahasiswa baru dari berbagai wilayah.

4. **Pengembangan Fitur Prototype**

   Prototype yang telah dikembangkan dalam penelitian ini bersifat demonstratif. Untuk implementasi produksi, disarankan pengembangan fitur tambahan seperti *conversation history management*, *feedback mechanism* untuk *continuous improvement*, *monitoring* dan *analytics dashboard* untuk memantau performa sistem secara *real-time*, penyematan *widget chat* pada halaman utama *website* PPMB UPN "Veteran" Jatim agar dapat diakses langsung oleh publik.

5. **Studi Longitudinal**

   Mengingat perkembangan pesat dalam bidang *large language models*, penelitian longitudinal yang membandingkan performa model-model baru yang dirilis secara berkala dapat memberikan *insight* berharga tentang tren perkembangan kemampuan *reasoning* dan implikasinya terhadap aplikasi RAG.