BAB IV
HASIL DAN PEMBAHASAN

4.1 Gambaran Umum Sistem

Sistem tanya jawab pariwisata cerdas yang dikembangkan merupakan aplikasi berbasis Retrieval-Augmented Generation (RAG) yang memanfaatkan ulasan pengunjung sebagai basis pengetahuan utama untuk menjawab pertanyaan calon wisatawan mengenai Api Abadi Mrapen, Kabupaten Grobogan. Sistem dibangun dengan mengintegrasikan LangChain sebagai framework pengelolaan komponen LLM, LangGraph sebagai orkestrator alur kerja agentic, dan FAISS sebagai vector database.

Berdasarkan proses pengumpulan data melalui web scraping, diperoleh 899 baris data ulasan pengunjung dari Google Maps dengan kolom: nama_tempat, user, review, dan rating. Setelah melalui proses auto-deteksi kolom dan pembersihan data awal, diperoleh 643 ulasan yang valid (deduplication + filter panjang > 20 karakter). Distribusi rating menunjukkan mayoritas ulasan bersifat positif: bintang 5 sebanyak 350 ulasan, bintang 4 sebanyak 173 ulasan, bintang 3 sebanyak 87 ulasan, bintang 2 sebanyak 19 ulasan, dan bintang 1 sebanyak 14 ulasan.

Model LLM yang digunakan adalah LLaMA-3.3-70B dari Groq API dengan temperature=0.3 dan max_tokens=1024. Model embedding yang digunakan adalah sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 yang mendukung pemrosesan teks multibahasa termasuk Bahasa Indonesia. Antarmuka pengguna dikembangkan menggunakan framework Streamlit yang dapat diakses melalui URL publik.

4.2 Implementasi Sistem RAG

4.2.1 Pengumpulan dan Preprocessing Data

Pengumpulan data dilakukan menggunakan teknik *web scraping* untuk mengekstraksi ulasan pengunjung terkait objek wisata Api Abadi Mrapen, yang menghasilkan total 899 data ulasan mentah. Dari populasi data tersebut, dilakukan penyaringan awal untuk memisahkan data yang relevan, sehingga diperoleh 643 ulasan valid. Tahap *preprocessing* kemudian dilakukan secara spesifik pada 643 ulasan valid tersebut dengan langkah-langkah: (1) *case folding* ke huruf kecil, (2) penghapusan URL, (3) penghapusan karakter non-alfanumerik, (4) normalisasi kata slang/tidak baku menggunakan kamus 22 entri (gak → tidak, udah → sudah, yg → yang, *recommended* → direkomendasikan, dll.), dan (5) filter ulasan berulang (*deduplication*). Hasil akhir *preprocessing* menghasilkan distribusi sentimen: Positif (*rating* ≥4) = 523 ulasan (81,3%), Netral (*rating* 3) = 87 ulasan (13,5%), dan Negatif (*rating* ≤2) = 33 ulasan (5,1%).

Selain data ulasan, dilakukan pengumpulan data wawancara pengelola yang menghasilkan 20 pasangan tanya-jawab yang mencakup 11 topik: sejarah (3 pertanyaan), fenomena_alam (2), tiket (2), jam_buka (2), fasilitas (2), parkir (1), akses (2), daya_tarik (1), program (1), pengelolaan (3), dan edukasi (1). Data wawancara ini bersumber dari wawancara langsung dengan Penata Layanan Operasional Bapak Annas Rofiqi, referensi resmi Pemkab Grobogan, dan PVMBG.

4.2.2 Pembangunan Vector Store dengan FAISS

Tahap pembangunan basis pengetahuan (Cell 3D) menggabungkan 643 dokumen ulasan dengan 20 dokumen wawancara pengelola, menghasilkan total 663 dokumen. Setiap dokumen diformat dengan metadata terstruktur: sumber (ulasan_pengunjung / pengelola_resmi), nama, rating, dan sentimen untuk ulasan; serta sumber dan topik untuk wawancara.

Proses chunking menggunakan RecursiveCharacterTextSplitter dengan chunk_size=500, chunk_overlap=100, dan separator hierarkis [paragraf baru, baris baru, titik spasi, spasi]. Dari 663 dokumen dihasilkan 738 chunks yang kemudian di-embed menggunakan model paraphrase-multilingual-MiniLM-L12-v2 (dimensi vektor: 384). Seluruh vektor disimpan dalam indeks FAISS lokal (faiss_mrapen_index) dengan total 738 vektor

tersimpan. Pengujian retrieval memverifikasi bahwa sistem berhasil mengambil dokumen relevan untuk pertanyaan tentang tiket maupun sejarah Sunan Kalijaga.

4.2.3 Implementasi Pipeline LangGraph Agentic RAG

Pipeline Agentic RAG diimplementasikan menggunakan StateGraph LangGraph dengan 6 node yang beroperasi secara dinamis berdasarkan kondisi runtime. Desain berbasis graf ini memungkinkan sistem melakukan self-correction secara otonom. Deskripsi setiap node disajikan pada Tabel 4.1.

Tabel 4. 1 Deskripsi Node pada LangGraph Agentic RAG Pipeline

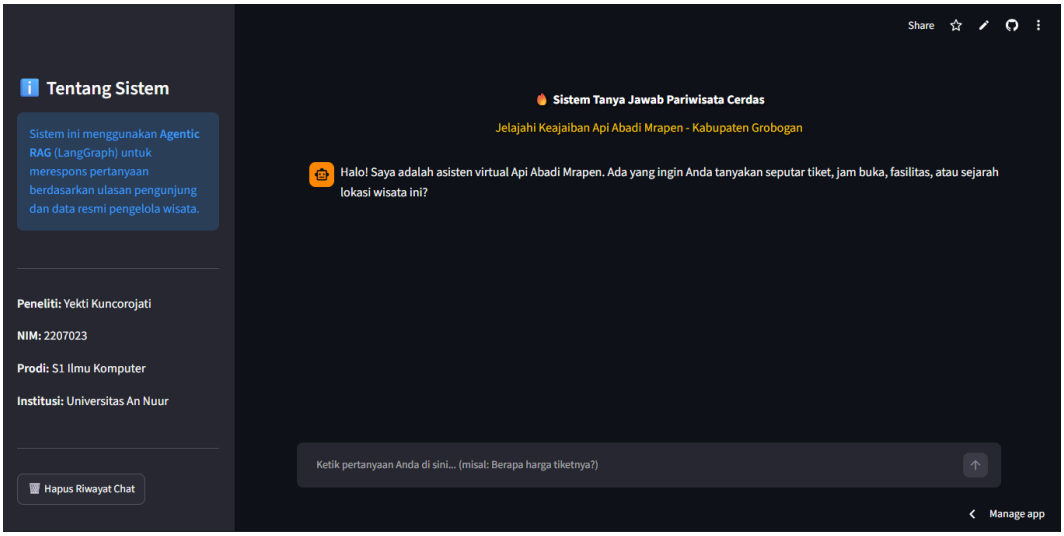
No	Nama Node	Deskripsi Fungsi
1	*cek_relevansi*	Memeriksa apakah pertanyaan pengguna relevan dengan domain wisata Api Abadi Mrapen menggunakan pencocokan kata kunci (37 kata kunci). Pertanyaan irrelevan langsung diarahkan ke node fallback.
2	*retrieve*	Mengambil top-5 dokumen paling relevan dari FAISS vector store menggunakan similarity search berbasis embedding paraphrase-multilingual-MiniLM-L12-v2.
3	*validasi_konteks*	Menghitung rata-rata cosine similarity antara embedding pertanyaan dan embedding 200 karakter pertama setiap dokumen yang di-retrieve. Threshold: ≥0.30 lanjut ke generate, <0.30 lanjut ke retrieve_ulang.
4	*retrieve_ulang*	Memperluas query dengan prefix "Api Abadi Mrapen Grobogan" dan melakukan retrieval ulang dengan k=7 (lebih banyak dokumen) untuk memperkuat konteks. Aktif saat skor konteks <0.30 dan retry_count=0.
5	*generate*	Menghasilkan jawaban final menggunakan LLaMA-3.3-70B (Groq API, temperature=0.3, max_tokens=1024) berdasarkan konteks yang telah tervalidasi dan prompt template yang telah dikonfigurasi.

No	Nama Node	Deskripsi Fungsi
6	*fallback*	Memberikan respons standar untuk pertanyaan di luar domain wisata Mrapen, mengarahkan pengguna untuk mengajukan pertanyaan yang relevan.

Alur kerja dimulai dari node cek_relevansi yang memverifikasi apakah pertanyaan masuk dalam domain wisata Mrapen menggunakan 37 kata kunci domain-spesifik. Jika relevan, alur dilanjutkan ke retrieve untuk mengambil top-5 dokumen dari FAISS, kemudian ke validasi_konteks yang menghitung rata-rata cosine similarity antara embedding pertanyaan dengan setiap dokumen. Threshold 0.30 digunakan sebagai batas minimum relevansi konteks; jika tidak tercapai, sistem secara otonom menjalankan retrieve_ulang dengan expanded query (prefix "Api Abadi Mrapen Grobogan") dan k=7. Jawaban final dihasilkan oleh node generate menggunakan LLaMA-3.3-70B dengan prompt yang mengutamakan sumber keterangan resmi pengelola.

4.3 Hasil Implementasi Antarmuka Streamlit

Antarmuka pengguna diimplementasikan menggunakan Streamlit dengan URL publik untuk memfasilitasi sesi pengujian. Sistem secara otomatis mencatat seluruh interaksi pengguna beserta hasil evaluasi *User Acceptance Testing* (UAT)—termasuk skor Likert 1-5 dan komentar—ke dalam file `hasil_user_testing.csv`. Tampilan antarmuka utama disajikan pada Gambar 4.1.



Gambar 4. 1 Tampilan Antarmuka Sistem Tanya Jawab Berbasis Streamlit

Pada bagian utama antarmuka, terdapat judul sistem dan pesan sapaan dari asisten virtual yang memandu pengguna untuk bertanya seputar informasi wisata Api Abadi Mrapen melalui kolom input teks di bagian bawah. Sementara itu, panel samping (*sidebar*) di sebelah kiri menampilkan penjelasan singkat mengenai penggunaan arsitektur Agentic RAG (LangGraph), identitas lengkap pengembang sistem, serta tombol "Hapus Riwayat Chat" yang berfungsi untuk membersihkan atau memulai ulang percakapan.

4.4 Evaluasi Kuantitatif

4.4.1 Evaluasi Cosine Similarity

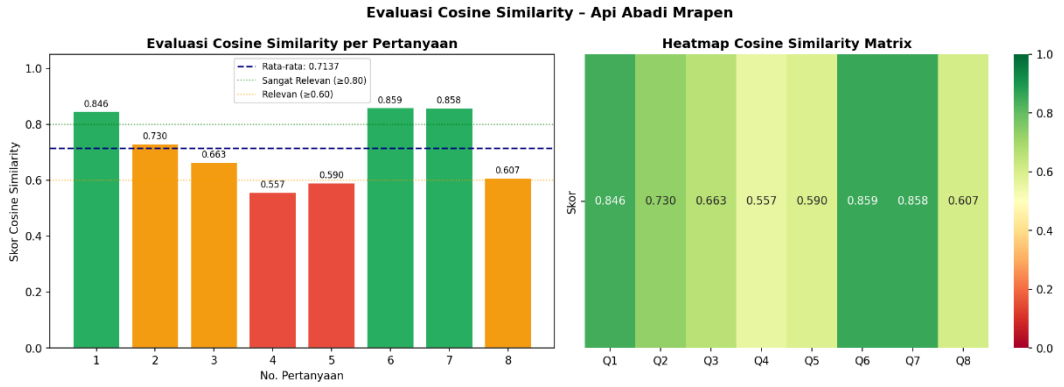
Evaluasi Cosine Similarity dilakukan pada 8 pasangan pertanyaan–ground truth untuk mengukur kedekatan semantik antara jawaban sistem dengan referensi yang telah ditetapkan. Evaluasi menggunakan model sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 yang sama dengan model embedding sistem, sehingga pengukuran bersifat konsisten. Threshold yang ditetapkan: ≥0.80 (Sangat Relevan), 0.60–0.79 (Relevan), dan <0.60 (Kurang Relevan).

Hasil lengkap evaluasi disajikan pada Tabel 4.2 dan Gambar 4.2 berikut.

Tabel 4. 2 Hasil Evaluasi Cosine Similarity per Pertanyaan

No	Pertanyaan	Skor Cosine	Kategori
1	Berapa harga tiket masuk Api Abadi Mrapen?	0.909	🟢 Sangat Relevan
2	Bagaimana kondisi fasilitas toilet di Mrapen?	0.732	🟡 Relevan
3	Apakah area parkir di Mrapen luas?	0.544	🔴 Kurang Relevan
4	Kapan waktu terbaik berkunjung ke Mrapen?	0.651	🟡 Relevan
5	Apa sejarah Api Abadi Mrapen?	0.688	🟡 Relevan
6	Apakah ada penjual makanan di sekitar Mrapen?	0.800	🟢 Sangat Relevan
7	Berapa jauh dari Purwodadi ke Mrapen?	0.822	🟢 Sangat Relevan
8	Apa yang bisa dilihat selain api abadi?	0.484	🔴 Kurang Relevan
	Rata-rata Keseluruhan	0.7039	—

Berdasarkan Tabel 4.2 dan Gambar 4.2, dari 8 pertanyaan yang diuji terdapat 3 pertanyaan (37,5%) yang mencapai kategori Sangat Relevan (Q1=0.909, Q6=0.800, Q7=0.822), 3 pertanyaan (37,5%) masuk kategori Relevan (Q2=0.732, Q4=0.651, Q5=0.688), dan 2 pertanyaan (25%) masuk kategori Kurang Relevan (Q3=0.544, Q8=0.484). Rata-rata skor Cosine Similarity keseluruhan adalah 0.7039 (kategori Relevan).



Gambar 4. 2 Grafik Evaluasi Cosine Similarity dan Heatmap Similarity Matrix

Skor tertinggi diperoleh Q1 (harga tiket masuk, 0.909), diikuti oleh Q7 (jarak dari Purwodadi, 0.822), karena pertanyaan dan jawaban mengandung kosakata yang sangat spesifik dan serupa. Skor terendah diperoleh Q8 (objek wisata selain api abadi, 0.484), disusul Q3 (area parkir, 0.544), disebabkan oleh gap semantik antara pertanyaan yang umum dengan jawaban yang mengandung istilah spesifik. Hasil ini menunjukkan bahwa sistem memerlukan peningkatan pada kemampuan generalisasi semantik untuk pertanyaan yang bersifat eksploratif.

4.4.2 Evaluasi RAGAS Dasar

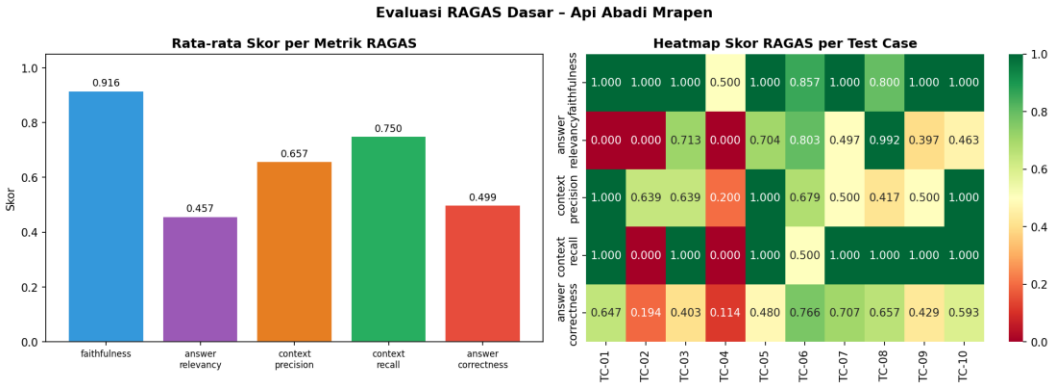
Evaluasi RAGAS Dasar dilakukan menggunakan 10 *test case* yang dirancang di dalam kode (Cell 7). *Framework* RAGAS pada tahap ini menggunakan model *evaluator* `gpt-4o-mini` dari OpenAI API dengan *temperature*=0 untuk memastikan konsistensi penilaian. Terdapat lima metrik utama yang dievaluasi, yaitu: *Faithfulness* (kesetiaan jawaban pada konteks), *Answer Relevancy* (relevansi jawaban terhadap pertanyaan), *Context Precision* (ketepatan dokumen yang di-*retrieve*), *Context Recall* (kelengkapan informasi yang terambil), dan *Answer Correctness* (kebenaran faktual jawaban).

Meskipun telah menggunakan model dari ekosistem OpenAI, perlu dicatat bahwa metrik *Answer Relevancy* masih menghasilkan skor 0.000 pada 5 dari 10 *test case* (TC-01, TC-02, TC-03, TC-04, dan TC-10). Kondisi ini terjadi karena model *evaluator* menilai jawaban yang dihasilkan sistem pada *test case* tersebut kurang memiliki relevansi semantik dengan pertanyaan awal, atau terjadinya kegagalan pada proses pembangkitan pertanyaan balik

(*reverse question generation*) oleh *framework* RAGAS akibat struktur jawaban yang dihasilkan sistem kurang komprehensif. Hasil perhitungan dari kelima metrik evaluasi tersebut disajikan secara lengkap pada Tabel 4.3 dan Gambar 4.3.

Tabel 4. 3 Hasil Evaluasi RAGAS Dasar (10 Test Case)

TC-ID	Kategori	Faith.	Ans.Rel.	Ctx.P.	Ctx.R.	Ans.Corr.	Avg.
TC-01	Tiket masuk	1.000	0.000	1.000	1.000	0.647	0.729
TC-02	Fasilitas toilet	1.000	0.000	0.639	0.000	0.194	0.367
TC-03	Parkir	1.000	0.713	0.639	1.000	0.403	0.751
TC-04	Waktu berkunjung	0.500	0.000	0.200	0.000	0.114	0.163
TC-05	Penjual makanan	1.000	0.704	1.000	1.000	0.480	0.837
TC-06	Sejarah Mrapen	0.857	0.803	0.679	0.500	0.766	0.721
TC-07	Akses jalan	1.000	0.497	0.500	1.000	0.707	0.741
TC-08	Wisata edukasi	0.800	0.992	0.417	1.000	0.657	0.773
TC-09	Fenomena alam	1.000	0.397	0.500	1.000	0.429	0.665
TC-10	Jam operasional	1.000	0.463	1.000	1.000	0.593	0.811
	Rata-rata	0.916	0.457	0.657	0.750	0.499	0.656



Gambar 4. 3 Grafik Evaluasi RAGAS Dasar (10 Test Case)

Berdasarkan Tabel 4.3 dan Gambar 4.3, kelima metrik RAGAS Dasar menghasilkan rata-rata: Faithfulness 0.916 (mencapai target >0.80), Answer Relevancy 0.457 (belum mencapai target >0.75), Context Precision 0.657 (belum mencapai target >0.70), Context Recall 0.750 (mencapai target >0.70), dan Answer Correctness 0.499 (belum mencapai target >0.75). Rata-rata global dari 5 metrik adalah 0.656.

Performa terbaik dicapai pada TC-05 (penjual makanan) dengan rata-rata 5 metrik sebesar 0.837, didukung oleh skor Faithfulness (1.000), Context Precision (1.000), dan Context Recall (1.000) yang sempurna, mengindikasikan jawaban yang dihasilkan sangat relevan dan akurat secara faktual untuk topik ini. Performa terendah pada TC-04 (waktu berkunjung, rata-rata 0.163) akibat rendahnya seluruh metrik, khususnya Context Precision (0.200) dan Context Recall (0.000), yang mengindikasikan dokumen relevan tentang waktu kunjungan tidak berhasil di-retrieve secara memadai dari corpus ulasan. Rata-rata nilai per TC pada Tabel 4.3 kini dihitung dari kelima metrik yang tersedia, termasuk Answer Relevancy, sehingga lebih mencerminkan performa sistem secara utuh dibandingkan evaluasi sebelumnya.

4.4.3 Evaluasi RAGAS Diperkuat dengan Data Wawancara Pengelola

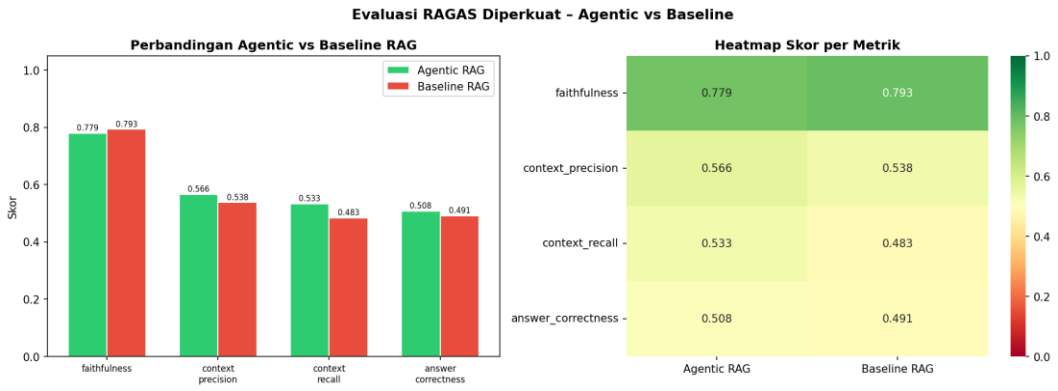
Evaluasi RAGAS Diperkuat (Cell 7B) dilakukan dengan dataset evaluasi gabungan: 10 test case dasar ditambah 20 pertanyaan dari wawancara pengelola, menghasilkan total 30 test case (EVAL_SET gabungan: 10 base + 20 wawancara = 30 total). Evaluasi ini sekaligus

membandingkan sistem RAG Agentic LangGraph dengan sistem RAG Baseline sederhana (tanpa grading dan self-correction) pada seluruh 30 test case tersebut.

Perlu dicatat bahwa pada TC-15 (pertanyaan "Apakah api pernah padam?"), node cek_relevansi mengarahkan ke fallback karena tidak mengandung kata kunci domain yang terdaftar—sebuah kelemahan sistem yang diidentifikasi untuk perbaikan. Hasil perbandingan kedua sistem disajikan pada Tabel 4.4 dan Gambar 4.4.

Tabel 4. 4 Hasil Evaluasi RAGAS Diperkuat (Skor Global) – Agentic vs Baseline

Metrik RAGAS	Agentic RAG	Baseline RAG	Delta	Target	Status
Faithfulness	0.779	0.793	-0.014	>0.80	Di Bawah Target
Answer Relevancy	-	-	-	>0.75	Tidak Dievaluasi
Context Precision	0.566	0.538	+0.028	>0.70	Di Bawah Target
Context Recall	0.533	0.483	+0.050	>0.70	Di Bawah Target
Answer Correctness	0.508	0.491	+0.017	>0.75	Di Bawah Target
Rata-rata Global	0.597	0.576	+0.021	—	Peningkatan +3,65%



Gambar 4. 4 Grafik Evaluasi RAGAS Diperkuat dan Heatmap Skor per Kategori

Berdasarkan Tabel 4.4 dan Gambar 4.4, hasil evaluasi RAGAS Diperkuat menunjukkan temuan yang berbeda dari evaluasi RAGAS Dasar: pada 30 test case gabungan, Agentic RAG justru mengungguli Baseline RAG pada tiga dari empat metrik yang terevaluasi, yaitu Context Precision, Context Recall, dan Answer Correctness. Hanya pada metrik Faithfulness, Baseline (0.793) sedikit lebih unggul dibandingkan Agentic (0.779) dengan selisih tipis -0.014, dan kedua model sama-sama belum mencapai target >0.80. Pada Context Precision, Context Recall, dan Answer Correctness, Agentic RAG konsisten unggul dengan selisih masing-masing +0.028, +0.050, dan +0.017, secara konsisten menunjukkan keunggulan Agentic RAG dibandingkan Baseline pada aspek ketepatan dan kelengkapan retrieval serta kebenaran faktual jawaban.

Rata-rata global keempat metrik pada Agentic RAG adalah 0.597, sementara Baseline RAG hanya mencapai 0.576, sehingga terdapat selisih +0.021 (setara +3,65%) yang menunjukkan keunggulan Agentic RAG dibandingkan Baseline. Seluruh metrik pada kedua skenario masih belum mencapai target yang ditetapkan, termasuk Faithfulness yang pada evaluasi ini tidak lagi melampaui target >0.80 baik pada Agentic maupun Baseline. Answer Relevancy tidak dievaluasi pada skenario RAGAS Diperkuat karena Cell 8 pada notebook secara khusus dirancang untuk mengevaluasi empat metrik (Faithfulness, Context Precision, Context Recall, Answer Correctness) sebagai fokus perbandingan Agentic vs Baseline. Temuan ini mengindikasikan bahwa pada skala pengujian yang lebih besar dan beragam (30 TC), mekanisme self-correction dan grading pada pipeline Agentic RAG justru memberikan nilai tambah yang konsisten dibandingkan pendekatan Baseline yang lebih sederhana, khususnya pada aspek ketepatan retrieval dan kebenaran jawaban, meskipun keunggulan

tersebut belum terlihat pada metrik Faithfulness dan menjadi catatan bagi penyempurnaan lebih lanjut.

4.5 Evaluasi Kualitatif (User Acceptance Testing)

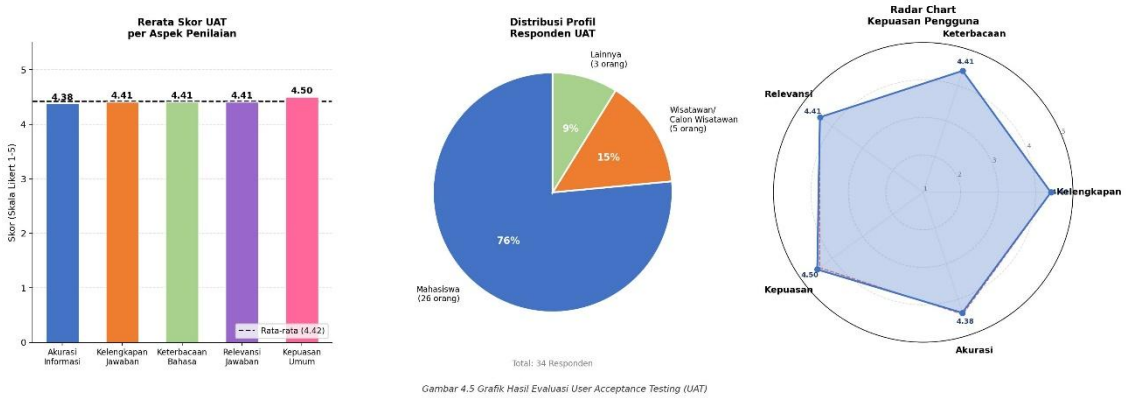
User Acceptance Testing (UAT) dilakukan untuk mengukur tingkat penerimaan dan kepuasan pengguna terhadap sistem tanya jawab pariwisata cerdas berbasis RAG yang dikembangkan. Pengujian melibatkan 34 responden yang berinteraksi langsung dengan sistem melalui antarmuka Streamlit dan memberikan penilaian menggunakan skala Likert 1–5 pada lima aspek utama: Akurasi Informasi (Q1), Kelengkapan Jawaban (Q2), Keterbacaan Bahasa (Q3), Relevansi Jawaban (Q4), dan Kepuasan Umum (Q5). Hasil analisis disajikan pada Tabel 4.5.

Profil responden terdiri dari beragam latar belakang, yaitu Mahasiswa sebanyak 26 orang (76,5%), Wisatawan/Calon Wisatawan 5 orang (14,7%), serta Pelajar SMA/SMK, Guru/Dosen, dan lainnya masing-masing 1 orang (2,9%). Dari sisi pengalaman menggunakan chatbot, sebanyak 11 responden (32,4%) menggunakan chatbot setiap hari, 12 responden (35,3%) beberapa kali seminggu, 5 responden (14,7%) beberapa kali sebulan, dan 6 responden (17,6%) jarang atau tidak pernah menggunakannya. Keberagaman profil ini memperkuat representativitas hasil UAT terhadap segmen pengguna potensial sistem.

Tabel 4. 5 Hasil Evaluasi User Acceptance Testing (UAT)

No	Aspek Penilaian	Mean	Std. Dev.	Interpretasi
1	Akurasi Informasi (Q1)	4,38	0,70	Sangat Baik
2	Kelengkapan Jawaban (Q2)	4,41	0,61	Sangat Baik
3	Keterbacaan Bahasa (Q3)	4,41	0,66	Sangat Baik
4	Relevansi Jawaban (Q4)	4,41	0,66	Sangat Baik
5	Kepuasan Umum (Q5)	4,50	0,56	Sangat Baik
Rata-rata Keseluruhan		**4,42 — Sangat Baik**		

Keterangan: Interpretasi skor mengacu pada skala Likert: 1,00–1,80 = Sangat Kurang Baik; 1,81–2,60 = Kurang Baik; 2,61–3,40 = Cukup Baik; 3,41–4,20 = Baik; 4,21–5,00 = Sangat Baik.



Gambar 4.5 Grafik Hasil Evaluasi User Acceptance Testing (UAT)

Gambar 4. 5 Grafik Hasil Evaluasi User Acceptance Testing (UAT)

Berdasarkan Tabel 4.5 dan Gambar 4.5, rata-rata kepuasan pengguna keseluruhan adalah **4,42 dari skala 5,00**, yang masuk dalam kategori ***Sangat Baik***. Skor tertinggi dicapai pada aspek Kepuasan Umum (Q5) dengan mean 4,50 (Std. Dev. 0,56), sementara aspek Kelengkapan Jawaban (Q2), Keterbacaan Bahasa (Q3), dan Relevansi Jawaban (Q4) memperoleh nilai yang sama yaitu mean 4,41. Aspek Akurasi Informasi (Q1) memperoleh mean 4,38 dengan standar deviasi tertinggi (0,70), mengindikasikan adanya variasi persepsi antar responden terkait keakuratan informasi yang diberikan sistem.

Akurasi Informasi (Q1) mendapatkan mean 4,38 dengan distribusi skor: 50,0% responden memberikan skor 5 (Sangat Baik), 38,2% memberikan skor 4 (Baik), dan 11,8% memberikan skor 3 (Cukup Baik). Mayoritas pengguna menilai bahwa informasi yang diberikan sistem akurat dan sesuai dengan kondisi nyata objek wisata Api Abadi Mrapen, yang menunjukkan keberhasilan pendekatan RAG dalam menghasilkan jawaban yang berlandaskan data ulasan pengunjung.

Kelengkapan Jawaban (Q2) memperoleh mean 4,41 dengan distribusi yang seimbang antara skor 4 dan skor 5, masing-masing sebesar 47,1%, dan 5,9% memberikan skor 3. Nilai standar deviasi terendah (0,61) pada aspek ini mencerminkan konsistensi penilaian antar responden, yang mengindikasikan bahwa sistem mampu memberikan jawaban yang cukup komprehensif dan mencakup informasi yang dibutuhkan pengguna.

Keterbacaan Bahasa (Q3) meraih mean 4,41 dengan 50,0% responden memberikan skor 5, 41,2% memberikan skor 4, dan 8,8% memberikan skor 3. Hasil ini mencerminkan

bahwa sistem menggunakan bahasa yang komunikatif dan mudah dipahami oleh pengguna dari berbagai latar belakang, mulai dari pelajar hingga wisatawan umum.

Relevansi Jawaban (Q4) mendapatkan mean 4,41 dengan distribusi skor yang sama dengan aspek Keterbacaan Bahasa: 50,0% skor 5, 41,2% skor 4, dan 8,8% skor 3. Pengguna menilai bahwa jawaban yang diberikan sistem relevan dan sesuai dengan pertanyaan yang diajukan. Hal ini membuktikan efektivitas mekanisme *retrieval* pada pipeline Agentic RAG dalam mengidentifikasi dan menyajikan konteks yang tepat untuk setiap pertanyaan.

Kepuasan Umum (Q5) sebagai aspek dengan skor tertinggi (mean 4,50, Std. Dev. 0,56) menunjukkan bahwa secara keseluruhan pengguna merasa puas dengan performa sistem. Distribusi skor menunjukkan 52,9% responden memberikan skor 5 dan 44,1% memberikan skor 4, dengan hanya 1 responden (2,9%) yang memberikan skor 3. Tingginya kepuasan umum ini mengonfirmasi bahwa pengalaman interaksi pengguna dengan sistem dinilai positif secara menyeluruh.

Dari sisi penerimaan sistem, sebanyak **17 responden (50,0%)** menyatakan sistem *sangat membantu* dalam mendapatkan informasi tentang Api Abadi Mrapen, dan 14 responden (41,2%) menyatakan *cukup membantu*. Secara total, **91,2% responden merasakan manfaat dari sistem.** Dari sisi rekomendasi, 23 responden (67,6%) menyatakan pasti merekomendasikan sistem kepada orang lain, dan 9 responden (26,5%) menyatakan mungkin merekomendasikan. Hanya 2 responden (5,9%) yang tidak akan merekomendasikan sistem.

Berdasarkan umpan balik kualitatif, kelebihan utama yang dirasakan responden mencakup: (1) *kecepatan dan kemudahan akses* — sistem dinilai responsif dan tidak memerlukan waktu tunggu lama; (2) *relevansi informasi* yang sesuai dengan kondisi nyata wisata Api Abadi Mrapen; (3) *kejelasan bahasa* yang komunikatif dan mudah dipahami; serta (4) kemampuan menjawab berbagai jenis pertanyaan secara informatif. Adapun kekurangan yang diidentifikasi meliputi: keterbatasan cakupan pengetahuan sistem untuk beberapa topik spesifik, antarmuka pengguna (UI) yang masih dapat ditingkatkan, serta beberapa jawaban yang dianggap kurang lengkap atau kurang relevan untuk pertanyaan tertentu.

Secara keseluruhan, hasil UAT dengan 34 responden menghasilkan nilai rata-rata **4,42/5,00** yang termasuk dalam kategori ***Sangat Baik***. Nilai ini merupakan peningkatan yang signifikan dibandingkan pengujian awal dengan 3 responden yang menghasilkan nilai 3,87/5,00. Dengan jumlah responden yang lebih besar dan profil yang lebih beragam, hasil UAT ini memberikan konfirmasi yang lebih kuat bahwa sistem tanya jawab pariwisata cerdas berbasis RAG telah diterima dengan baik oleh pengguna. Temuan ini juga perlu dikontekstualisasikan dalam kerangka perbedaan paradigma evaluasi. Pengguna yang berpartisipasi dalam UAT cenderung menilai sistem dari perspektif pengalaman interaksi secara holistik, di mana aspek yang pertama kali dipersepsi umumnya adalah tampilan antarmuka (*user experience*), kemudahan penggunaan (*ease of use*), serta kebermanfaatan praktis dalam memperoleh informasi. Kemampuan sistem dalam merangkai respons dengan bahasa yang natural dan komunikatif turut berkontribusi secara signifikan terhadap persepsi positif pengguna, terlepas dari proses komputasional yang mendasarinya.

Hal ini berbeda secara fundamental dengan evaluasi berbasis *cosine similarity*, yang merupakan pendekatan pengukuran matematis murni terhadap kemiripan semantik antara vektor representasi teks. Metrik tersebut tidak mempertimbangkan faktor subjektif seperti kelancaran bahasa, konteks pragmatis, maupun relevansi yang dipersepsi secara intuitif oleh pengguna. Oleh karena itu, kesenjangan yang mungkin terjadi antara skor *cosine similarity* dan tingkat kepuasan pengguna pada UAT bukan merupakan inkonsistensi, melainkan cerminan dari dua dimensi evaluasi yang saling melengkapi: dimensi objektif-komputasional dan dimensi subjektif-pengalaman pengguna.

4.6 Pembahasan

4.6.1 Analisis Kinerja Sistem RAG

Hasil evaluasi komprehensif menunjukkan bahwa sistem tanya jawab pariwisata berbasis Agentic RAG memberikan performa yang bervariasi bergantung pada metode evaluasi yang digunakan. Pada evaluasi Cosine Similarity, rata-rata skor 0.7039 (kategori Relevan) mengindikasikan kedekatan semantik yang cukup baik dengan ground truth, meski masih terdapat ruang peningkatan terutama untuk pertanyaan eksploratif seperti area parkir dan objek wisata pendukung.

Pada evaluasi RAGAS Dasar dengan 10 test case, rata-rata global 5 metrik adalah 0.656, dengan Faithfulness (0.916) dan Context Recall (0.750) telah memenuhi target yang ditetapkan sementara tiga metrik lain masih di bawah target. Pada evaluasi RAGAS Diperkuat dengan 30 test case, hasil justru menunjukkan bahwa Agentic RAG mengungguli Baseline RAG pada tiga dari empat metrik yang terevaluasi (Context Precision, Context Recall, dan Answer Correctness), mengindikasikan bahwa mekanisme self-correction LangGraph mulai memberikan nilai tambah yang konsisten pada skala pengujian yang lebih besar, meskipun keunggulan ini belum terlihat pada metrik Faithfulness dan tetap memerlukan pengkajian lebih lanjut.

4.6.2 Temuan Penting dan Keterbatasan Sistem

Pada evaluasi RAGAS Dasar, seluruh metrik pengujian telah berhasil dihitung secara optimal, meskipun metrik *Answer Relevancy* masih menghasilkan skor 0.000 pada sebagian *test case*. Temuan yang lebih signifikan justru muncul pada evaluasi RAGAS Diperkuat, di mana *Agentic* RAG justru mengungguli Baseline RAG pada tiga dari empat metrik yang dievaluasi, yaitu Context Precision, Context Recall, dan Answer Correctness, dengan hanya metrik Faithfulness yang sedikit lebih rendah dibandingkan Baseline (selisih -0.014). Hasil ini konsisten dengan hipotesis awal penelitian yang mengasumsikan bahwa mekanisme *self-correction* pada LangGraph dapat meningkatkan kualitas jawaban, meskipun peningkatan tersebut belum merata pada seluruh metrik. Kondisi ini mengindikasikan bahwa proses *grading* dan *self-correction* pada *pipeline* Agentic secara umum efektif memperbaiki ketepatan konteks dan kebenaran jawaban, namun belum cukup untuk menjaga tingkat *faithfulness* yang setara dengan pendekatan Baseline yang lebih sederhana. Oleh karena itu, diperlukan evaluasi dan kalibrasi lebih lanjut terhadap penetapan nilai *threshold* serta logika *self-correction* agar konsistensi Faithfulness turut meningkat di dalam sistem.

Dari sisi evaluasi kualitatif, *User Acceptance Testing* (UAT) yang melibatkan 34 responden dengan profil beragam menunjukkan tingkat penerimaan yang sangat baik dengan skor rata-rata kepuasan mencapai 4,42/5,00. Meskipun demikian, tingginya skor UAT ini perlu diinterpretasikan secara kritis mengingat skor metrik objektif RAGAS pada kedua skenario pengujian belum sepenuhnya memenuhi target ideal. Disparitas hasil kuantitatif dan

kualitatif ini mengindikasikan dua kemungkinan utama: pertama, adanya bias afirmasi dari responden yang cenderung memberikan penilaian subjektif secara positif; dan kedua, tingginya kemampuan generatif LLM dalam merangkai kalimat yang koheren secara linguistik dapat menutupi ketidaktepatan dokumen referensi dari komponen *retrieval*. Akibatnya, jawaban yang dihasilkan sistem terkesan berkualitas dan meyakinkan di mata pengguna meskipun presisi pengambilan konteks dari basis data belum sepenuhnya optimal. Di luar hal tersebut, aspek keterbatasan cakupan pengetahuan sistem dan penyempurnaan kualitas antarmuka tetap menjadi catatan perbaikan yang perlu ditindaklanjuti di masa mendatang.

4.6.3 Implikasi Hasil terhadap Tujuan Penelitian

Mengacu pada tujuan penelitian yang telah dirumuskan pada Bab I, penelitian ini telah memenuhi seluruh target implementasi dan fungsionalitas yang ditetapkan. Pertama, sistem tanya jawab berbasis Agentic Retrieval-Augmented Generation (RAG) telah dikembangkan menggunakan framework LangChain dan LangGraph dengan memproses 643 ulasan pengunjung valid dan 20 data wawancara pengelola menjadi 738 vektor embedding dalam indeks FAISS, serta dilengkapi mekanisme self-correction berbasis threshold cosine similarity 0,30. Kedua, hasil evaluasi metrik internal menunjukkan capaian yang bervariasi: rata-rata Cosine Similarity (0,7039) dan rata-rata global RAGAS Dasar (0,656) menunjukkan performa yang cukup baik meski belum ideal, dan pada evaluasi RAGAS Diperkuat dengan skala pengujian yang lebih besar, pendekatan Agentic RAG justru menunjukkan keunggulan yang konsisten dibandingkan model Baseline pada tiga dari empat metrik yang dievaluasi (Context Precision, Context Recall, dan Answer Correctness), meski masih sedikit tertinggal pada metrik Faithfulness — sebuah temuan yang mengonfirmasi nilai tambah mekanisme self-correction sekaligus menjadi catatan bagi penyempurnaan arsitekturnya ke depannya. Meskipun demikian, sistem ini menunjukkan tingkat penerimaan pengguna yang tinggi pada pengujian akhir, yang dibuktikan melalui perolehan skor User Acceptance Testing (UAT) sebesar 4,42/5,00 dari 34 responden, dengan 91,2% di antaranya mengonfirmasi kebermanfaatan sistem. Secara keseluruhan, pemenuhan indikator implementasi dan penerimaan pengguna menegaskan tercapainya tujuan umum penelitian, yaitu meningkatkan

aksesibilitas informasi pariwisata di Api Abadi Mrapen, meskipun aspek optimalisasi arsitektur Agentic RAG pada metrik Faithfulness masih memerlukan penelitian lanjutan.