Jurnal Pengembangan Teknologi Informasi dan Komunikasi

JUPTIK

2026 Vol. 4 No.1 Hal:123-129 e-ISSN: 2987-7016

Chatbot Hybrid Fatwa MUI Menggunakan Retrieval Augmented Generation dan Large Language Model

Surya Hidayatullah Firdaus [1], Nazruddin Safaat H [2], Yelfi Vitriani [3], Novriyanto [4]

[1,2,3,4] Fakultas Sains dan Teknologi, Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim, Pekanbaru, Indonesia

[1] 12250111759@students.uin-suska.ac.id, [2] nazruddin.safaat@uin-suska.ac.id*, [3] yelfi.vitriani@ uin-suska.ac.id, [4]novriyanto@uin-suska.ac.id

Abstract

The fragmented accessibility of digital Indonesian Ulema Council (MUI) fatwas makes conventional information retrieval less effective. Furthermore, single-corpus AI question-answering systems are prone to generating inaccurate responses (hallucinations) for out-of-domain queries. This study develops a Hybrid Fatwa MUI Chatbot using a tiered Hybrid Retrieval architecture integrating two knowledge sources: MUI Fatwa documents as the primary corpus and 12,370 Bukhari-Muslim hadiths as a fallback mechanism. The system implements semantic search, automatic topic verification by a language model, and dynamic routing to the hadith database if the fatwa context is deemed irrelevant. Evaluation results reveal a 13.23% improvement in answer semantic similarity (from 0.6664 to 0.7546) and a 10.57% increase in faithfulness to reference sources (from 85.37% to 94.39%), with an identical abstain rate of 26.83%. This multi-corpus approach is proven to significantly enhance response relevance and accuracy compared to standard single-source RAG systems.

Keywords: *chatbot hybrid, MUI fatwa, retrieval-augmented generation (RAG), large language model (LLM), hybrid retrieval*

Abstrak

Aksesibilitas dokumen digital Fatwa Majelis Ulama Indonesia (MUI) yang terfragmentasi membuat pencarian informasi kurang efektif. Di sisi lain, sistem tanya jawab AI berbasis satu sumber dokumen (*single-corpus*) rentan menghasilkan jawaban tidak akurat (*halusinasi*) pada pertanyaan di luar domain. Penelitian ini mengembangkan Chatbot Hybrid Fatwa MUI menggunakan arsitektur *Hybrid Retrieval* bertingkat dengan dua sumber pengetahuan: dokumen Fatwa MUI sebagai korpus utama dan 12.370 hadis Bukhari-Muslim sebagai mekanisme cadangan (*fallback*). Sistem ini menerapkan pencarian semantik, verifikasi topik otomatis oleh model bahasa, dan pengalihan ke basis data hadis jika konteks fatwa dinilai tidak relevan. Hasil evaluasi menunjukkan peningkatan kesamaan makna jawaban sebesar 13,23% (dari 0,6664 menjadi 0,7546) dan peningkatan kesetiaan pada rujukan (*faithfulness*) sebesar 10,57% (dari 85,37% menjadi 94,39%), dengan tingkat penolakan (*abstain rate*) identik sebesar 26,83%. Pendekatan multi-korpus ini terbukti signifikan meningkatkan relevansi dan keakuratan jawaban dibandingkan RAG standar.

Kata kunci: chatbot hybrid, fatwa MUI, retrieval-augmented generation (RAG), large language model (LLM), hybrid retrieval

1. Pendahuluan

Perkembangan teknologi kecerdasan buatan atau *Artificial Intelligence* (AI) mengalami akselerasi yang sangat masif dalam beberapa tahun terakhir, khususnya pada ranah pemrosesan bahasa alami (*Natural Language Processing*/NLP). Evolusi model bahasa besar (*Large Language Model* atau LLM) telah mengubah paradigma interaksi antara manusia dan mesin [1]. Kemampuan AI generatif dalam memproses kueri yang kompleks menjadikan teknologi ini mampu bertindak sebagai asisten pencarian informasi cerdas, yang umumnya diwujudkan dalam bentuk *chatbot*. Xu *et al.* [2] mendefinisikan *chatbot* sebagai sistem berbasis AI yang mampu berinteraksi secara alami dengan pengguna dalam berbagai domain layanan. Lebih lanjut, Ardimansyah dan Widianto [3] membuktikan bahwa *platform* pesan instan yang diintegrasikan dengan AI dapat menjadi medium penyampaian informasi yang interaktif dan aksesibel.

Dalam konteks demografi masyarakat Muslim di Indonesia, kebutuhan terhadap kepastian hukum syar'i senantiasa membutuhkan rujukan yang otoritatif, yakni dokumen fatwa Majelis Ulama Indonesia (MUI). Fatwa MUI memiliki otoritas moral, sosial, dan keagamaan yang kuat sebagai panduan bagi umat Islam [4]. Kajian Hasanuddin *et al.*[4] dan Shuhufi *et al.* [5] memperkuat pandangan bahwa fatwa MUI memberikan kontribusi strategis dalam membentuk tata kelola kebijakan berbasis syariah yang responsif terhadap dinamika era modern. Namun, aksesibilitas terhadap dokumen fatwa digital yang umumnya berformat PDF masih tersebar

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

JUPTIK

2026 Vol. 4 No.1 Hal:123-129 e-ISSN: 2987-7016

tanpa satu sistem pencarian terpusat. Akibatnya, proses penelusuran manual membutuhkan waktu lama serta pemahaman konteks hukum yang umumnya tidak dimiliki oleh masyarakat awam.

Penerapan AI untuk memfasilitasi akses informasi hukum Islam ini memunculkan perdebatan akademis krusial. Dewi [6] mengidentifikasi bahwa AI tidak dapat menggantikan peran ulama dalam berijtihad karena keterbatasan dalam memahami *maqashid syariah* secara holistik. Sejalan dengan rekomendasi Putra [7] mengenai pentingnya pembatasan peran AI, sistem yang dikembangkan dalam penelitian ini secara ketat memposisikan AI murni sebagai fasilitator akses informasi tekstual fatwa, bukan sebagai entitas pemberi fatwa. Di sisi lain, adopsi teknologi AI generatif memiliki risiko fatal terkait fenomena halusinasi model bahasa (*AI hallucination*) [8]. Untuk mengatasi kendala tersebut, arsitektur *Retrieval-Augmented Generation* (RAG) secara komprehensif telah terbukti menjadi solusi yang tangguh dengan memisahkan pengetahuan faktual dari parameter pelatihan LLM, sehingga model dapat menggabungkan kemampuan generatifnya dengan fleksibilitas pencarian informasi eksternal guna meminimalkan halusinasi [9]. Pendekatan lebih lanjut seperti kerangka *Self-RAG* juga membuktikan bahwa mekanisme pengalihan pencarian adaptif dan evaluasi mandiri (*self-reflection*) sangat krusial untuk meningkatkan keakuratan dan kesetiaan jawaban pada dokumen sumber [10].

Pada domain sistem informasi publik, integrasi RAG terbukti mumpuni. Fauzi dan Sutabri [11] berhasil merancang *PublicTalk* dengan akurasi pemahaman teks di atas 88%, sementara Cahyanti dan Raya [12] mengembangkan *Chatku AI* yang efektif memproses data eksternal PDF dengan skor penerimaan pengguna 99,2%. Dalam ranah literatur keislaman, Rahayu *et al.* [13] mengembangkan sistem tanya jawab Fikih Empat Madzhab berbasis LangChain pada dataset PDF. Sejalan dengan itu, Helviansyah *et al.* [14] memanfaatkan LLM berbasis *similarity search* untuk data fiqih kontemporer. Pencarian dalil primer juga dieksplorasi oleh Herwanza *et al.* [15] untuk mengekstraksi hadis, serta Haekal *et al.* [16] yang merancang sistem tanya jawab *closed-domain* terhadap dokumen fatwa menggunakan RAG konvensional. Lebih lanjut, evaluasi komprehensif LLM pada domain penalaran hukum Islam (seperti ilmu mawaris) menunjukkan bahwa model standar rentan mengalami kegagalan nalar fundamental dan salah interpretasi sumber, sehingga memvalidasi pentingnya integrasi sumber eksternal yang solid [17].

Untuk memastikan sistem AI bekerja sesuai rujukan faktualnya, evaluasi kualitas sistem tanya jawab tidak cukup hanya dengan pencocokan kata. Machado *et al.*[18] menjelaskan bahwa metrik *BERTScore* sangat efektif untuk evaluasi teks karena mampu memanfaatkan representasi kontekstual untuk menangkap kesamaan makna secara semantik. Di samping itu, Dobslaw *et al.* [19] memperkenalkan *Giskard* sebagai *framework* evaluasi yang dirancang khusus untuk menguji aplikasi berbasis LLM, termasuk pengujian *faithfulness* (kesetiaan pada dokumen sumber) dan deteksi halusinasi.

Meskipun penelitian terdahulu telah meletakkan fondasi teknis yang solid, terdapat kelemahan konseptual mendasar, yakni tingginya ketergantungan pada korpus dokumen tunggal (*single corpus dependency*). Karakteristik fatwa MUI pada dasarnya bersifat insidentil dan berbasis penyelesaian masalah (*problem-based*), di mana dokumen tidak dirancang untuk mencakup seluruh aspek hukum Islam. Konsekuensinya, jika pengguna menanyakan persoalan di luar cakupan dokumen PDF fatwa yang tersedia, sistem RAG standar berisiko mengalami *out-of-domain query* dan memaksakan pencocokan konteks yang salah.

Berdasarkan kesenjangan literatur (*research gap*) tersebut, penelitian ini bertujuan mengembangkan Chatbot Hybrid Fatwa MUI dengan menerapkan arsitektur *Hybrid Retrieval*. Kebaruan (*novelty*) utama dalam sistem ini adalah hadirnya mekanisme *fallback* pencarian berjenjang: sistem memprioritaskan pencarian semantik pada dokumen Fatwa MUI, namun apabila verifikasi LLM menilai tidak ada konteks fatwa yang relevan, sistem secara otomatis mengalihkan penelusuran leksikal ke pangkalan data sekunder berupa 12.370 hadis sahih terkurasi. Penelitian ini memberikan kontribusi teoretis berupa rancangan arsitektur integrasi RAG multi-korpus, sekaligus berkontribusi praktis dalam penyediaan sistem layanan informasi hukum Islam digital yang meminimalkan halusinasi dan dapat dipertanggungjawabkan secara syar'i.

2. Metode Penelitian

2.1. Kerangka Dasar Penelitian

Penelitian ini merupakan penelitian pada bidang *engineering* yang berfokus pada pengembangan sistem (*system development*) serta pengujian performa model dalam konteks sistem tanya jawab berbasis kecerdasan buatan. Pendekatan yang digunakan adalah eksperimen komparatif, yaitu membandingkan kinerja sistem *Retrieval-Augmented Generation* (RAG) standar dengan sistem yang diusulkan berupa *Hybrid Retrieval* yang mengintegrasikan dua sumber pengetahuan: dokumen Fatwa MUI sebagai korpus utama dan basis data hadis sebagai mekanisme *fallback*. Penelitian ini tidak melibatkan responden karena berorientasi pada pengujian sistem berbasis data dan evaluasi otomatis.

Hipotesis penelitian ini adalah bahwa penerapan metode *Hybrid Retrieval* dengan mekanisme pencarian berjenjang mampu meningkatkan kualitas jawaban sistem dibandingkan dengan pendekatan RAG standar berbasis *single corpus*. Variabel independen adalah metode retrieval yang digunakan, yaitu RAG standar dan *Hybrid Retrieval* (multi-korpus dengan mekanisme *fallback*). Variabel dependen adalah kualitas jawaban

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal:123-129 e-ISSN: 2987-7016

yang diukur menggunakan BERTScore (precision, recall, dan F1-Score), tingkat *faithfulness*, serta potensi halusinasi yang dievaluasi menggunakan framework Giskard.

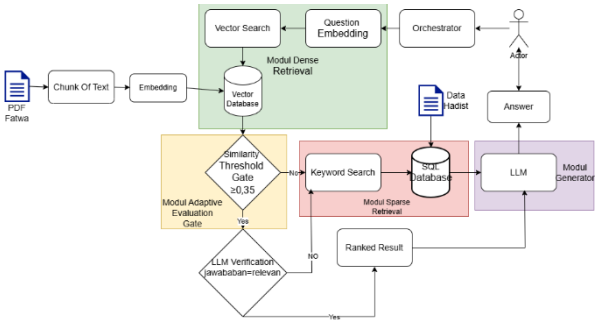
Kerangka pemikiran penelitian ini didasarkan pada permasalahan keterbatasan cakupan dokumen fatwa yang bersifat insidentil (*problem-based*), sehingga tidak mampu menjawab seluruh pertanyaan pengguna. Untuk mengatasi hal tersebut, diusulkan pendekatan *Hybrid Retrieval* yang bekerja secara berjenjang, dimulai dari pencarian pada korpus Fatwa MUI menggunakan *semantic similarity*, dilanjutkan dengan proses verifikasi kesesuaian topik menggunakan *Large Language Model*. Apabila tidak ditemukan konteks yang relevan, sistem secara otomatis mengalihkan pencarian ke basis data hadis sebagai sumber pengetahuan sekunder.

2.2. Tahapan Penelitian

Penelitian ini terdiri dari enam tahap. Tahap pertama adalah analisis masalah untuk mengidentifikasi kendala utama dalam pemanfaatan dokumen Fatwa MUI, meliputi struktur dokumen PDF yang tidak terorganisir (*unstructured data*) serta keterbatasan cakupan fatwa yang bersifat insidentil. Tahap kedua adalah studi literatur dengan mengkaji berbagai penelitian terkait sistem tanya jawab berbasis LLM, RAG, *Hybrid Retrieval*, serta teknik evaluasi seperti BERTScore dan Giskard. Tahap ketiga adalah pengumpulan data, yaitu dokumen Fatwa MUI dalam format PDF yang diunduh melalui www.fatwamui.com/data-fatwa sebagai korpus utama, dan 7.008 hadis riwayat Bukhari serta 5.362 hadis riwayat Muslim dari repositori publik GitHub https://github.com/irsyadulibad sebagai korpus sekunder. Tahap keempat adalah analisis dan perancangan arsitektur *Hybrid Retrieval* yang mencakup alur *embedding*, *vector storage*, *semantic search*, dan mekanisme *fallback*. Tahap kelima adalah implementasi sistem chatbot berbasis web. Tahap keenam adalah pengujian dan evaluasi menggunakan BERTScore dan framework Giskard.

2.3. Analisa dan Perancangan Sistem

Sistem ini menempatkan Large Language Model (LLM) sebagai mesin penalaran (*reasoning engine*) sekaligus komponen evaluatif dalam pengendalian relevansi konteks. Model Llama 3.1 8B dipilih sebagai komponen LLM utama karena kapabilitas pemahaman lintas bahasa (*multilingual understanding*) yang unggul, khususnya dalam memproses teks Bahasa Indonesia dan Arab yang dominan dalam domain keilmuan Islam. Pemilihan ini juga didasari oleh arsitektur keluarga model Llama 3 yang secara bawaan (*natively*) telah dioptimalkan untuk mendukung penalaran tingkat lanjut (*reasoning*) dan penggunaan alat bantu (*tool usage*) dengan kualitas yang sangat kompetitif dibandingkan model *state-of-the-art* lainnya [20]. Kapabilitas penalaran yang tinggi ini menjadikannya sangat ideal dan adaptif untuk diimplementasikan sebagai mesin evaluatif (*Adaptive Decision Gate*) yang bertugas memverifikasi kesesuaian topik secara semantik.



Gambar 1. Arsitektur Sistem Chatbot Hybrid Fatwa MUI

Integrasi sistem diperkuat oleh kerangka kerja *LangChain* yang berfungsi sebagai orchestrator dalam arsitektur modular. Secara spesifik, sistem menerapkan strategi penyimpanan hibrida (*hybrid storage strategy*). Dokumen PDF Fatwa MUI (yang dipra-pemrosesan menggunakan metode *Recursive Character Text Splitting* dengan overlap antar-segmen) disimpan dalam Vector Store untuk mendukung pencarian semantik (dense retrieval). Sementara itu, data Hadis dikelola dalam RDBMS MySQL yang dioptimalkan dengan indeks teks penuh (full-text index) untuk mendukung pencarian leksikal (sparse retrieval).

Berbeda dengan RAG konvensional, perancangan logika sistem ini menggunakan mekanisme Sequential Hybrid Retrieval dengan pengambilan keputusan bertingkat (multi-level decision making) melalui beberapa tahapan kritis. Tahapan kritis pada sistem ini diawali dengan tahap pertama yaitu primary retrieval (pencarian semantik), di mana sistem melakukan dense retrieval pada koleksi vektor Fatwa MUI yang hasilnya melalui penyaringan awal menggunakan evaluasi ambang batas relevansi (similarity threshold). Selanjutnya pada tahap kedua diterapkan LLM-based relevance verification sebagai gerbang keputusan, di mana potongan dokumen fatwa tidak langsung digeneralisasi menjadi jawaban, melainkan komponen adaptive decision gate menggunakan LLM untuk memverifikasi secara semantik apakah topik fatwa benar-benar sesuai dengan intensi kueri pengguna. Terakhir pada tahap ketiga diaktifkan dynamic routing dan fallback mechanism; apabila verifikasi menghasilkan konteks yang relevan maka jawaban akan disintesis, namun apabila LLM menilai konteks fatwa tidak memadai atau tidak relevan, sistem akan secara otomatis mengaktifkan mekanisme fallback dengan mengalihkan rute pencarian (routing) menggunakan sparse retrieval berbasis kata kunci pada basis data Hadis sekunder.

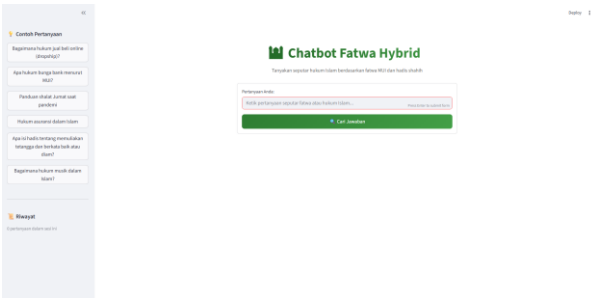
2.4. Implementasi Sistem

Antarmuka sistem diimplementasikan menggunakan framework *Streamlit*, yang berfungsi sebagai

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal:123-129 e-ISSN: 2987-7016

penghubung antara pengguna dan sistem chatbot. Tampilan utama aplikasi terdiri atas panel navigasi di sisi kiri dan panel interaksi utama di bagian tengah. Panel navigasi memuat Contoh Pertanyaan (enam topik representatif: jual beli *online*, hukum bunga bank, panduan shalat Jumat pada masa pandemi, hukum asuransi, tata cara wudhu, dan hukum musik dalam Islam) serta Riwayat percakapan. Panel interaksi utama menyediakan kolom teks masukan dan tombol "Cari Jawaban" yang memicu proses *retrieval* dan generasi jawaban secara *real time*. Desain antarmuka yang minimalis dan berwarna hijau dipilih untuk mencerminkan identitas visual yang bernuansa Islami.



Gambar 2. Antarmuka Pengguna Chatbot Hybrid Fatwa MUI Berbasis Streamlit

Implementasi dilakukan pada perangkat keras dengan spesifikasi: Processor Intel® Core™ i5-13450HX, RAM 20GB, dan SSD 512GB, dengan sistem operasi Windows 11 dan bahasa pemrograman Python. Sistem menggunakan model LLM Llama 3.1 8B yang dijalankan secara lokal melalui runtime Ollama, sehingga beroperasi independen tanpa memerlukan API Key dari pihak ketiga, menghindari batasan kuota token, dan menjamin privasi data pengguna. Parameter *temperature default* (0.8) dipilih agar model memiliki keluwesan linguistik yang cukup untuk merangkum potongan-potongan dokumen Fatwa dan Hadis menjadi jawaban yang natural. *Context window* ditetapkan sebesar 2048 token guna menyeimbangkan efisiensi penggunaan memori pada perangkat keras lokal dengan kecukupan kapasitas untuk menampung *system prompt*, kueri pengguna, serta dokumen hasil *retrieval*, sekaligus mencegah penurunan kecepatan inferensi dan hilangnya fokus informasi (*lost in the middle*). Model embedding all-MiniLM-L6-v2 dengan dimensi vektor 384 dipilih karena keseimbangan optimal antara kecepatan komputasi dan kualitas representasi semantik.

Dense Retriever menggunakan FAISS dengan indeks *IndexFlatIP* untuk pencarian semantik berbasis *cosine similarity*. *Sparse Retriever* memanfaatkan MySQL *Full-Text Search* pada basis data Hadis (7.008 hadis Bukhari dan 5.362 hadis Muslim). Mekanisme *routing* dikendalikan oleh nilai *SIMILARITY_THRESHOLD = 0.35*: apabila skor kemiripan *dense retrieval* tidak melampaui ambang batas tersebut, sistem mengalihkan pencarian ke jalur *sparse retrieval*. Selain itu, sistem menerapkan *keyword bypass* berupa *hard-rule* yang langsung mengaktifkan jalur *sparse retrieval* untuk pertanyaan yang secara eksplisit merujuk pada sumber Hadis.

2.5. Pengujian Sistem

Tahap pengujian dirancang untuk membandingkan kualitas keluaran jawaban antara dua konfigurasi sistem: sistem *Chatbot Hybrid* (Fatwa MUI + *fallback* Hadis) dan sistem RAG standar berbasis korpus tunggal. Pengujian difokuskan pada evaluasi *end-to-end*, yaitu menilai kualitas jawaban yang dihasilkan secara keseluruhan. Seluruh komponen lain seperti model LLM, konfigurasi *prompt*, parameter *temperature*, dan antarmuka sistem dibuat identik pada kedua konfigurasi.

Untuk mengakomodasi karakteristik metrik pengujian yang berbeda, penelitian ini menggunakan dua kelompok dataset yang terpisah. Pertama, evaluasi relevansi jawaban dilakukan menggunakan metrik BERTScore pada 10 sampel pertanyaan spesifik. Evaluasi ini secara khusus menggunakan anotasi jawaban referensi (*human-annotated gold standard*) yang disusun secara manual untuk mengukur kedalaman kesamaan semantik antara jawaban sistem (*hypothesis*) dan jawaban referensi (*reference*), guna menghasilkan nilai *Precision*, *Recall*, dan *F1-Score*. Kedua, evaluasi keandalan faktual dan potensi halusinasi dilakukan menggunakan *framework* Giskard dengan RAGET (Retrieval Augmented Generation Evaluation Toolkit) pada skala yang lebih luas, yakni 50 sampel pertanyaan terpisah. Evaluasi ini mengukur metrik *faithfulness* (kesetiaan pada dokumen sumber) dan *abstain rate*. Ke-50 pertanyaan ini mencakup dua kategori representatif: pertanyaan dalam domain Fatwa MUI dan pertanyaan di luar domain Fatwa MUI yang memerlukan rujukan dari basis data Hadis melalui mekanisme *fallback*.

3. Hasil dan Pembahasan

3.1. Hasil Evaluasi BERTScore

BERTScore digunakan untuk mengukur kesamaan semantik antara jawaban yang dihasilkan oleh sistem dengan jawaban referensi. Evaluasi dilakukan pada 10 pertanyaan yang mencakup berbagai topik hukum Islam dari domain fatwa MUI dan hadis.

Hasil pengujian menunjukkan peningkatan signifikan dalam kualitas jawaban sistem setelah penerapan metode *Hybrid Retrieval*. Pada sistem RAG standar, nilai rata-rata F1-Score mencapai 0,6664 dengan precision 0,6770 dan recall 0,6630. Sementara itu, sistem *Hybrid Retrieval* menghasilkan F1-Score sebesar 0,7546 dengan precision 0,7634 dan recall 0,7506. Metrik *recall* mengalami kenaikan sebesar 13,20% (dari 0,6630 menjadi 0,7506), menunjukkan bahwa sistem *Hybrid Retrieval* lebih tanggap dalam menangkap aspek-aspek penting dari jawaban referensi. Secara keseluruhan, F1-Score mencatatkan peningkatan sebesar 13,23%, mencerminkan keseimbangan yang lebih baik antara *precision* dan *recall*. Hasil selengkapnya disajikan pada Tabel 1.

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal:123-129 e-ISSN: 2987-7016

Tabel 1. Hasil Evaluasi BERTScore pada 10 Sampel Pertanyaan

Sistem	Precision	Recall	F1-Score
Standard RAG	0,6770	0,6630	0,6664
Hybrid RAG	0,7634	0,7506	0,7546
Peningkatan	+12,75%	+13,20%	+13,23%

3.2. Hasil Evaluasi RAGET: Faithfulness dan Abstain Rate

Untuk mengevaluasi keandalan jawaban dan potensi halusinasi, penelitian ini menggunakan *framework* Giskard dengan RAGET (*Retrieval Augmented Generation Evaluation Toolkit*). Evaluasi dilakukan pada 50 sampel pertanyaan yang telah dianotasi sebagai *gold standard*, dengan fokus pada dua metrik utama: *faithfulness* dan *abstain rate*. Selama proses evaluasi otomatis, beberapa baris dataset tidak berhasil dievaluasi akibat kendala teknis. Untuk memastikan perbandingan yang adil (*apple-to-apple comparison*), analisis akhir dilakukan pada 41 pertanyaan yang sama-sama berhasil dievaluasi oleh kedua sistem.

Faithfulness mengukur seberapa akurat jawaban sistem dalam mengikuti konteks yang diperoleh dari dokumen referensi. Hasil pengujian pada 41 pertanyaan irisan menunjukkan bahwa sistem Standard RAG memiliki nilai *faithfulness* rata-rata 0,8537 (85,37%), sedangkan sistem *Hybrid Retrieval* menghasilkan nilai *faithfulness* sebesar 0,9439 (94,39%), mencerminkan peningkatan sebesar 10,57%. *Abstain Rate* yang mengukur proporsi pertanyaan yang tidak dapat dijawab menghasilkan nilai identik sebesar 0,2683 (26,83%) pada kedua sistem. Temuan ini mengungkapkan bahwa keunggulan utama *Hybrid Retrieval* tidak terletak pada peningkatan jumlah pertanyaan yang dapat dijawab, melainkan pada peningkatan kualitas dan akurasi jawaban yang dihasilkan. Hasil selengkapnya disajikan pada Tabel 2.

Tabel 2. Hasil Evaluasi Giskard RAGET

Metrik	Standard RAG	Hybrid RAG	Perubahan
Faithfulness	0,8537 (85,37%)	0,9439 (94,39%)	+10,57%
Abstain Rate	0,2683 (26,83%)	0,2683 (26,83%)	0,00%
Evaluated Rows	46	43	41 (irisan)

3.3. Analisis Komparatif Sistem Standard vs Hybrid

Perbandingan antara kedua sistem pada 41 pertanyaan irisan menunjukkan bahwa pendekatan *Hybrid Retrieval* memberikan kontribusi yang sangat signifikan dalam meningkatkan kualitas jawaban dari chatbot fatwa, meskipun tidak meningkatkan jumlah pertanyaan yang dapat dijawab. Pada dimensi *Answer Relevance*, peningkatan F1-Score sebesar 13,23% menunjukkan bahwa jawaban sistem *Hybrid* lebih relevan dan komprehensif. Pada dimensi *Faithfulness*, peningkatan sebesar 10,57% (dari 85,37% menjadi 94,39%) membuktikan bahwa sistem mampu menghasilkan jawaban yang jauh lebih akurat dan setia terhadap konteks referensi, meminimalkan potensi halusinasi. Pada dimensi *Coverage*, abstain rate identik sebesar 26,83% menunjukkan bahwa keunggulan *Hybrid Retrieval* terletak pada kualitas jawaban, bukan sekadar kuantitas pertanyaan yang dijawab.

3.4. Pembahasan dan Perbandingan Literatur

Hasil yang diperoleh secara konsisten memvalidasi hipotesis awal bahwa penerapan metode *Hybrid Retrieval* dengan mekanisme pencarian berjenjang mampu meningkatkan kualitas jawaban sistem dibandingkan dengan pendekatan RAG standar. F1-Score meningkat 13,23% dan *faithfulness* meningkat 10,57%, membuktikan bahwa Hybrid Retrieval merupakan pendekatan yang efektif untuk meningkatkan keandalan faktual jawaban, meskipun belum mampu memperluas cakupan kuantitas pertanyaan yang dapat dijawab.

Dibandingkan dengan Haekal et al. [6] yang menerapkan RAG pada dokumen fatwa dengan pendekatan *single corpus*, penelitian ini berkontribusi dengan mengintegrasikan mekanisme *Hybrid Retrieval* yang memungkinkan sistem secara otomatis beralih ke basis data hadis ketika konteks fatwa tidak memadai. Rahayu et al. [3] mencapai rata-rata F1-Score (BERTScore) sebesar 81% menggunakan LangChain pada Fikih Empat Mazhab, namun masih mengandalkan *retrieval* linier berbasis kesamaan vektor tanpa verifikasi *semantic relevance* dan *fallback*. Penelitian ini mengusulkan arsitektur *dual-stage filtering* yang terbukti mampu menjamin tingkat *faithfulness* hingga 94,39%, sekaligus mengamankan sistem dari *out-of-domain query* melalui mekanisme *fallback* ke basis data hadis.

Peningkatan performa ini juga konsisten dengan temuan [16] yang menunjukkan bahwa penggabungan beberapa sumber pengetahuan secara konsisten menghasilkan performa lebih baik dibandingkan model tunggal. Pendekatan ini juga sejalan dengan [17] yang membuktikan bahwa kombinasi *sparse* dan *dense retrieval* yang dilengkapi *reranking* berbasis LLM secara signifikan meningkatkan akurasi sistem tanya jawab pada domain biomedis, yang memiliki karakteristik serupa dengan domain hukum Islam.

3.5. Signifikansi dan Implikasi Praktis

Secara teoretis, penelitian ini memperkenalkan arsitektur *Hybrid Retrieval* multi-korpus yang memperkaya literatur sistem tanya jawab berbasis kecerdasan buatan dalam domain hukum Islam, dengan menyediakan pendekatan berjenjang yang menggabungkan *semantic similarity search* dan

verifikasi relevansi berbasis LLM. Secara praktis, mekanisme *Hybrid Retrieval* berhasil meningkatkan tingkat keakuratan jawaban (*faithfulness*) dari 85,37% menjadi 94,39%, meminimalkan risiko penyebaran informasi hukum Islam yang keliru. Hal ini sangat penting mengingat fatwa MUI memiliki peran sebagai kompas moral bagi umat. Sistem ini pun berpotensi diadaptasi oleh institusi keislaman lainnya, lembaga pendidikan berbasis pesantren, maupun platform layanan publik berbasis AI yang memerlukan integrasi sumber pengetahuan ganda dari domain spesifik.

3.6. Keterbatasan dan Implikasi untuk Penelitian Lanjutan

Meskipun hasil penelitian ini menunjukkan peningkatan yang konsisten dan terukur, terdapat beberapa keterbatasan yang perlu diakui. Pertama, evaluasi BERTScore dilakukan hanya pada 10 sampel pertanyaan dan RAGET pada 50 sampel, yang relatif terbatas untuk merepresentasikan seluruh spektrum pertanyaan hukum Islam. Kedua, sistem saat ini belum melibatkan evaluasi dari pengguna nyata (*user testing*) maupun validasi dari pakar hukum Islam.

Selain itu, tingkat penolakan (*abstain rate*) yang masih tertahan di angka 26,83% mengindikasikan bahwa mekanisme *fallback* saat ini belum sepenuhnya mencakup seluruh variasi pertanyaan hukum Islam. Oleh karena itu, untuk penelitian lanjutan disarankan adanya integrasi sumber pengetahuan tersier seperti kitab fikih klasik atau ensiklopedia hukum Islam guna memperkecil angka *abstain rate* tersebut. Disarankan pula penggunaan model *embedding* berbasis bahasa Arab-Indonesia yang lebih spesifik, serta penerapan teknik *fine-tuning* pada model bahasa dengan dataset fatwa dan hadis.

4. Kesimpulan

Penelitian ini berhasil mengembangkan Chatbot Hybrid Fatwa MUI dengan menerapkan arsitektur *Hybrid Retrieval* yang mengintegrasikan dokumen Fatwa MUI sebagai korpus utama dan basis data hadis sebagai mekanisme *fallback* otomatis. Hasil evaluasi secara konsisten memvalidasi hipotesis penelitian bahwa pendekatan *Hybrid Retrieval* mampu meningkatkan kualitas jawaban sistem secara signifikan dibandingkan RAG standar berbasis *single corpus.* Hal ini dibuktikan melalui tiga dimensi: peningkatan F1-Score BERTScore sebesar 13,23% yang mencerminkan relevansi semantik jawaban yang lebih tinggi; peningkatan *faithfulness* sebesar 10,57% (dari 85,37% menjadi 94,39%) yang mengindikasikan jawaban yang lebih setia terhadap konteks referensi; serta *abstain rate* identik sebesar 26,83% pada kedua sistem, menunjukkan bahwa keunggulan utama *Hybrid Retrieval* terletak pada peningkatan kualitas jawaban, bukan pada peningkatan jumlah pertanyaan yang dapat dijawab.

Meskipun demikian, penelitian ini masih memiliki keterbatasan, terutama pada tingkat *abstain rate* yang belum mengalami penurunan (tetap 26,83%), serta jumlah sampel evaluasi yang relatif terbatas tanpa adanya validasi dari pakar hukum Islam. Untuk penelitian selanjutnya, guna mengatasi tingginya angka *abstain rate* tersebut, disarankan untuk mengintegrasikan sumber pengetahuan tersier (seperti kitab fikih klasik), di samping penggunaan model *embedding* yang lebih spesifik terhadap bahasa Arab-Indonesia, penerapan *fine-tuning* pada dataset keislaman, serta pelibatan evaluasi pakar ulama dan mekanisme *feedback loop* guna meningkatkan kualitas sistem secara berkelanjutan.

Daftar Rujukan

[1] W. X. Zhao *et al.*, “A Survey of Large Language Models,” *arXiv Prepr. arXiv2303.18223*, pp. 1–144, 2026, doi: https://doi.org/10.48550/arXiv.2303.18223.

[2] L. Xu, L. Sanders, K. Li, and J. C. L. Chow, “Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review,” *JMIR Cancer*, vol. 7, no. 4, 2021, https://doi.org/10.2196/27850

[3] M. I. Ardimansyah and M. H. Widianto, “Development of online learning media based on Telegram Chatbot (Case studies: Programming courses),” *J. Phys. Conf. Ser.*, vol. 1987, no. 1, 2021, https://doi.org/10.1088/1742-6596/1987/1/012006

[4] M. Hasanuddin, M. S. Bin Sharuddin, and J. W. Mahri, “Pemodelan Fatwa Ekonomi Syariah dan Karakteristiknya di Indonesia,” *Asy-Syari’ah, Vol*, vol. 25, no. 1, pp. 1–16, 2023, https://doi.org/10.15575/as.v25i1.25373

[5] M. Shuhufi and F. Muhammad, “I Fatwas and Sharia-Based Policy Governance in Islamic Education Management in Indonesia,” vol. 10, no. 1, pp. 184–200, 2026, doi: 10.22373/sjhk.v10.i1.28762.

[6] R. R. Dewi, “Problematika Artificial Intelligence Sebagai Pemberi Fatwa Dalam Perspektif Hukum Islam,” *J. Anal. Huk.*, vol. 7, no. 2, pp. 209–223, 2024, https://doi.org/10.38043/jah.v7i2.5137

[7] S. A. Putra, “AD-DUSTUR Jurnal Hukum dan Konstitusi AD-DUSTUR Jurnal Hukum dan Konstitusi,” *AD-Dustur J. Huk. dan Konstitusi*, vol. 1, no. 1, pp. 1–17, 2024. https://doi.org/10.58326/jad.v1i1.195

[8] S. M. T. I. Tonmoy, S. M. M. Zaman, A. Rani, V. Rawte, A. Chadha, and A. Das, “A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models,” *arXiv Prepr. arXiv2401.01313*, 2023, doi: https://doi.org/10.48550/arXiv.2401.01313.

[9] Y. Gao *et al.*, “Retrieval-Augmented Generation for Large Language Models : A Survey,” pp. 1–21, doi: https://doi.org/10.48550/arXiv.2312.10997.

[10] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, “SELF-RAG: LEARNING TO RETRIEVE, GENERATE, AND CRITIQUE THROUGH SELF-REFLECTION,” *arXiv Prepr. arXiv2509.01081*, no. Step 1, 2023, doi: https://doi.org/10.48550/arXiv.2310.11511.

[11] M. Erlangga Fauzi and Tata Sutabri, “PublicTalk: Sistem Chatbot Pintar Berbasis Natural Language Processing untuk Layanan Pemerintahan Digital,” *J. Sains Student Res.*, vol. 3, no. 2, pp. 426–433, 2025, https://doi.org/10.61722/jssr.v3i2.4325

[12] F. L. D. Cahyanti and R. D. A. Raya, “Perancangan Sistem Informasi Chatbot Retrieval Augmented Generation Berbasis Website Pada PT. Revolusi Cita Edukasi,” *Indones. J. Comput. Sci.*, vol. 4, no. 1, pp. 15–21, 2025,

https://doi.org/10.31294/m75d4782

[13] S. Rahayu, N. S. Harahap, S. Agustian, and P. Pizaini, "Penerapan Teknologi LangChain pada Question Answering System Fikih Empat Madzhab," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, pp. 974–983, 2024, https://doi.org/10.57152/malcom.v4i3.1397

[14] T. Helviansyah, N. S. Harahap, M. Irsyad, and B. S. Negara, "Sistem Tanya Jawab Berbasis Chatbot Website Menggunakan Gemini Ai Pada Data Fiqih Kontemporer," *J. Inf. Syst. Manag.*, vol. 7, no. 1, pp. 38–47, 2025, https://doi.org/10.24076/joism.2025v7i1.2082

[15] N. A. M. Herwanza, N. S. Harahap, F. Yanto, and F. Insani, "Penerapan Langchain Retriever dengan Model Chat Openai dalam Pengembangan Sistem Chatbot Hadis Berbasis Telegram," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 6, no. 1, pp. 70–83, 2024, https://doi.org/10.35746/jtim.v6i1.514

[16] F. I. Haekal, R. Setya Perdana, and P. P. Adikara, "Sistem Tanya Jawab Closed-Domain terhadap Dokumen Fatwa menggunakan Retrieval Augmented Generation dan Large Language Model," vol. 9, no. 5, pp. 2548–964, 2025, [Online]. Available: http://j-ptiik.ub.ac.id

[17] A. Bouchekif, S. Rashwani, H. Sbahi, S. Gaben, M. Al-khatib, and M. Ghaly, "Assessing Large Language Models on Islamic Legal Reasoning: Evidence from Inheritance Law Evaluation," 2025, doi: https://doi.org/10.48550/arXiv.2509.01081.

[18] L. Machado, R. Miyaji, R. Moulin, and S. Monc, "Evaluating RAG-based QA Systems : A Comparative Analysis of LLM as a Judge , Traditional Metrics , and Human Alignment," 2024.

[19] F. Dobslaw, R. Feldt, J. Yoon, and S. Yoo, "Challenges in Testing Large Language Model Based Software : A Faceted Taxonomy," *rXiv Prepr. arXiv2503.00481*, vol. 1, no. 1, pp. 1–20, 2025, doi: https://doi.org/10.48550/arXiv.2503.00481.

[20] L. Team and A. I. Meta, "The Llama 3 Herd of Models," *arXiv Prepr. arXiv2407.21783*, pp. 1–92, 2024, doi: https://doi.org/10.48550/arXiv.2407.21783.