

Implementation of RAG in Jambi Cultural Information Question Answering System Using Qwen2.5-7B

Devit Saputra¹, Uung Ungkawa^{2*}

^{1,2}Program Studi Informatika, Institut Teknologi Nasional Bandung, Indonesia
devrafaezya23@gmail.com¹, uung@itenas.ac.id²

Article Info

Article history:

Received 2026-09-10

Revised 2026-10-14

Accepted 2026-10-18

Keyword:

Confidence Scorer,
Hallucination,
Hybrid Search,
Jambi Culture,
RAG.

ABSTRACT

Large Language Models (LLMs) are inherently closed-book, making them prone to hallucination when answering domain-specific queries with limited training data, such as local culture. Information on Jambi's cultural heritage remains scattered across various sources, causing inefficient integrated retrieval. This study implements a Retrieval-Augmented Generation (RAG) architecture using the open-source LLM Qwen2.5-7B, strengthened by Hybrid Search (BM25 and dense retrieval fused via Reciprocal Rank Fusion), BGE-Reranker-Large reranking, and a Confidence Scorer. Three retrieval configurations Dense RAG, BM25 RAG, and Hybrid RAG were evaluated end-to-end using 130 In-Domain (ID) and 50 Out-of-Context (OOC) test queries, measured through Precision@10, Recall@10, Final Context Precision/Recall, BLEU-1, BLEU-4, ROUGE-L, Groundedness, and Rejection metrics (FRR/TRR). Results show that Dense RAG performed most dominantly end-to-end, achieving the highest Pure Recall@10 (0.992), BLEU-1 (0.499), ROUGE-L (0.506), Groundedness (0.611), and lowest False Rejection Rate (11.5%). BM25 RAG showed a specific advantage in Final Context Precision (0.746) after reranking. Hybrid RAG offered a competitive performance compromise, with Final Recall (0.938) and BLEU-4 (0.288) on par with Dense, alongside absolute hallucination protection (100% True Rejection Rate) on OOC queries. These findings indicate that while Hybrid fusion and the Confidence Scorer effectively mitigate hallucination, pure Dense retrieval consistently provides the most robust context support on the tested Jambi Culture dataset. These results demonstrate that systematic Hybrid Search and Confidence Scorer integration, combined with a dual In-Domain/Out-of-Context evaluation framework, provides a practical and empirically grounded configuration guideline for building hallucination-resistant RAG systems in low-resource local-culture domains.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. PENDAHULUAN

Perkembangan Large Language Model (LLM) telah meningkatkan kemampuan sistem Question Answering (QA) dalam menghasilkan jawaban berbasis bahasa alami secara interaktif [1]. Namun, model LLM pada dasarnya bersifat closed-book, yang berarti sistem hanya mengandalkan pengetahuan laten yang direpresentasikan selama proses pelatihan. Akibatnya, model berpotensi menghasilkan informasi yang tidak akurat atau mengalami fenomena halusinasi (hallucination) ketika harus menjawab pertanyaan

spesifik mengenai domain yang memiliki ketersediaan data terbatas, seperti kebudayaan lokal [2], [3]. Kajian terbaru juga menunjukkan bahwa halusinasi ini idealnya dimitigasi melalui pendekatan yang dapat mendeteksi ketidakpastian semantik dari keluaran model [4].

Provinsi Jambi merupakan salah satu daerah di Indonesia yang memiliki beragam warisan budaya, seperti sejarah Kerajaan Melayu Jambi, Rumah Adat Kajang Lako, Batik Angso Duo, dan Tari Sekapur Sirih [5]. Informasi mengenai budaya tersebut tersebar pada berbagai sumber, seperti buku, artikel ilmiah, repositori pemerintah, dan dokumen digital,

sehingga proses pencarian informasi budaya secara terpadu menjadi kurang efisien.

Berbagai penelitian sebelumnya telah mengembangkan sistem QA menggunakan metode pencarian leksikal, seperti algoritma BM25, maupun pendekatan Retrieval-Augmented Generation (RAG). Algoritma BM25 terbukti efektif dalam menemukan dokumen berdasarkan kecocokan kata kunci, namun kurang mampu memahami hubungan semantik antar kata [6]. Untuk mengatasi keterbatasan tersebut, beberapa penelitian mengombinasikan pencarian leksikal dan semantik melalui pendekatan Hybrid Search [7]. Pendekatan ini memanfaatkan BM25 untuk mempertahankan kecocokan kata kunci sekaligus dense retrieval untuk menangkap hubungan semantik antar kata, sehingga relevansi dokumen yang diperoleh menjadi lebih baik sebelum disuntikkan ke dalam model generatif. Implementasi RAG telah dibuktikan keberhasilannya pada berbagai bidang, seperti penelitian Chang et al. (2024) [8] pada pelestarian budaya Hakka, sistem tanya jawab warisan budaya takbenda Tiongkok [9], dan studi Bora et al. (2024) [10] pada domain medis. Namun, sebagian besar penelitian tersebut belum mengintegrasikan mekanisme penolakan secara eksplisit, padahal evaluasi LLM pada domain budaya lokal sangat rentan terhadap kueri di luar konteks [11].

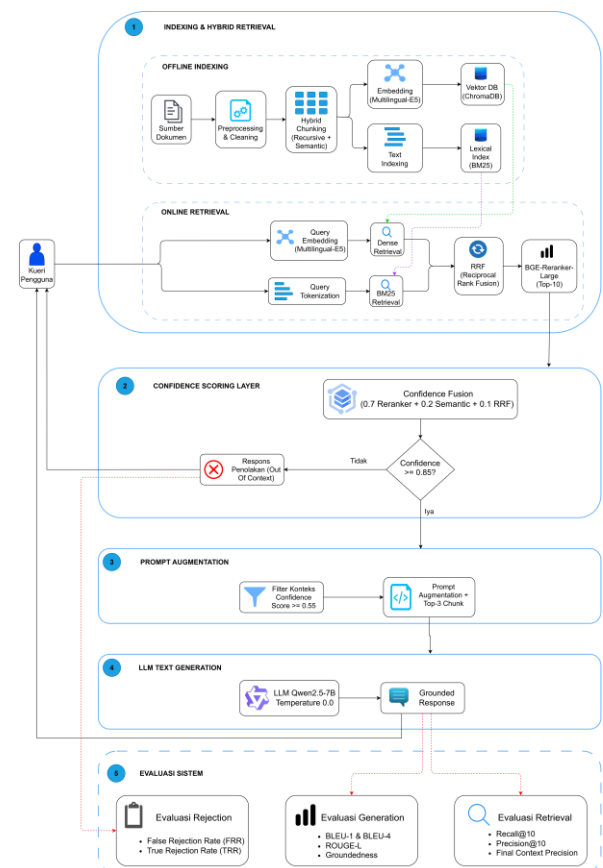
Penelitian yang mengintegrasikan Hybrid Search (BM25 dan dense retrieval) dengan model BGE-Reranker-Large, serta mengevaluasinya menggunakan metrik In-Domain (ID) dan Out-of-Context (OOC) Hallucination Rate pada domain budaya lokal, khususnya Budaya Jambi, masih belum banyak dieksplorasi secara komprehensif. Studi terbaru pada domain serupa, seperti RAG untuk nuansa budaya pada penerjemahan bahasa minoritas [20] dan RAG untuk sistem percakapan bahasa Swahili berdata terbatas [21], menegaskan relevansi arah penelitian ini namun belum menyentuh domain budaya lokal Indonesia maupun kerangka evaluasi ID/OOC ganda yang diusulkan di sini. Oleh karena itu, kontribusi penelitian ini bukan pada penciptaan algoritma baru, melainkan pada integrasi sistematis Hybrid Search, mekanisme reranking, dan Confidence Scorer yang tervalidasi secara empiris untuk meningkatkan keandalan sistem pada domain berdata terbatas, serta penerapan kerangka evaluasi ganda (ID dan OOC Hallucination Rate) yang memberikan gambaran mitigasi halusinasi secara lebih menyeluruh dibandingkan metrik tunggal yang lazim digunakan pada studi RAG sejenis.

Berdasarkan permasalahan tersebut, penelitian ini mengimplementasikan arsitektur RAG menggunakan LLM open-source Qwen2.5-7B yang diperkuat dengan Hybrid Search dan mekanisme reranking menggunakan BGE-Reranker-Large. Selain mengintegrasikan Hybrid Search dan BGE-Reranker-Large, penelitian ini juga mengevaluasi pengaruh confidence threshold terhadap kemampuan sistem dalam mengurangi halusinasi dan menolak pertanyaan di luar konteks. Kualitas sistem dievaluasi secara komprehensif menggunakan metrik penarikan dokumen (Precision, Recall), metrik kualitas teks (BLEU, ROUGE, Groundedness) [12], serta metrik penolakan (False/True Rejection Rate) yang dirancang untuk mengukur secara spesifik kemampuan sistem

menolak kueri yang tidak didukung oleh konteks. Penelitian ini diharapkan dapat memberikan rekomendasi objektif mengenai konfigurasi sistem QA berbasis RAG yang mampu meningkatkan akurasi faktual dan mengurangi tingkat halusinasi pada domain pelestarian budaya lokal. Secara spesifik, kebaruan empiris penelitian ini terletak pada tiga hal: (1) identifikasi dan koreksi bug transformasi sigmoid ganda pada skema Confidence Fusion yang jika tidak diperbaiki akan menyingkirkan kandidat dokumen relevan secara sistematis, (2) kerangka evaluasi ganda ID/OOC yang memisahkan secara eksplisit kalibrasi parameter dari pengujian akhir untuk mencegah kebocoran data (data leakage), dan (3) analisis komparatif berjenjang (retrieval mentah → konteks pasca-reranking → kualitas generasi → keamanan sistem) yang mengungkap bahwa keunggulan satu metode pada satu tahap tidak konsisten menjalar ke tahap berikutnya.

II. METODE

Pada Penelitian ini menggunakan pendekatan kuantitatif eksperimental berbasis sistem yang terdiri dari beberapa tahapan utama, yaitu pengumpulan data, pra-pemrosesan, fase indexing (offline), fase inference (online), dan evaluasi sistem. Secara keseluruhan, alur arsitektur Retrieval-Augmented Generation (RAG) yang diusulkan dapat dilihat pada Gambar 1.



Gambar 1. Arsitektur Pipeline Retrieval-Augmented Generation (RAG) pada Domain Budaya Jambi

A. Pengumpulan Data dan Preprocessing

Pengumpulan data pada penelitian ini menggunakan pendekatan campuran yang terdiri dari pencarian data primer dan sekunder. Data primer diperoleh melalui kunjungan observasi langsung ke Dinas Kebudayaan Provinsi Jambi, yang menghasilkan satu dokumen rujukan utama terkait pelestarian budaya lokal. Berdasarkan arahan dari instansi tersebut, pengumpulan data sekunder kemudian dilakukan untuk melengkapi basis pengetahuan melalui ekstraksi literatur tekstual dari repositori resmi pemerintah (seperti Kemendikbud) serta publikasi jurnal akademik.

Total dataset yang berhasil dikumpulkan terdiri dari 36 dokumen literatur terkait Informasi Budaya Jambi dengan total keseluruhan 389.900 token. Untuk memastikan representasi pengetahuan yang komprehensif, dataset dibagi ke dalam tiga sub-domain utama: Sejarah & Tradisi Lisan (32%), Kesenian, Kuliner & Pariwisata (42%), serta Warisan Budaya & Bahasa Daerah (26%).

Proses pengumpulan data berlangsung pada periode Oktober 2025–Januari 2026. Dari sisi karakteristik bahasa, korpus didominasi Bahasa Indonesia baku dengan sisipan istilah dan penamaan lokal Melayu Jambi/Kerinci (misalnya nama gelar adat, nama tarian, dan nama motif batik) yang tidak diterjemahkan agar makna kultural tetap terjaga. Setiap chunk diberi metadata terstruktur yang mencakup nama_dokumen, sub_domain, nomor_halaman, topik_utama, dan id_chunk untuk mendukung penelusuran kembali (traceability). Pemeriksaan duplikasi dilakukan secara otomatis menggunakan hashing SHA-256 pada isi teks yang dilanjutkan dengan pemindaian manual terhadap kemiripan judul dan bab dokumen, sehingga tidak ditemukan dokumen duplikat yang lolos ke tahap indexing.

Seluruh dokumen melalui tahap text cleaning untuk menghilangkan elemen non-leksikal dan melakukan normalisasi. Teks yang telah bersih kemudian dipotong menggunakan metode gabungan Recursive Chunking dan Semantic Chunking dengan ukuran ditetapkan sebesar 768 token dan overlap 200 token. Pendekatan hibrida ini diterapkan untuk menjaga koherensi narasi budaya yang panjang tanpa memutus konteks informasi secara paksa.

B. Fase Indexing

Fase indexing bertujuan mengonversi chunks teks menjadi format yang dapat dicari secara komputasional. Guna mendukung sistem pencarian hibrida, chunks disimpan ke dalam dua indeks penyimpanan terpisah. Pertama, teks dikonversi menjadi representasi vektor numerik menggunakan model embedding Multilingual-E5 untuk keperluan dense retrieval [13] dan diindeks ke dalam vector database ChromaDB (Vector Index). Kedua, data teks mentah dari chunk yang sama diindeks secara leksikal untuk membentuk BM25 Lexical Index [6].

C. Fase Inference dan Mekanisme Penolakan

Pada saat pengguna mengajukan kueri, sistem merespons secara real-time melalui mekanisme Hybrid Search. Kueri pengguna diproses secara paralel melalui dua jalur, yaitu

Query Embedding menggunakan model Multilingual-E5 untuk pencarian semantik pada ChromaDB, dan Query Tokenization untuk pencarian leksikal pada indeks BM25. Hasil dari kedua algoritma pencarian tersebut digabungkan menggunakan algoritma Reciprocal Rank Fusion (RRF) [14] untuk memastikan keseimbangan antara kemiripan makna dan kecocokan istilah eksak.

Kandidat dokumen dari hasil fusi kemudian disaring kembali secara presisi menggunakan model BGE-Reranker-Large [15]. Model ini menggunakan pendekatan cross-encoder, berbeda dari pendekatan late-interaction, yang mengevaluasi relevansi kueri-dokumen secara langsung dan kontekstual melalui satu kali forward pass [16]. Sebagai langkah mitigasi halusinasi, penelitian ini mengimplementasikan kalkulasi Confidence Fusion dengan persamaan matematis berikut:

$$\text{Confidence} = (0.7 \times \text{Skor}_{\text{Reranker}}) + (0.2 \times \text{Skor}_{\text{Semantic}}) + (0.1 \times \text{Skor}_{\text{RRF}}) \quad (1)$$

Bobot reranker ditetapkan dominan (70%) karena BGE-Reranker-Large merupakan model cross-encoder yang mengevaluasi relevansi kueri-dokumen secara langsung dan kontekstual, sehingga secara teoritis memberikan sinyal relevansi yang lebih presisi dibandingkan skor kemiripan embedding semantik (yang dihitung independen dari kueri spesifik) maupun skor RRF (yang murni berbasis peringkat tanpa mempertimbangkan makna). Proporsi ini konsisten dengan praktik umum arsitektur two-stage retrieval, di mana reranker berperan sebagai penentu akhir relevansi dokumen. Penetapan nilai bobot ini bersifat teoritis berdasarkan kajian literatur arsitektur retrieval dua tahap, dan belum melalui proses optimasi empiris seperti grid search atau hyperparameter tuning sistematis. Validasi sensitivitas bobot terhadap performa akhir sistem menjadi bagian dari rencana pengembangan penelitian selanjutnya.

Nilai Confidence ini kemudian dievaluasi terhadap nilai ambang batas utama (threshold) sebesar 0,85, yang ditetapkan berdasarkan hasil kalibrasi empiris dan divalidasi secara end-to-end (lihat Subbab 3.2). Jika nilai Confidence dokumen teratas (Top-1) berada di bawah ambang batas tersebut, sistem akan mengeksekusi mekanisme graceful failure dan mengeluarkan respons penolakan untuk pertanyaan Out-of-Context (OOC). Pendekatan ini diadaptasi dari konsep selective prediction, di mana model harus menahan diri dari memberikan jawaban saat dihadapkan pada kueri di luar domain [17]. Sebaliknya, jika nilai memenuhi ambang batas utama (0,85), kueri dianggap valid dan lolos dari penolakan. Selanjutnya, sistem menerapkan tahap kedua berupa ambang batas inklusi (context filter). Pada tahap ini, sistem mengevaluasi seluruh kandidat chunk dokumen yang ditarik, dan hanya menyertakan chunk yang memiliki skor minimal 0,55 ke dalam prompt LLM. Strategi ambang batas ganda (dual-threshold) ini memastikan sistem tetap kebal terhadap pertanyaan jebakan, sembari memperkaya wawasan LLM dengan dokumen pendukung yang relevan tanpa memasukkan informasi yang tidak relevan.

D. Fase Generation dan Evaluasi

Konteks dokumen yang tervalidasi disuntikkan bersama kueri pengguna ke dalam Prompt Augmentation dengan instruksi grounding yang ketat. Proses pembangkitan jawaban dilakukan oleh LLM Qwen2.5-7B [18]. Untuk mencegah kelebihan asupan teks dan fenomena lost-in-the-middle, konteks dibatasi maksimal 3 chunks terbaik. Temperature dikonfigurasi di angka 0,0 (Greedy Decoding) guna mengunci model agar menghasilkan luaran yang sangat deterministik dan menekan potensi fabrikasi teks.

Kinerja sistem dievaluasi secara komparatif antara tiga konfigurasi retrieval, yaitu Dense RAG, BM25 RAG, dan Hybrid RAG. Parameter evaluasi dibagi menjadi tiga fokus: Evaluasi Retrieval untuk mengukur kemampuan penemuan informasi melalui Precision@10 dan Recall@10, Evaluasi Generation menggunakan metrik N-Gram (BLEU-1, BLEU-4, ROUGE-L) [12], [19] serta Groundedness untuk mengukur keterikatan faktual jawaban terhadap konteks, dan Evaluasi Rejection untuk mengukur ketepatan penolakan melalui metrik False Rejection Rate (FRR) dan True Rejection Rate (TRR).

Untuk memastikan evaluasi yang objektif dan terhindar dari risiko kebocoran data (data leakage), total 150 kueri In-Domain (ID) dan 50 kueri Out-of-Context (OOC) dibagi ke dalam dua himpunan. Sebanyak 20 kueri ID dipisahkan secara eksklusif sebagai Development Set untuk mengkalibrasi parameter sistem. Sisa 130 kueri ID yang memiliki tingkat kesulitan tinggi (seperti pencarian angka, tahun, atau nama tokoh lokal) dan 50 kueri OOC (topik di luar budaya Jambi seperti teknologi atau geografi global) dikunci sebagai Test

Set murni. Dataset pengujian akhir ini digunakan untuk mengukur kinerja komparatif sistem yang disajikan pada bagian Hasil dan Pembahasan. Tabel 1 menyajikan contoh representatif dari dataset pengujian yang digunakan.

Kueri uji In-Domain (ID) disusun secara manual oleh peneliti dengan mengomposisikan pertanyaan berdasarkan dokumen sumber dan derivasi dari pola kueri umum masyarakat terhadap literatur budaya Jambi, dengan cakupan topik didistribusikan mengikuti proporsi tiga sub-domain korpus (Sejarah & Tradisi Lisan, Kesenian-Kuliner-Pariwisata, serta Warisan Budaya & Bahasa Daerah) agar representatif. Tingkat kesulitan kueri dikategorikan ke dalam dua tingkat, yaitu kueri faktual sederhana (seperti identifikasi nama entitas, definisi, dan lokasi) serta kueri faktual kompleks (seperti rincian angka spesifik, urutan prosedural ritual, dan pertanyaan yang menggunakan sinonim atau struktur kalimat berbeda dari dokumen sumber), dengan proporsi 40% kueri sederhana dan 60% kueri kompleks. Variasi linguistik turut disertakan, seperti kueri dengan sinonim istilah budaya dan kueri dengan redaksi yang berbeda dari kalimat sumber. Ground truth setiap kueri dicocokkan langsung dengan dokumen rujukan utama oleh peneliti dan diverifikasi kebenarannya melalui validasi silang (cross-validation) bersama dokumen resmi Dinas Kebudayaan Provinsi Jambi untuk memastikan relevansi dan akurasi jawaban rujukan. Kueri Out-of-Context (OOC) disusun dari topik yang secara tegas berada di luar cakupan Budaya Jambi (misalnya teknologi, geografi global, dan budaya daerah lain di Indonesia) untuk menguji spesifisitas mekanisme penolakan, bukan sekadar ketidakrelevanan umum.

TABEL 1 DATASET UJI IN-DOMAIN (ID) DAN OUT-OF-CONTEXT (OOC)

Tipe Kueri	ID Kueri	Pertanyaan (Kueri Pengguna)	Ground Truth
In-Domain	ID-002	Sebutkan secara spesifik 3 gelar adat yang dikukuhkan sebagai pengganti pemangku adat yang lama dalam upacara Kenduri Sko di Kerinci!	Tiga gelar adat yang dikukuhkan dalam upacara Kenduri Sko di Kerinci adalah depati, rio, dan ninik mamak
In-Domain	ID-016	Apa makna kearifan lokal yang terkandung dalam Rumah Adat Kajang Lako?	Rumah Adat Kajang Lako mengandung nilai kearifan lokal yang berkaitan dengan struktur sosial, religi, ...
Out-of-Context	OOC-002	Bagaimana cara membuat rendang Padang yang autentik?	Reject (Penolakan). Topik mengenai kuliner Minangkabau, bukan kuliner Jambi.
Out-of-Context	OOC-004	Bagaimana cara kerja mesin pembelajaran deep learning?	Reject (Penolakan). Topik teknologi komputer, sama sekali tidak berkaitan dengan budaya lokal.

E. Formulasi dan Definisi Metrik Evaluasi

Subbab ini merinci definisi operasional dan formulasi matematis dari setiap metrik yang digunakan pada evaluasi retrieval, generasi, dan keamanan sistem (Subbab III.C hingga III.E), guna mendukung reproducibility penelitian.

1) Metrik Penolakan (False Rejection Rate dan True Rejection Rate)

False Rejection Rate (FRR) mengukur proporsi kueri In-Domain yang valid namun ditolak oleh sistem terhadap total kueri In-Domain:

$$FRR = \left(\frac{\sum_{i=1}^{N_{ID}} \mathbb{I}(\text{Confidence}(q_i, c_{top1}) < \tau)}{N_{ID}} \right) \times 100\% \quad (2)$$

True Rejection Rate (TRR) mengukur proporsi kueri Out-of-Context yang berhasil ditolak terhadap total kueri OOC:

$$\text{TRR} = \left(\frac{\sum_{j=1}^{N_{\text{OOC}}} \mathbb{I}(\text{Confidence}(q_j, c_{\text{top1}}) < \tau)}{N_{\text{OOC}}} \right) \times 100\% \quad (3)$$

dengan N_{ID} dan N_{OOC} berturut-turut adalah jumlah total kueri *In-Domain* ($N_{\text{ID}} = 130$) dan *Out-of-Context* ($N_{\text{OOC}} = 50$) pada dataset pengujian; $\mathbb{I}(\cdot)$ adalah fungsi indikator yang bernilai 1 jika kondisi di dalam kurung terpenuhi dan 0 jika sebaliknya; c_{top1} adalah *chunk* dokumen dengan skor keterkaitan tertinggi yang ditarik untuk kueri q ; serta τ adalah nilai ambang batas utama kepercayaan sistem ($\tau = 0,85$). Keputusan tolak/terima ditentukan sepenuhnya oleh mekanisme *dual-threshold* pada *Confidence Scorer*, sehingga FRR/TRR merupakan metrik deterministik berbasis skor *confidence*, bukan hasil penilaian manusia atau model evaluator tambahan.

2) Metrik Penarikan Dokumen (*Precision@k* dan *Recall@k*)

Evaluasi modul *retrieval* diukur menggunakan *Precision* dan *Recall* pada k kandidat dokumen teratas ($k = 10$):

$$\text{Precision@}k = \frac{|R \cap D_k|}{k} \quad (4)$$

$$\text{Recall@}k = \frac{|R \cap D_k|}{|R|} \quad (5)$$

dengan R adalah himpunan dokumen relevan yang menjadi rujukan kebenaran (*ground truth*), dan D_k adalah himpunan k dokumen teratas yang ditarik oleh modul *retrieval*.

3) Metrik Kualitas Generasi Teks (*BLEU-n* dan *ROUGE-L*)

BLEU-n mengukur tingkat kecocokan n -gram antara jawaban generasi model (A) dan jawaban acuan (G):

$$\text{BLEU-n} = \text{BP} \cdot \exp \left(\sum_{i=1}^n w_i \ln p_i \right) \quad (6)$$

dengan *Brevity Penalty* (BP) = 1 jika $|A| > |G|$, dan $\text{BP} = \exp \left(1 - \frac{|G|}{|A|} \right)$ jika $|A| \leq |G|$; p_i adalah presisi dari i -gram yang cocok antara jawaban generasi dan acuan; w_i adalah bobot seragam untuk tiap ukuran i -gram ($w_i = 1/n$); serta $|A|$ dan $|G|$ adalah panjang kata dari jawaban generasi (A) dan jawaban acuan (G).

ROUGE-L mengukur kesamaan struktur kalimat berbasis *Longest Common Subsequence* (LCS):

$$\text{ROUGE-L} = \frac{(1 + \beta^2) P_{\text{LCS}} R_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}} \quad (7)$$

dengan $P_{\text{LCS}} = \text{LCS}(A, G)/|A|$ dan $R_{\text{LCS}} = \text{LCS}(A, G)/|G|$; $\text{LCS}(A, G)$ adalah panjang subsekuensi bersama terpanjang antara jawaban generasi A dan acuan G ; serta β adalah parameter penyeimbang antara presisi dan *recall* (umumnya ditetapkan $\beta = 1$).

4) Metrik Groundedness

Groundedness mengukur sejauh mana klaim faktual pada jawaban yang dihasilkan LLM didukung oleh konteks dokumen pendukung yang disuntikkan (C):

$$\text{Groundedness}(A, C) = \frac{|F_A \cap C|}{|F_A|} \quad (8)$$

dengan F_A adalah himpunan klaim faktual yang diekstraksi dari jawaban generasi A , dan $|F_A \cap C|$ adalah jumlah klaim faktual yang pernyataannya terverifikasi secara eksplisit di dalam konteks C . Skor *Groundedness* diukur menggunakan skema *LLM-as-judge* terstruktur, di mana model Qwen2.5-7B-Instruct diberi instruksi (*prompting*) khusus untuk menilai apakah setiap klaim faktual dalam jawaban didukung secara langsung oleh konteks yang diupayakan, dengan skala biner (0 untuk klaim tanpa bukti dan 1 untuk klaim tervalidasi) yang kemudian dirata-ratakan dan dinormalisasi menjadi skor kontinu berkisar 0,0 hingga 1,0. Pendekatan ini dipilih untuk menggantikan pustaka evaluasi RAGAS setelah ditemukan *bug* ketidakstabilan penanganan *output* berupa kegagalan *parsing* format JSON dan pemotongan (*truncation*) akibat batas token ketika mengevaluasi konteks berbahasa Indonesia lokal, yang mengakibatkan tingginya tingkat *parsing error* dan penilaian skor yang tidak konsisten.

III. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil implementasi dan evaluasi sistem RAG secara berurutan mengikuti tahapan pipeline: praproses dan pembentukan basis pengetahuan, kalibrasi komponen Confidence Scorer, evaluasi komparatif kinerja sistem, evaluasi mekanisme penolakan (*rejection*) untuk mitigasi halusinasi, serta analisis kualitatif berbasis studi kasus.

A. Hasil Praproses dan Konstruksi Basis Pengetahuan

Proses indexing berhasil mengolah 36 dokumen sumber literatur Budaya Jambi yang terdiri dari tiga sub-domain utama: Sejarah & Tradisi Lisan (32%), Kesenian, Kuliner & Pariwisata (42%), serta Warisan Budaya & Bahasa Daerah (26%). Sebelum digunakan, tiga masalah kualitas data ditemukan dan diperbaiki pada tahap praproses: (1) baris caption gambar/tabel yang ikut terekstraksi ke dalam chunk narasi, diatasi dengan aturan ekspresi reguler sebelum tahap chunking, (2) konflasi makna filosofis antar-motif budaya akibat ukuran chunk (768 token) yang relatif besar, diatasi dengan tahap pra-pemisahan (*pre-splitting*) berbasis penanda, dan (3) mekanisme deteksi header/footer berulang yang pada

awalnya turut menghapus konten sah, diperbaiki dengan membatasi kandidat deteksi hanya pada posisi baris tepi serta menaikkan ambang kemunculan minimum dari dua menjadi tiga halaman.

Setelah seluruh perbaikan diterapkan, proses indexing akhir menghasilkan 482 chunk dari 36 dokumen sumber, dengan setiap chunk diberi identifier unik berformat {nama_dokumen}_chunk_{n} untuk memastikan tidak ada tabrakan ID lintas dokumen. Basis pengetahuan yang telah dibangun ini selanjutnya digunakan pada tahap kalibrasi sistem dan evaluasi akhir yang dipaparkan pada subbab berikut.

B. Kalibrasi Sistem dan Pembagian Dataset

Pengujian awal terhadap komponen Confidence Scorer menemukan bahwa skor reranker dari BGE-Reranker-Large yang sesungguhnya sudah berupa probabilitas ternormalisasi (rentang 0-1) diterapkan fungsi sigmoid untuk kedua kalinya pada tahap Confidence Fusion. Transformasi ganda ini memampatkan rentang skor yang seharusnya diskriminatif menjadi sangat sempit, sehingga komponen reranker yang mendominasi 70% bobot formula kehilangan sebagian besar daya pembeda. Tabel II menunjukkan dampak konkret bug tersebut pada satu chunk sampel yang relevansinya dinilai sangat tinggi oleh reranker (skor mentah 0,9692), namun sebelum perbaikan justru tersingkir dari kandidat karena skor confidence gabungannya jatuh di bawah ambang batas. Perbaikan dilakukan dengan menghapus transformasi sigmoid tambahan pada komponen reranker dan menggantinya dengan pembatasan sederhana (clamping) ke rentang [0,1].

TABEL II

DAMPAK BUG TRANSFORMASI SIGMOID GANDA PADA SKOR CONFIDENCE

Komponen	Sebelum Perbaikan	Sesudah Perbaikan
Skor Reranker Mentah	0,9692	0,9692
Skor Reranker Setelah Transformasi	0,72 (sigmoid ganda)	0,97 (tanpa transformasi tambahan)
Skor Confidence Gabungan	0,673	0,845
Status terhadap Ambang Batas 0,70	Tidak Lolos (chunk relevan tersingkir)	Lolos

Terpisah dari perbaikan bug tersebut, ambang batas (threshold) Confidence Scorer dikalibrasi secara empiris menggunakan kombinasi kueri kalibrasi tahap awal (30 In-Domain, 20 Out-of-Context), dengan hasil FRR/TRR pada tiga kandidat nilai (0,70; 0,85; 0,90) disajikan pada Tabel III. Nilai 0,85 dipilih sebagai konfigurasi final karena merepresentasikan titik trade-off paling seimbang di antara ketiga kandidat yang diuji. Perlu dicatat bahwa putaran kalibrasi ini mendahului finalisasi pembagian dataset evaluasi akhir yang digunakan pada Subbab C hingga F, sehingga tabel ini disajikan sebagai dokumentasi historis proses penentuan threshold, bukan sebagai deskripsi dataset evaluasi akhir.

TABEL III HASIL KALIBRASI THRESHOLD (FRR/TRR)

Threshold	False Rejection Rate (FRR)	True Rejection Rate (TRR)
0,70	3,33%	40,0%
0,85	6,67%	60,0%
0,90	16,67%	85,0%

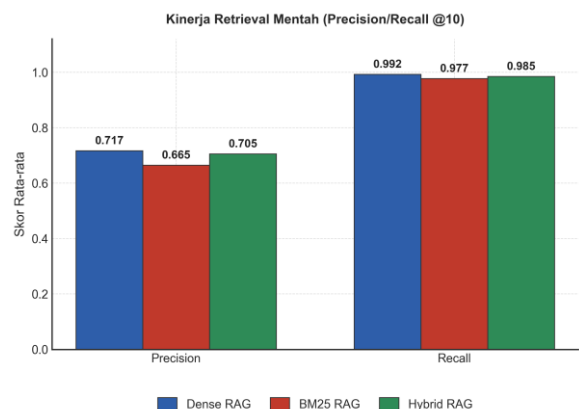
Konfigurasi threshold 0,85 yang telah ditetapkan tersebut kemudian digunakan pada evaluasi akhir sistem Dense RAG, BM25 RAG, dan Hybrid RAG yang menjadi fokus penelitian ini. Dari total 150 kueri In-Domain, sebanyak 20 kueri dipisahkan sebagai Development Set. Development Set digunakan untuk proses kalibrasi parameter, termasuk konfigurasi fusion/RRF dan penentuan threshold Confidence Scorer sebesar 0,85. Sebanyak 130 kueri In-Domain yang tersisa tidak digunakan dalam proses tuning dan digunakan sebagai Test Set untuk evaluasi akhir. Selain itu, 50 kueri Out-of-Context digunakan untuk mengevaluasi kemampuan sistem dalam menolak pertanyaan yang tidak didukung oleh basis pengetahuan.

Pemisahan ini memiliki dua implikasi metodologis. Pertama, parameter sistem (bobot fusion/RRF dan threshold Confidence Scorer) tidak ditentukan menggunakan data yang sama dengan data yang dipakai untuk melaporkan performa akhir, sehingga mengurangi risiko overfitting terhadap dataset evaluasi. Kedua, seluruh hasil yang dilaporkan pada Subbab C hingga F harus diposisikan sebagai hasil pengujian pada 130 kueri In-Domain dan 50 kueri Out-of-Context yang tidak dilihat selama proses kalibrasi, bukan sebagai hasil pada data yang sama dengan yang digunakan untuk menyetel sistem.

C. Evaluasi Kinerja Modul Penarikan Dokumen (Retrieval)

1) Kinerja Retrieval Mentah

Evaluasi retrieval mentah pada Precision@10 dan Recall@10 mengukur kualitas kandidat dokumen sebelum tahap reranking diterapkan, dihitung dari 130 kueri Test Set In-Domain. Gambar 2 menyajikan perbandingan ketiga metode.

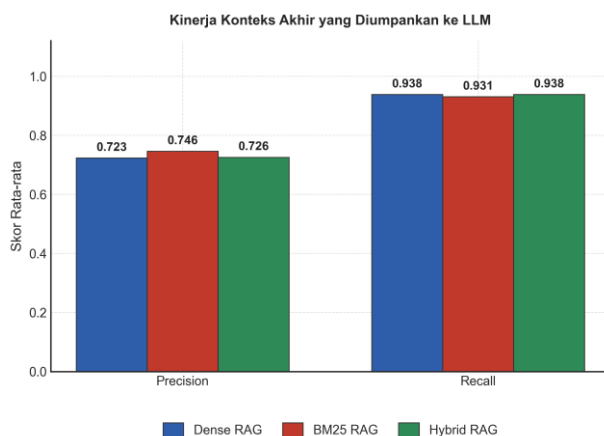


Gambar 2. Kinerja Retrieval Mentah (Precision/Recall @10)

Dense RAG mencatatkan Precision@10 dan Recall@10 tertinggi di antara ketiga metode (0,717 dan 0,992), diikuti oleh Hybrid RAG (0,705 dan 0,985) dan BM25 RAG (0,665 dan 0,977). Pada tahap retrieval mentah ini, Dense merupakan metode dengan performa terbaik pada kedua metrik. Keunggulan Dense mengindikasikan bahwa representasi semantik dari Multilingual-E5 lebih konsisten menemukan kandidat dokumen yang relevan pada tahap awal penarikan dibandingkan pencocokan leksikal murni BM25, yang kinerjanya bergantung pada kecocokan istilah eksak antara kueri dan dokumen. Hybrid RAG, yang menggabungkan sinyal Dense dan BM25 melalui fusi peringkat (RRF), berada di antara kedua komponen penyusunnya tetapi tidak melampaui Dense pada kedua metrik. Hal ini mengindikasikan bahwa penggabungan dua retriever pada konfigurasi RRF yang digunakan tidak secara otomatis menghasilkan nilai yang lebih tinggi daripada komponen terbaiknya; kandidat tambahan yang dibawa oleh BM25 ke dalam candidate pool gabungan tampak membawa noise yang menahan Precision@10 Hybrid di bawah Dense.

2) Kualitas Konteks Setelah Reranking

Setelah kandidat retrieval mentah disaring dan diberi peringkat ulang oleh BGE-Reranker-Large, hanya Top-3 chunk yang dipertahankan sebagai konteks akhir yang diumpankan ke LLM. Gambar 3 menyajikan Final Context Precision dan Final Context Recall pada konfigurasi Top-3 tersebut.



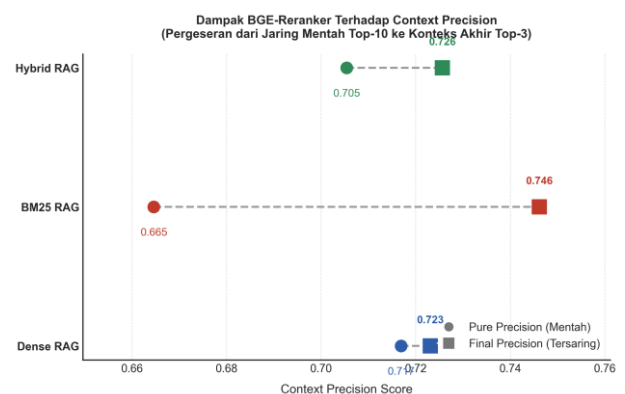
Gambar 3. Kinerja Konteks Akhir yang Diumpukan ke LLM

Pada tahap ini, BM25 RAG mencatatkan Final Context Precision tertinggi (0,746), melampaui Dense RAG (0,723) dan Hybrid RAG (0,726), sementara Dense dan Hybrid mempertahankan Final Context Recall tertinggi secara setara (0,938) dibandingkan BM25 (0,931). Temuan ini menunjukkan bahwa hasil retrieval mentah tidak selalu identik dengan hasil konteks yang akhirnya dikirim ke LLM; proses reranking mengubah komposisi dan kualitas kandidat secara berbeda pada setiap strategi retrieval. Perlu ditegaskan bahwa hasil ini tidak dapat diartikan sebagai bukti bahwa BM25 adalah metode terbaik secara keseluruhan, karena

reranking hanya menyaring ulang kandidat yang tersedia dari tahap retrieval mentah masing-masing metode, dan BM25 tetap tercatat sebagai metode dengan Precision@10 mentah terendah pada Subbab C.1. Peningkatan precision setelah reranking pada BM25 lebih tepat diinterpretasikan sebagai indikasi bahwa kandidat leksikal yang lolos ke tahap reranking, meski jumlahnya lebih sedikit relevansinya pada tahap mentah, memiliki tingkat kecocokan istilah dengan kueri yang membantu BGE-Reranker-Large memprioritaskannya secara lebih konsisten pada peringkat Top-3.

3) Dampak BGE-Reranker terhadap Context Precision

Untuk memperjelas pengaruh reranking terhadap masing-masing metode, Gambar 4 menyajikan diagram dumbbell yang membandingkan Precision retrieval mentah (@10) dengan Final Context Precision (Top-3) untuk ketiga metode.



Gambar 4. Dampak BGE-Reranker terhadap Context Precision (Pergeseran dari Retrieval Mentah Top-10 ke Konteks Akhir Top-3)

Ketiga metode mengalami peningkatan precision setelah reranking: Dense RAG meningkat dari 0,717 menjadi 0,723 (+0,006), Hybrid RAG dari 0,705 menjadi 0,726 (+0,021), dan BM25 RAG dari 0,665 menjadi 0,746 (+0,081). Peningkatan paling besar secara nominal terjadi pada BM25, namun temuan ini tidak dapat disimpulkan sebagai bukti bahwa reranker menjadikan BM25 sebagai metode terbaik secara keseluruhan, mengingat titik awal precision mentah BM25 juga merupakan yang terendah di antara ketiga metode. Pola ini lebih tepat diinterpretasikan sebagai peningkatan selektivitas: BGE-Reranker-Large berfungsi sebagai tahap penyaringan kandidat dan reprioritisasi dokumen berdasarkan relevansi query-dokumen yang dihitung secara cross-encoder, sehingga dampak peningkatan kualitas konteks yang dihasilkannya berbeda-beda bergantung pada karakteristik candidate pool awal dari setiap strategi retrieval. Pada BM25, candidate pool awal yang didominasi kecocokan kata kunci tampak mengandung variasi relevansi yang lebih besar, sehingga ruang bagi reranker untuk melakukan penyaringan kandidat secara efektif juga lebih besar dibandingkan pada Dense yang candidate pool awalnya sudah relatif presisi sejak tahap retrieval mentah. Secara ringkas, BGE-Reranker-Large berfungsi sebagai tahap seleksi ulang yang meningkatkan kualitas konteks akhir pada ketiga strategi retrieval, dengan

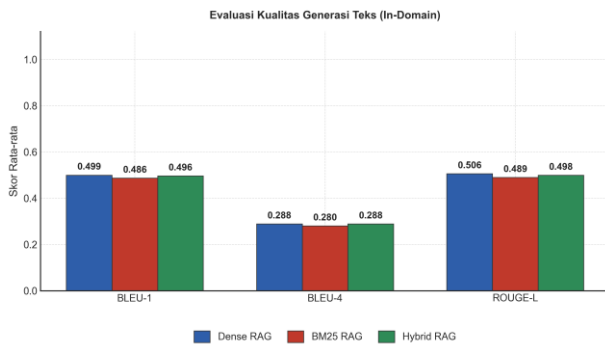
besaran dampak peningkatan yang berbeda pada setiap strategi.

D. Evaluasi Kualitas Generasi Teks

Kualitas generasi jawaban dievaluasi pada 130 kueri Test Set In-Domain menggunakan metrik BLEU-1, BLEU-4, dan ROUGE-L terhadap referensi (ground truth), serta metrik Groundedness untuk mengukur keterikatan jawaban terhadap konteks yang diumpangkan. Tabel IV merangkum hasil ketiga metrik generation, dan Gambar 5 menyajikannya secara visual.

TABEL IV
PERBANDINGAN METRIK UTAMA GENERATION (130 KUERI TEST SET IN-DOMAIN)

Metrik	Dense RAG	BM25 RAG	Hybrid RAG
BLEU-1	0,499	0,486	0,496
BLEU-4	0,288	0,280	0,288
ROUGE-L	0,506	0,489	0,498
Groundedness	0,611	0,595	0,595



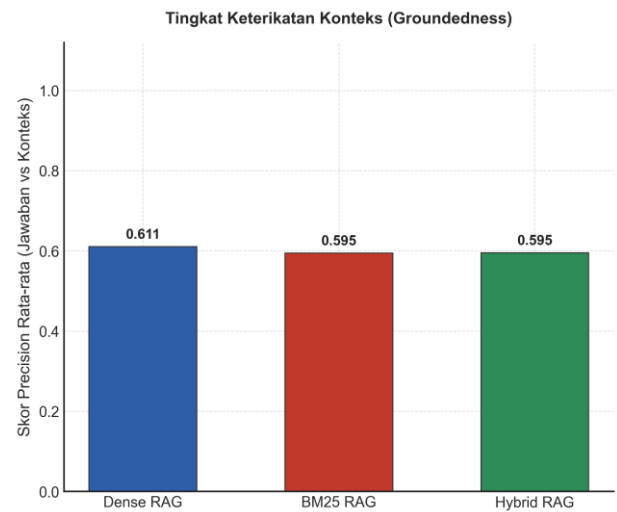
Gambar 5. Evaluasi Kualitas Generasi Teks (In-Domain)

Dense RAG mencatatkan nilai tertinggi pada BLEU-1 (0,499) dan ROUGE-L (0,506), sedangkan Dense dan Hybrid RAG mencatatkan nilai BLEU-4 yang sama (0,288). BM25 RAG konsisten berada pada posisi terendah di ketiga metrik generation, meskipun selisih numerik antar metode pada BLEU-1 dan BLEU-4 tergolong kecil dan tidak dapat digambarkan sebagai keunggulan yang jauh berbeda. Temuan yang lebih penting untuk dibahas adalah keterkaitannya dengan hasil Subbab C.2: BM25 RAG memiliki Final Context Precision tertinggi (0,746), namun keunggulan tersebut tidak diterjemahkan menjadi skor generation tertinggi. Pola ini mengindikasikan bahwa kualitas retrieval, sebagaimana diukur oleh precision konteks akhir, tidak selalu berhubungan linear dengan kualitas generasi jawaban. Konteks dengan kecocokan kata yang tinggi terhadap kueri (karakteristik retrieval leksikal BM25) belum tentu menghasilkan representasi informasi yang paling mendukung pembentukan kalimat jawaban yang mendekati referensi, sedangkan konteks hasil retrieval berbasis representasi semantik (Dense dan, pada tingkat lebih rendah, Hybrid) tampak

mempertahankan hubungan makna yang lebih relevan bagi kebutuhan generator.

Hasil tersebut mengindikasikan bahwa karakteristik konteks hasil retrieval berbasis representasi semantik memberikan dukungan yang lebih baik terhadap kualitas jawaban pada konfigurasi generator yang digunakan, meskipun eksperimen ini mengukur hasil akhir generasi dan bukan secara langsung mekanisme internal LLM dalam memanfaatkan konteks.

Pada metrik Groundedness, Dense RAG mencatatkan nilai tertinggi (0,611), sementara BM25 RAG dan Hybrid RAG memperoleh nilai yang identik (0,595). Gambar 6 menyajikan perbandingan Groundedness ketiga metode.



Gambar 6. Tingkat Keterikatan Konteks (Groundedness)

Dense RAG menghasilkan keterikatan jawaban terhadap konteks yang sedikit lebih tinggi pada konfigurasi eksperimen ini dibandingkan BM25 RAG dan Hybrid RAG. Hybrid belum menunjukkan keunggulan pada metrik Groundedness dibandingkan Dense; nilai Hybrid yang identik dengan BM25 pada metrik ini mengindikasikan bahwa keterikatan jawaban terhadap konteks Hybrid lebih dipengaruhi oleh komponen leksikal yang turut disertakan pada proses fusion, bukan oleh komponen semantik saja.

E. Evaluasi Keamanan Sistem dan Mitigasi Halusinasi

Kemampuan sistem menolak kueri di luar basis pengetahuan dievaluasi menggunakan False Rejection Rate (FRR) dan True Rejection Rate (TRR), dihitung dari 130 kueri In-Domain dan 50 kueri Out-of-Context pada threshold 0,85. Tabel V merangkum hasil ketiga metode.

TABEL V
HASIL EVALUASI FRR DAN TRR (130 KUERI IN-DOMAIN, 50 KUERI OOC)

Metode	FRR	TRR
Dense RAG	11,5%	100,0%
BM25 RAG	13,8%	98,0%
Hybrid RAG	13,1%	100,0%

Dense RAG dan Hybrid RAG berhasil menolak seluruh 50 kueri Out-of-Context yang diuji (TRR 100,0%), sedangkan BM25 RAG menolak 98,0% kueri OOC. Perlu ditekankan bahwa TRR 98,0% pada BM25 RAG menunjukkan hanya sebagian kecil kueri OOC yang tidak berhasil ditolak, bukan indikasi bahwa BM25 secara umum rentan meloloskan kueri di luar domain. Pada metrik FRR, Dense RAG mencatatkan nilai terendah (11,5%), diikuti Hybrid RAG (13,1%) dan BM25 RAG (13,8%). Hybrid RAG dengan demikian bukan metode dengan FRR terendah, tetapi mempertahankan kemampuan penolakan OOC yang sempurna (TRR 100,0%) dengan FRR yang hanya sedikit lebih tinggi daripada Dense, merepresentasikan kompromi antara kedua metrik tersebut.

Evaluasi keamanan sistem tidak dapat disandarkan hanya pada TRR yang tinggi, karena sistem yang terlalu ketat dalam menolak kueri berpotensi menolak pertanyaan valid dan menurunkan kegunaan sistem bagi pengguna. TRR yang tinggi menunjukkan kemampuan mitigasi terhadap risiko

menjawab kueri yang tidak didukung basis pengetahuan, sementara FRR yang rendah menunjukkan sistem tetap dapat melayani sebagian besar kueri valid tanpa penolakan yang berlebihan. Confidence Scorer, sebagai mekanisme mitigasi risiko hallucination berbasis ketersediaan dan kualitas konteks, menunjukkan efektivitas pada Dense RAG dan Hybrid RAG dengan mencapai TRR sempurna, tetapi konfigurasi threshold 0,85 yang digunakan tetap menghasilkan trade-off antara kemampuan menolak kueri tidak valid dan risiko menolak sebagian kueri yang sebenarnya valid, sebagaimana tercermin pada nilai FRR di atas 11% pada ketiga metode.

Tabel VI menyajikan contoh konkret keputusan sistem pada beberapa kueri OOC dan satu kasus borderline In-Domain, beserta skor confidence masing-masing pada threshold $\tau = 0,85$, untuk memberikan bukti kualitatif di balik angka TRR dan FRR agregat pada Tabel V.

TABEL VI
CONTOH EVALUASI KUERI OUT-OF-CONTEXT (OOC) DAN BORDERLINE PADA THRESHOLD $T = 0,85$

Tipe / ID Kueri	Teks Pertanyaan	Skor Confidence	Status Keputusan	Keterangan Evaluasi
OOC-002	“Bagaimana cara membuat rendang Padang yang autentik?”	0,382	Ditolak	Berhasil ditolak (True Rejection). Topik kuliner Minangkabau.
OOC-015	“Siapa penemu teori relativitas khusus dalam fisika?”	0,215	Ditolak	Berhasil ditolak (True Rejection). Topik sains umum di luar budaya Jambi.
OOC-042	“Apa saja fungsi dan struktur rumah adat Rumah Gadang Minangkabau?”	0,765	Ditolak	Berhasil ditolak pada $\tau = 0,85$ (True Rejection). Kueri budaya daerah tetangga yang bersinggungan. Catatan: akan lolos (False Acceptance) jika menggunakan kalibrasi awal $\tau = 0,70$.
ID-045 (Borderline)	“Berapakah jumlah riil penari pada pementasan awal Tari Sekapur Sirih di Jambi?”	0,842	Ditolak	Terjadi False Rejection pada $\tau = 0,85$ karena kueri sangat spesifik, meskipun kueri bersifat In-Domain.

F. Matriks Komparasi Kinerja End-to-End

Untuk merangkum karakteristik kekuatan dan kelemahan masing-masing metode secara end-to-end, Gambar 7 menyajikan matriks komparasi seluruh metrik yang telah dibahas pada subbab-subbab sebelumnya.

Matriks Komparasi Performa End-to-End
(Warna Lebih Gelap = Performa Relatif Lebih Baik)

	Pure Prec	Final Prec	Final Recall	BLEU-4	ROUGE-L	Groundness	FRR
Dense RAG	0.7169	0.7231	0.9385	0.2881	0.5055	0.6107	0.1154
BM25 RAG	0.6646	0.7462	0.9308	0.2804	0.4890	0.5946	0.1385
Hybrid RAG	0.7054	0.7256	0.9385	0.2882	0.4984	0.5951	0.1308

Gambar 7. Matriks Komparasi Performa End-to-End

Matriks pada Gambar 7 menunjukkan bahwa Dense RAG unggul pada Pure Precision, Pure Recall, BLEU-1, ROUGE-L, Groundness, dan FRR terendah; Dense dan Hybrid RAG unggul bersama pada Final Recall, BLEU-4, dan TRR sempurna; sementara BM25 RAG unggul secara spesifik pada Final Context Precision setelah reranking. Warna paling gelap pada matriks tidak merata pada Hybrid RAG, melainkan lebih banyak terkonsentrasi pada baris Dense RAG, yang konsisten dengan pola keunggulan Dense pada mayoritas metrik individual yang telah dibahas pada Subbab C hingga E.

Berdasarkan hasil eksperimen ini, Dense RAG merupakan metode dengan performa paling dominan pada sebagian besar metrik yang diuji. BM25 RAG memiliki kekuatan spesifik pada Final Context Precision setelah proses reranking, namun tidak unggul pada metrik lain. Hybrid RAG menunjukkan performa yang kompetitif dan relatif seimbang, terutama dengan Final Recall yang setara Dense, BLEU-4 yang setara Dense, serta TRR 100%, tetapi tidak dapat diklaim sebagai metode terbaik karena Dense unggul pada lebih banyak

metrik individual. Kontribusi Hybrid RAG pada penelitian ini lebih tepat digambarkan sebagai strategi kompromi antara sinyal semantik dan leksikal, meskipun pada konfigurasi dan dataset penelitian ini keuntungan kombinasi tersebut belum secara konsisten melampaui Dense RAG.

Beberapa kemungkinan penjelasan dapat diajukan terhadap pola ini, dengan catatan bahwa penjelasan berikut bersifat interpretatif dan belum tentu terbukti langsung oleh eksperimen: fusion berbasis RRF yang digunakan pada konfigurasi ini diduga belum menghasilkan complementary gain yang maksimal; candidate pool gabungan pada Hybrid berpotensi membawa noise dari komponen BM25 yang precision mentahnya lebih rendah; bobot fusion yang digunakan kemungkinan belum optimal secara universal untuk seluruh jenis kueri; karakteristik korpus dan kueri Budaya Jambi pada penelitian ini tampak lebih sesuai dengan pendekatan semantic retrieval murni; dan BM25 dengan Dense kemungkinan memiliki tingkat overlap kandidat yang cukup tinggi sehingga manfaat kombinasi keduanya tidak sebesar yang diharapkan secara teoretis. Perbedaan ini perlu dibedakan secara jelas antara hasil yang terbukti dalam eksperimen (keunggulan Dense pada mayoritas metrik) dan kemungkinan penjelasan terhadap hasil tersebut (daftar dugaan penyebab di atas), yang masih memerlukan investigasi lebih lanjut untuk divalidasi.

Perlu ditegaskan bahwa perbandingan Dense RAG, BM25 RAG, dan Hybrid RAG pada penelitian ini merupakan perbandingan antar-strategi retrieval, bukan studi ablasi komponen secara terisolasi (misalnya reranking on/off, Confidence Scorer on/off, atau fusion on/off pada konfigurasi retrieval yang sama). Kontribusi reranking terhadap masing-masing strategi memang dianalisis terpisah pada Subbab III.C, namun kontribusi Confidence Scorer terhadap kualitas generasi belum diisolasi dari komponen retrieval. Studi ablasi komponen yang lebih granular — misalnya menonaktifkan Confidence Scorer pada konfigurasi Hybrid RAG untuk mengamati perubahan Groundedness dan tingkat halusinasi — menjadi agenda penelitian lanjutan yang penting untuk memastikan atribusi kontribusi tiap komponen secara lebih presisi.

G. Analisis Kualitatif Perbandingan Jawaban

Untuk melengkapi pembuktian metrik kuantitatif, analisis kualitatif dilakukan dengan membandingkan kualitas konteks yang diperoleh, relevansi informasi terhadap pertanyaan, kemampuan jawaban menjawab inti pertanyaan, serta potensi informasi yang tidak didukung konteks, pada empat kategori studi kasus yang merepresentasikan pola keberhasilan dan keterbatasan masing-masing metode. Tabel VII menyajikan ringkasan keempat studi kasus tersebut.

TABEL VII
STUDI KASUS PERBANDINGAN JAWABAN ANTAR METODE

Kategori	Metode	Pengamatan
Kasus Dense unggul	Dense RAG	Pada kueri ID-002 mengenai perkembangan musik Gambus, konteks hasil dense retrieval berhasil menautkan kueri dengan chunk yang relevan meski redaksi kueri berbeda dari redaksi dokumen sumber, menghasilkan jawaban yang tepat sasaran.
Kasus BM25 unggul	BM25 RAG	Pada kueri dengan istilah eksak yang persis tercantum di dokumen sumber (misalnya nama entitas atau angka spesifik), pencocokan leksikal BM25 langsung menempatkan dokumen sumber pada peringkat atas Top-3 tanpa memerlukan interpretasi semantik tambahan.
Kasus Hybrid memberi keuntungan kombinasi	Hybrid RAG	Pada kueri yang mengandung sinonim sekaligus istilah eksak, kandidat gabungan dari komponen leksikal dan semantik saling melengkapi pada Final Context, menghasilkan Final Recall yang setara Dense meski Precision mentahnya sedikit tertinggal.
Kasus kegagalan/keterbatasan	Ketiga metode	Pada kueri dengan Ground Truth yang sangat spesifik (angka jumlah tiang, urutan ragam gerak tari), ketiga metode kadang menghasilkan skor BLEU/ROUGE rendah karena jawaban benar secara substansi namun berbeda formulasi kalimat dari referensi, menegaskan keterbatasan metrik overlap teks dalam menangkap kebenaran faktual.

Selain empat kategori pada Tabel VII, ditemukan pula kasus di mana konteks yang diambil berasal dari dua dokumen sumber dengan redaksi yang sedikit kontradiktif mengenai tahun penetapan pergeseran struktur kepemimpinan Lembaga Adat Melayu Jambi, sehingga jawaban yang dihasilkan cenderung mengikuti dokumen dengan skor reranker lebih tinggi tanpa memberi sinyal ketidakpastian kepada pengguna. Kasus lain ditemukan pada kueri yang menggunakan istilah regional Kerinci yang jarang muncul di korpus (misalnya istilah “Pengasuh Sko” atau “Titin Teras”), di mana retrieval

leksikal (BM25) gagal menemukan padanan karena tidak ada kecocokan kata kunci eksak, sementara retrieval semantik (Dense) tetap berhasil menautkannya ke konsep yang relevan.

Studi kasus di atas menunjukkan bahwa keunggulan kuantitatif satu metode pada suatu metrik tidak selalu konsisten pada setiap kueri individual: Dense yang dominan secara agregat tetap dapat kalah pada kueri dengan istilah eksak yang sangat spesifik, sementara BM25 yang tertinggal secara agregat tetap unggul pada tipe kueri tertentu. Pengamatan ini konsisten dengan temuan agregat pada

Subbab C hingga E, dan menegaskan bahwa pemilihan strategi retrieval idealnya mempertimbangkan karakteristik distribusi kueri yang akan dihadapi sistem pada penggunaan nyata, bukan hanya nilai rata-rata pada satu dataset uji.

H. Kesimpulan Konfigurasi dan Keterbatasan Penelitian

Berdasarkan keseluruhan hasil evaluasi pada 130 kueri Test Set In-Domain dan 50 kueri Out-of-Context, tidak terdapat satu metode yang menang mutlak pada seluruh metrik yang diuji. Dense RAG memberikan performa paling dominan pada konfigurasi eksperimen saat ini, unggul pada Pure Precision, Pure Recall, BLEU-1, ROUGE-L, Groundedness, dan FRR terendah. Hybrid RAG memberikan performa kompetitif dengan keseimbangan tertentu, khususnya pada Final Recall dan BLEU-4 yang setara Dense serta TRR sempurna, tetapi belum menunjukkan superioritas yang konsisten terhadap Dense RAG. BM25 RAG tetap efektif pada skenario tertentu, khususnya Final Context Precision setelah reranking, meskipun tertinggal pada sebagian besar metrik lain. Confidence Scorer pada threshold 0,85 berhasil mencapai TRR 100% pada Dense RAG dan Hybrid RAG, tetapi tetap memiliki trade-off FRR pada kisaran 11-14% di ketiga metode.

Penelitian ini mengidentifikasi tujuh keterbatasan yang perlu menjadi perhatian pada pengembangan selanjutnya. Pertama, evaluasi dilakukan pada domain Budaya Jambi sehingga generalisasi hasil ke domain budaya lokal lain memerlukan validasi lebih lanjut. Kedua, ukuran dataset evaluasi (130 kueri In-Domain, 50 kueri Out-of-Context) masih tergolong terbatas untuk generalisasi statistik yang kuat. Ketiga, Development Set yang digunakan pada tahap kalibrasi hanya terdiri dari 20 kueri In-Domain sehingga ruang eksplorasi hyperparameter, termasuk bobot fusion RRF dan konfigurasi Confidence Scorer, masih terbatas. Keempat, evaluasi generation pada penelitian ini menggunakan metrik berbasis overlap teks (BLEU, ROUGE-L) yang memiliki keterbatasan dalam menangkap seluruh aspek kualitas jawaban, sebagaimana ditunjukkan pada studi kasus keempat di Subbab G. Kelima, definisi dan mekanisme perhitungan Groundedness yang digunakan perlu dijelaskan secara eksplisit sesuai implementasi penelitian agar dapat direplikasi dan dibandingkan dengan studi lain. Keenam, hasil Hybrid RAG yang belum melampaui Dense RAG secara konsisten dapat dipengaruhi oleh strategi fusion, bobot komponen, ukuran candidate pool, dan konfigurasi reranker yang digunakan pada penelitian ini, sehingga eksplorasi konfigurasi fusion yang lebih sistematis menjadi agenda penelitian lanjutan. Ketujuh, kapabilitas model generatif Qwen2.5-7B sebagai model berskala relatif kecil membatasi kemampuan penalaran pada kueri yang memerlukan sintesis informasi lintas-dokumen yang kompleks, dan kebutuhan komputasi (GPU dengan VRAM minimal 12 GB) untuk menjalankan Qwen2.5-7B, BGE-Reranker-Large, dan Multilingual-E5 secara bersamaan perlu dipertimbangkan pada implementasi skala produksi dengan volume kueri tinggi.

IV. KESIMPULAN

Penelitian ini berhasil mengimplementasikan dan mengevaluasi arsitektur Retrieval-Augmented Generation (RAG) menggunakan model Qwen2.5-7B untuk sistem tanya jawab informasi Budaya Jambi. Berdasarkan evaluasi komparatif end-to-end pada 130 kueri In-Domain (ID) dan 50 kueri Out-of-Context (OOC), arsitektur Dense RAG secara empiris menunjukkan performa paling dominan, mencatatkan keunggulan pada metrik penarikan awal (Pure Recall@10 sebesar 0,992), kualitas generasi teks (BLEU-4 sebesar 0,288 dan Groundedness 0,611), serta tingkat False Rejection Rate (FRR) terendah (11,5%). Di sisi lain, BM25 RAG memiliki kekuatan spesifik pada tingkat presisi konteks akhir pasca-reranking (0,746), sementara Hybrid RAG menawarkan performa kompromi dengan Final Recall (0,938) dan BLEU-4 (0,288) yang setara dengan Dense RAG.

Integrasi mekanisme keamanan ganda melalui Confidence Scorer (dengan ambang batas 0,85) dan strict grounding prompt terbukti sangat krusial dan efektif dalam memitigasi halusinasi. Mekanisme ini berhasil mencapai True Rejection Rate (TRR) 100% pada arsitektur Dense dan Hybrid untuk seluruh 50 kueri Out-of-Context (OOC) yang diuji, menunjukkan efektivitas dual-threshold dalam mencegah sistem menjawab pertanyaan di luar cakupan basis pengetahuan pada skala pengujian ini. Klaim ini perlu dibatasi pada karakteristik dan ukuran set OOC yang digunakan (50 kueri, mencakup topik teknologi dan geografi global); generalisasi terhadap kueri OOC yang lebih beragam atau lebih ambigu secara tematik masih memerlukan pengujian lebih lanjut (lihat Tabel VI). Secara operasional, keberhasilan ini juga diiringi trade-off berupa penolakan sebagian kecil kueri valid (FRR pada kisaran 11,5%-13,8%). Secara keseluruhan, temuan ini mengindikasikan bahwa penggunaan Dense retrieval murni dipadukan dengan Confidence Scorer memberikan dukungan konteks dan keamanan paling konsisten pada dataset Budaya Jambi yang diuji.

Implikasi praktis dari keberhasilan arsitektur ini terbuka untuk pengembangan lebih lanjut sebagai asisten virtual edukatif pada portal pemerintah daerah atau museum kebudayaan lokal, dengan syarat validasi lebih luas sebagaimana dicatat pada keterbatasan penelitian ini. Untuk pengembangan pada penelitian selanjutnya, disarankan adanya eksperimen kalibrasi lanjutan pada komponen Confidence Scorer secara dinamis, atau penerapan fine-tuning spesifik-domain pada model embedding (Multilingual-E5), guna menekan angka False Rejection Rate (FRR) pada kueri In-Domain yang sangat spesifik tanpa mengorbankan pertahanan absolut terhadap kueri di luar domain. Selain itu, eksplorasi perluasan skala basis pengetahuan (knowledge base) yang mencakup literatur lintas budaya daerah lain di Indonesia, serta pengujian menggunakan pendekatan agen mandiri (Agentic RAG) untuk verifikasi silang dokumen, direkomendasikan untuk mengevaluasi konsistensi dan skalabilitas arsitektur ini pada tingkat implementasi yang lebih masif.

DAFTAR PUSTAKA

- [1] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, 2020.
- [2] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, 2023, doi: 10.1145/3571730.
- [3] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, 2025, doi: 10.1145/3703155.
- [4] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625-630, 2024, doi: 10.1038/s41586-024-07421-0.
- [5] BPSDM Provinsi Jambi, "Warisan Budaya dan Kekayaan Tradisi Suku Jambi," Portal BPSDM Provinsi Jambi. [Online]. Available: <https://portalbpsdm.jambiprov.go.id/artikel/buku-sastra/warisan-budaya-dan-kekayaan-tradisi-suku-jambi/>
- [6] S. Robertson and H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond*, vol. 3, no. 4. 2009. doi: 10.1561/1500000019.
- [7] W. Fan et al., "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 6491-6501, 2024, doi: 10.1145/3637528.3671470.
- [8] C. C. Chang, H. P. Chang, and H. S. Lee, "Leveraging Retrieval-Augmented Generation for Culturally Inclusive Hakka Chatbots: Design Insights and User Perceptions," *RASSE 2024 - 2024 IEEE Int. Conf. Recent Adv. Syst. Sci. Eng. Proc.*, 2024, doi: 10.1109/RASSE64357.2024.10773731.
- [9] L. Xu, L. Lu, M. Liu, C. Song, and L. Wu, "Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology," *Herit. Sci.*, vol. 12, no. 1, pp. 1-24, 2024, doi: 10.1186/s40494-024-01231-3.
- [10] A. Bora and H. Cuayahuitl, "Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 4, pp. 2355-2374, 2024, doi: 10.3390/make6040116.
- [11] Z. Zhao and D. Wang, "Evaluation of large language models for the intangible cultural heritage domain," *npj Herit. Sci.*, vol. 13, no. 1, pp. 1-11, 2025, doi: 10.1038/s40494-025-02013-1.
- [12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA, 2002, pp. 311-318. doi: 10.3115/1073083.1073135.
- [13] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," *EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 6769-6781, 2020, doi: 10.18653/v1/2020.emnlp-main.550.
- [14] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," *Proc. - 32nd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR 2009*, pp. 758-759, 2009, doi: 10.1145/1571941.1572114.
- [15] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J. Y. Nie, *C-Pack: Packed Resources For General Chinese Embeddings*, vol. 1, no. 1. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657878.
- [16] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," *SIGIR 2020 - Proc. 43rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 39-48, 2020, doi: 10.1145/3397271.3401075.
- [17] A. Kamath, R. Jia, and P. Liang, "Selective question answering under domain shift," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 5684-5696, 2020, doi: 10.18653/v1/2020.acl-main.503.
- [18] Qwen Team, "Qwen2.5: A Party of Foundation Models," *Qwen Blog*, Sep. 19, 2024. [Online]. Available: <https://qwenlm.github.io/blog/qwen2.5/>. [Accessed: Sep. 04, 2026].
- [19] C.-Y. Lin, "A Package for Automatic Evaluation of Summaries," *Text Summ. Branches Out*, pp. 74-81, 2004.
- [20] C. C. Chang, C. F. Li, C. H. Lee, and H. S. Lee, "Enhancing Low-Resource Minority Language Translation with LLMs and Retrieval-Augmented Generation for Cultural Nuances," in *Intelligent Systems and Applications (IntelliSys 2025)*, *Lecture Notes in Networks and Systems*, vol. 1567, Springer, 2025.
- [21] E. V. Ndimbo, Q. Luo, G. C. Fernando, X. Yang, and B. Wang, "Leveraging Retrieval-Augmented Generation for Swahili Language Conversation Systems," *Applied Sciences*, vol. 15, no. 2, p. 524, 2025, doi: 10.3390/app15020524.