Contents lists available at JESA

Journal of Engineering and Science Application

journal homepage: https://jesa.aks.or.id/index.php/jesa/index

Analisis Perbandingan *Large Language Model Meta AI* (LLaMA) dan Mistral pada Pendekatan *Retrieval-Augmented Generation* (RAG) untuk *Chatbot* Layanan Masyarakat Desa (Studi Kasus: Desa Bou)

Putri Angraini Aziz[a], Bambang Pramono[b], Adha Mashur Sajiah[c]
[a,b,c]*Universitas Halu Oleo, Jln. H.E.A. Mokodompit, Kota Kendari, Sulawesi Tenggara 93231*

Article Info

Keywords:

Chatbot
Retrieval-Augmented Generation (RAG)
LLaMA 3.1
Mistral 7B
Layanan Masyarakat Desa

ABSTRACT

Keterbatasan akses informasi layanan administrasi dan pemerintahan desa mendorong perlunya sistem informasi yang mampu memberikan jawaban berdasarkan dokumen resmi. Penelitian ini bertujuan mengembangkan TabeBot, *chatbot* layanan masyarakat Desa Bou berbasis *Retrieval-Augmented Generation* (RAG), serta membandingkan kinerja model LLaMA 3.1 8B dan Mistral 7B. Sistem dikembangkan menggunakan metode *Rational Unified Process* (RUP) dengan basis pengetahuan sebanyak 34 dokumen PDF yang menghasilkan 6.611 *chunk*. Arsitektur RAG menerapkan *chunking* dan *embedding* menggunakan *multilingual-e5-base*, penyimpanan vektor pada ChromaDB, serta *hybrid retrieval* yang menggabungkan BM25 dan *dense vector retrieval*. Kandidat hasil pencarian diurutkan menggunakan *Cross-Encoder Reranker* dan diproses dengan aturan khusus dokumen hukum berdasarkan metadata Registry Peraturan sebelum diberikan kepada LLM. Pengujian fungsional menggunakan *Black Box Testing*, sedangkan evaluasi kualitas jawaban dilakukan menggunakan RAGAS terhadap 125 pertanyaan. Hasil pengujian menunjukkan seluruh 19 skenario fungsional berhasil. LLaMA 3.1 memperoleh nilai *faithfulness* 0,84 dan *answer relevancy* 0,85, sedangkan Mistral 7B memperoleh masing-masing 0,82. Kedua model memperoleh nilai *context precision* 0,81 dan *context recall* 0,76. Berdasarkan hasil tersebut, LLaMA 3.1 dipilih untuk implementasi akhir TabeBot.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Putri Angraini Aziz
Jurusan Teknik Informatika, Fakultas Teknik
Universitas Halu Oleo
Jln. H.E.A. Mokodompit, Kota Kendari, Sulawesi Tenggara 93231
Email: putriangrainiaziz27@gmail.com

1. PENDAHULUAN

Transformasi digital menjadi salah satu upaya pemerintah dalam meningkatkan kualitas pelayanan publik melalui penerapan Sistem Pemerintahan Berbasis Elektronik (SPBE) yang diharapkan mewujudkan pelayanan yang responsif, efisien, dan transparan [1]. Namun, penerapan digitalisasi pada tingkat desa masih menghadapi kendala berupa keterlambatan penyampaian informasi, penggunaan mekanisme manual dalam administrasi, dan keterbatasan sumber daya manusia. Kondisi tersebut juga ditemukan di Desa Bou. Hasil

observasi awal menunjukkan bahwa masyarakat memperoleh informasi layanan dengan mendatangi kantor desa secara langsung, menghubungi perangkat desa, atau melalui komunikasi informal antarwarga. Mekanisme ini menyebabkan pertanyaan yang sama diajukan berulang kali kepada perangkat desa serta berpotensi menimbulkan informasi yang tidak seragam.

Perkembangan *Large Language Model* (LLM) membuka peluang pengembangan *chatbot* yang mampu memahami pertanyaan dalam bahasa alami. Namun, LLM menghasilkan jawaban berdasarkan data pelatihannya sehingga sulit diperbarui dan berpotensi menghasilkan halusinasi [2], [3]. Model umum juga tidak memiliki pengetahuan khusus mengenai prosedur dan kebijakan pada suatu desa. *Retrieval-Augmented Generation* (RAG) mengatasi keterbatasan tersebut dengan menggabungkan model bahasa dan basis dokumen eksternal yang diakses melalui *retriever*, sehingga jawaban lebih faktual dan pengetahuan dapat diperbarui dengan mengganti dokumen [3]. Pendekatan ini juga dapat diterapkan pada model dasar tanpa *fine-tuning*, meskipun halusinasi masih mungkin terjadi apabila dokumen yang diambil kurang relevan [2].

Penerapan RAG pada dokumen peraturan menghadapi tantangan tambahan, karena ketentuan hukum bersifat dinamis dan dapat mengalami perubahan maupun pencabutan dari waktu ke waktu [4]. Sistem RAG yang hanya mengambil dokumen berdasarkan kemiripan semantik tanpa mempertimbangkan masa berlaku suatu ketentuan berisiko menyajikan versi aturan yang tidak sesuai dengan konteks waktu pertanyaan [5], sehingga perubahan norma perlu dimodelkan secara eksplisit dalam proses *retrieval* [4]. Oleh karena itu, sistem yang dikembangkan dilengkapi Registry Peraturan untuk mengelola status keberlakuan, verifikasi, dan hubungan perubahan antarperaturan. Informasi tersebut diintegrasikan ke dalam metadata dokumen untuk penyaringan dan pemeringkatan hasil *retrieval* serta penyusunan konteks bagi LLM. Selain itu, sistem menerapkan *hybrid retrieval* yang menggabungkan BM25 sebagai metode pencarian berbasis kata kunci (*sparse retrieval*) dan *dense vector retrieval* sebagai metode pencarian berbasis representasi vektor. Metode berbasis kata kunci peka terhadap frasa yang spesifik, sedangkan *dense retrieval* lebih baik menangkap variasi leksikal dan hubungan semantik, sehingga keduanya bersifat saling melengkapi [6]. Penggabungan kedua metode tersebut bertujuan memanfaatkan kecocokan kata kunci dan kemiripan semantik dalam proses pencarian dokumen yang relevan dengan pertanyaan pengguna [7], [8]. Kandidat dokumen hasil *retrieval* selanjutnya diurutkan ulang menggunakan *Cross-Encoder Reranker*. Pada tahap ini, model memberikan penilaian terhadap setiap pasangan pertanyaan dan dokumen untuk menentukan urutan kandidat berdasarkan relevansinya [8].

Pada sisi pembangkitan jawaban, penelitian ini menggunakan LLaMA 3.1 8B [9] dan Mistral 7B [10], dua LLM *open weight* yang bobot modelnya dapat diunduh dan dijalankan pada infrastruktur sendiri. Pada RAG layanan informasi sekolah dengan 30 pertanyaan uji, LLaMA-3-8B-Instruct dan Mistral-7B-Instruct-v0.3 sama-sama mencapai akurasi 100% [11], sehingga kinerja keduanya belum dapat dibedakan. Pada RAG domain tanaman obat dengan 6 pertanyaan uji, Mistral 7B memperoleh skor METEOR lebih tinggi daripada LLaMA 2 7B, tetapi skor ROUGE *precision*-nya lebih rendah [12]. Kedua penelitian tersebut tidak menguji LLaMA 3.1, tidak berada pada domain pemerintahan, menggunakan sampel pertanyaan yang kecil, dan tidak memakai metrik evaluasi khusus RAG. Dengan demikian, perbandingan LLaMA 3.1 8B dan Mistral 7B pada layanan informasi pemerintahan desa masih terbatas.

Penelitian ini bertujuan mengembangkan *chatbot* layanan masyarakat Desa Bou berbasis RAG yang diberi nama TabeBot. Sistem ini dilengkapi dengan registry peraturan, *hybrid retrieval*, dan *Cross-Encoder Reranker*, serta membandingkan kinerja LLaMA 3.1 8B dan Mistral 7B dengan pertanyaan, hasil *retrieval*, dan konteks yang sama. Evaluasi sistem dilakukan menggunakan kerangka kerja RAGAS [13] dengan menerapkan empat metrik utama, yaitu *Faithfulness, Answer Relevancy*, *Context Precision*, dan *Context Recall* [14] . Kebaruan penelitian ini terletak pada integrasi registry peraturan ke dalam proses *retrieval* dan perbandingan kedua model dengan masukan yang setara pada domain pemerintahan desa. *Chatbot* ini berfungsi sebagai alat bantu informasi dan tidak menggantikan keputusan resmi pemerintah desa.

2. METODE PENELITIAN

2.1 Pengumpulan dan Persiapan Data

Pengumpulan data dilakukan melalui observasi dan wawancara dengan perangkat Desa Bou untuk mengidentifikasi kebutuhan informasi masyarakat serta memperoleh gambaran mengenai layanan dan informasi yang tersedia. Hasilnya digunakan untuk menyusun pertanyaan uji sebanyak 125. Sebagian pertanyaan bersumber dari pertanyaan yang disampaikan masyarakat, sedangkan sisanya disusun oleh peneliti berdasarkan isi dokumen basis pengetahuan dengan mengacu pada topik layanan yang diidentifikasi pada tahap observasi. Setiap pertanyaan dilengkapi jawaban referensi yang disusun langsung dari dokumen sumber dan digunakan sebagai acuan penilaian pada tahap evaluasi.

Basis pengetahuan sistem dibangun dari 34 dokumen PDF yang berasal dari dua sumber, yaitu dokumen yang diperoleh langsung dari kantor Desa Bou dan dokumen peraturan yang diunduh dari situs resmi pemerintah. Dari sisi jenisnya, dokumen tersebut meliputi profil desa, data perangkat desa, peraturan

dan keputusan kepala desa, serta peraturan perundang-undangan, mulai dari undang-undang, peraturan pemerintah, peraturan presiden, dan peraturan menteri hingga peraturan daerah dan peraturan bupati Kolaka Timur. Isi dokumen tersebut mencakup berbagai informasi layanan desa, antara lain administrasi umum dan informasi pemerintahan desa, bantuan sosial dan penanggulangan kemiskinan, kependudukan dan pencatatan sipil, pengelolaan keuangan dan pembangunan desa, serta administrasi perkawinan dan kesejahteraan keluarga. Dokumen berupa PDF berbasis teks (*native PDF*) dan PDF hasil pemindaian (*scanned PDF*), dan sebagian memuat tabel, gambar, bagan, maupun informasi visual lainnya.

Karena karakteristik tersebut, persiapan dokumen dilakukan melalui tahap *parsing* dan *preprocessing* dengan metode ekstraksi yang disesuaikan untuk setiap halaman. Pada halaman berbasis teks, PyMuPDF digunakan untuk mengekstraksi teks dan pdfplumber untuk mengekstraksi tabel. Halaman hasil pemindaian diproses menggunakan Tesseract OCR setelah melalui prapemrosesan citra, sedangkan OpenCV digunakan untuk mendukung pengolahan citra dan deteksi wilayah tabel. Informasi visual yang relevan, seperti bagan, grafik, peta, dan tabel berbentuk gambar, diproses dengan model *vision* Qwen2.5-VL-7B melalui Ollama. Hasil ekstraksi kemudian dibersihkan dan disusun menjadi data teks yang dilengkapi metadata berupa nama dokumen dan nomor halaman, lalu dipersiapkan untuk proses *chunking* dan pembentukan basis pengetahuan RAG.

2.2 Tahapan Pengembangan Sistem (RUP)

Pengembangan sistem menggunakan metode *Rational Unified Process* (RUP), yaitu kerangka kerja iteratif berbasis paradigma berorientasi objek yang menyediakan tahapan pengembangan terstruktur dan terdokumentasi agar perangkat lunak sesuai dengan kebutuhan pengguna [15]. RUP terdiri atas empat tahapan berikut:

a. *Inception*

Tahap ini dilakukan untuk mengidentifikasi permasalahan pelayanan informasi di Desa Bou, menentukan tujuan dan ruang lingkup pengembangan, serta mengidentifikasi kebutuhan sistem berdasarkan kondisi dan kebutuhan informasi masyarakat.

b. *Elaboration*

Tahap ini berfokus pada analisis kebutuhan dan perancangan sistem. Perancangan dilakukan menggunakan *flowchart* dan diagram UML yang meliputi *use case diagram*, *activity diagram*, dan *class diagram*. Selain itu, dilakukan perancangan antarmuka berbasis web untuk masyarakat dan halaman pengelolaan bagi administrator, termasuk pengelolaan basis pengetahuan dan Registry Peraturan.

c. *Construction*

Tahap ini merupakan implementasi hasil perancangan ke dalam aplikasi. Kegiatan meliputi pengolahan dokumen sebagai basis pengetahuan, implementasi alur RAG termasuk Registry Peraturan, integrasi model LLaMA 3.1 dan Mistral 7B melalui Ollama yang dijalankan pada lingkungan Kaggle dan dihubungkan ke aplikasi menggunakan ngrok, serta pengembangan aplikasi *chatbot* berbasis aplikasi web menggunakan Flask.

d. *Transition*

Tahap ini dilakukan untuk menguji dan mengevaluasi sistem. Pengujian fungsional menggunakan Black Box Testing untuk memastikan fitur berjalan sesuai kebutuhan, sedangkan kualitas jawaban dievaluasi menggunakan RAGAS untuk membandingkan LLaMA 3.1 dan Mistral 7B berdasarkan empat metrik, yaitu *faithfulness*, *answer relevancy*, *context precision*, dan *context recall*.

2.3 Arsitektur dan Alur Sistem *Retrieval-Augmented Generation* (RAG)

Implementasi RAG pada TabeBot terdiri atas dua proses utama, yaitu *indexing* dan *querying*. Pada proses *indexing*, dokumen hasil *preprocessing* dibagi menjadi *chunk* berukuran 800 karakter dengan *overlap* 100 karakter. Dokumen peraturan dibagi berdasarkan Pasal, ayat, dan butir agar ketentuan tidak terpotong, sedangkan dokumen umum dibagi berdasarkan judul, tabel, dan daftar. Setiap *chunk* dilengkapi metadata berupa nama dokumen, jenis dokumen, nomor halaman, dan konteks judul bagian. *Chunk* dikonversi menjadi vektor menggunakan intfloat/multilingual-e5-base dengan prefix *passage:* dan normalisasi vektor, lalu disimpan pada ChromaDB dengan jarak *cosine*.

Karena ketentuan hukum dapat berubah atau dicabut, pencarian yang hanya mengandalkan kemiripan semantik berisiko menyajikan pasal yang sudah tidak berlaku. Untuk itu, sistem dilengkapi Registry Peraturan yang dikelola administrator. Saat dokumen baru diindeks, sistem membuat entri draf secara otomatis dengan aturan berbasis ekspresi reguler, yaitu jenis peraturan dari judul atau nama berkas bila OCR gagal, nomor dan tahun dari pola "Nomor … Tahun …", serta hierarki dari jenis peraturan. Apabila judul menunjukkan peraturan perubahan, sistem mencocokkannya dengan peraturan terdaftar dan mengumpulkan nomor pasal dari kalimat "Ketentuan Pasal … diubah" dan "Pasal … dihapus" sebagai usulan pasal terdampak. Entri berstatus *pending* dan belum mengubah peraturan lama sampai administrator meninjau, mengoreksi, dan mengonfirmasinya. Setelah dikonfirmasi, peraturan lama berstatus diubah atau

dicabut dan metadata *chunk* di ChromaDB diperbarui tanpa *embedding* ulang. *Chunk* diberi penanda pasal usang apabila peraturannya dicabut atau pasalnya termasuk pasal terdampak, dan penanda konflik apabila pasalnya ditetapkan bertentangan dengan peraturan lain.

Pada proses *querying*, sistem menerima pertanyaan atau pesan dari pengguna sebagai masukan. Sistem terlebih dahulu memeriksa apakah pesan termasuk percakapan sederhana (*small talk*), seperti sapaan. Jika termasuk *small talk*, sistem memberikan respons tanpa menjalankan proses *retrieval*. Untuk pertanyaan umum yang dapat dipahami secara langsung, sistem melanjutkan ke proses *retrieval* untuk mencari informasi yang relevan pada basis pengetahuan. Sementara itu, apabila pertanyaan merupakan lanjutan dari percakapan sebelumnya dan masih bergantung pada konteks terdahulu, sistem menerapkan *query rewriting* dengan memanfaatkan riwayat percakapan agar pertanyaan menjadi lebih jelas dan dapat diproses secara mandiri. Sistem juga dapat menambahkan konteks Desa Bou pada pertanyaan tertentu yang memerlukan penyesuaian konteks lokasi. Pertanyaan yang akan dicari kemudian dikonversi menjadi *embedding* menggunakan model intfloat/multilingual-e5-base dengan prefix *query:*. Pada tahap *retrieval*, sistem menerapkan metode *hybrid retrieval* yang menggabungkan *sparse retrieval* menggunakan BM25 dan *dense retrieval* menggunakan *vector similarity*. Kedua metode tersebut digabungkan melalui *EnsembleRetriever* dengan bobot masing-masing 0,4 untuk BM25 dan 0,6 untuk pencarian vektor. Sistem mengambil 15 kandidat awal untuk diproses pada tahap pemeringkatan ulang (*reranking*). Khusus pertanyaan yang membutuhkan informasi menyeluruh, sistem melakukan pencarian tambahan pada setiap sumber dokumen untuk memperluas cakupan informasi yang diperoleh.

Kandidat hasil *retrieval* selanjutnya melalui proses diversifikasi sumber dokumen sebelum diproses menggunakan *Cross-Encoder Reranker* dengan model cross-encoder/mmarco-mMiniLMv2-L12-H384-v1 untuk mengurutkan potongan dokumen berdasarkan tingkat relevansinya terhadap pertanyaan. Proses *reranking* menghasilkan 10 kandidat teratas. Setelah itu, sistem menerapkan pemrosesan lanjutan berupa aturan khusus untuk dokumen hukum berdasarkan metadata, perluasan potongan dokumen pada bagian tabel melalui pengambilan *chunk* yang berdekatan, serta penggabungan *chunk* yang saling tumpang tindih. Selanjutnya, sistem memilih 4 *chunk* sebagai hasil akhir *retrieval* untuk pertanyaan biasa. Sementara itu, pertanyaan yang membutuhkan informasi menyeluruh menggunakan hingga 8 *chunk* serta pencarian tambahan pada setiap sumber dokumen. Hasil *retrieval* tersebut kemudian disusun menjadi konteks dengan menyertakan informasi sumber dokumen, nomor halaman, dan status ketentuan yang relevan.

Pada tahap pembangkitan jawaban (*generation*), konteks hasil *retrieval* digabungkan dengan pertanyaan pengguna, riwayat percakapan, dan instruksi pada *prompt template*. Selanjutnya, masukan tersebut diberikan kepada model LLaMA 3.1 atau Mistral 7B yang dijalankan melalui Ollama untuk menghasilkan jawaban dalam bahasa Indonesia. Kedua model menggunakan basis pengetahuan dan konfigurasi *retrieval* yang sama. Dalam pengujian komparatif, proses *retrieval* dilakukan satu kali untuk setiap pertanyaan, kemudian hasil konteks yang sama diberikan kepada kedua model. Dengan demikian, perbandingan dapat difokuskan pada kemampuan masing-masing model dalam menghasilkan jawaban berdasarkan konteks yang diperoleh.

2.4 Pengujian

Pengujian fungsional menggunakan *Black Box Testing* untuk memastikan fitur halaman pengguna dan admin, yaitu layanan *chatbot*, *login*, pengelolaan pengetahuan, riwayat percakapan, dan Registry Peraturan, berjalan sesuai kebutuhan. Kualitas jawaban dievaluasi menggunakan RAGAS pada 125 pertanyaan uji beserta jawaban referensinya, dengan LLaMA 3.1 dan Mistral 7B menerima pertanyaan dan konteks yang sama. Penilaian menggunakan Qwen2.5:14B sebagai *LLM judge* dan intfloat/multilingual-e5-base sebagai model *embedding*. Metrik yang digunakan adalah *faithfulness*, *answer relevancy*, *context precision*, dan *context recall*. Setiap metrik bernilai 0 sampai 1 dan nilai akhir model adalah rata-rata seluruh pertanyaan. Karena konteksnya sama, *context precision* dan *context recall* mencerminkan kualitas *retrieval*, sedangkan perbedaan antarmodel pada metrik *faithfulness* dan *answer relevancy*.

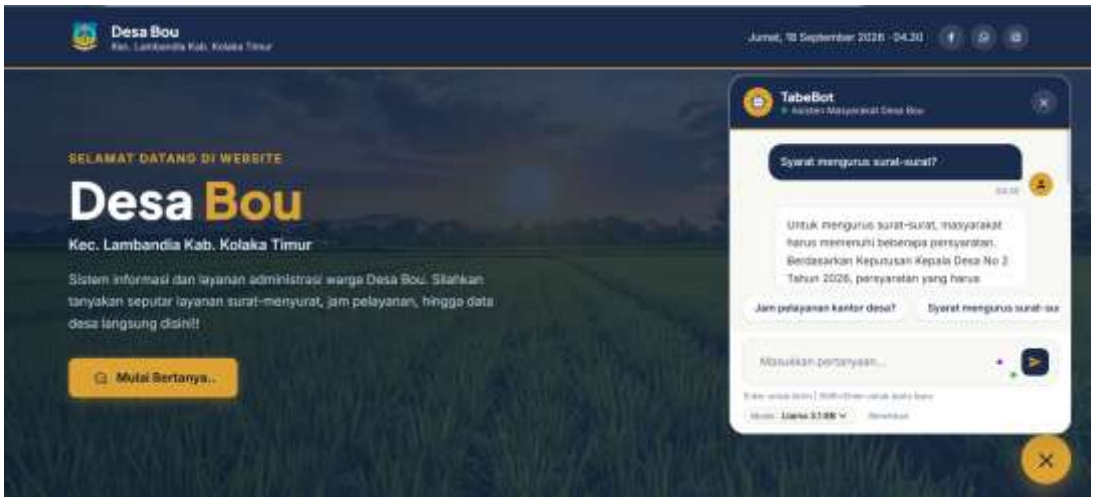
3. HASIL DAN PEMBAHASAN

3.1 Implementasi Sistem TabeBot

Sistem diimplementasikan sebagai aplikasi web TabeBot dengan dua kelompok pengguna, yaitu masyarakat dan operator desa (administrator). Masyarakat dapat mengakses layanan tanpa autentikasi, sedangkan fitur pengelolaan hanya dapat diakses administrator setelah *login*.

Antarmuka masyarakat berupa jendela *chatbot* yang terintegrasi pada halaman beranda website Desa Bou seperti yang ditunjukkan pada Gambar 1. Halaman beranda memuat penjelasan singkat mengenai layanan informasi dan administrasi desa serta tombol untuk memulai percakapan. Jendela *chatbot* dibuka melalui tombol tersebut atau ikon obrolan di sudut kanan bawah layar. Pada tampilan awal, jendela menampilkan pesan sambutan, contoh pertanyaan yang dapat dipilih langsung, kolom masukan pertanyaan,

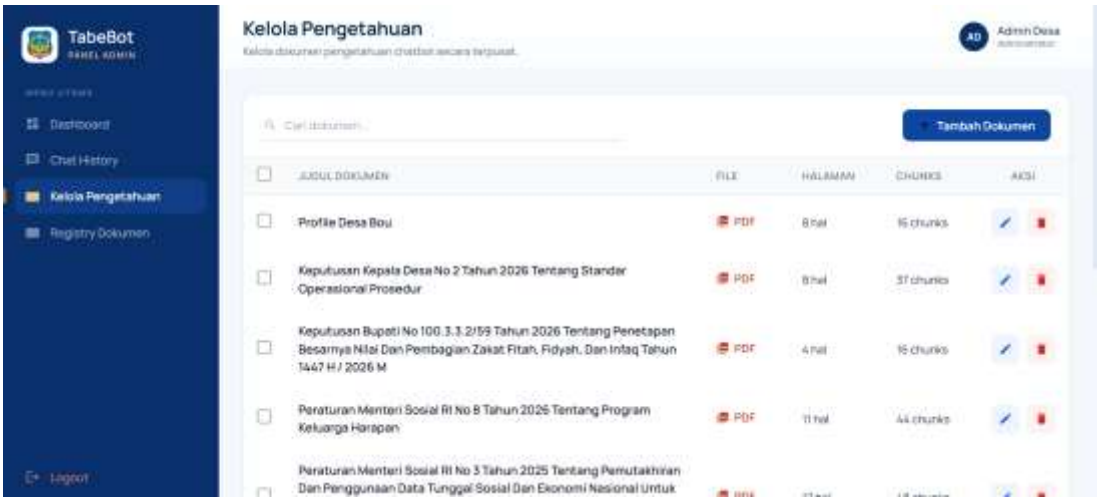
keterangan model bahasa yang digunakan, dan tombol untuk menghapus riwayat percakapan. Jawaban sistem ditampilkan dalam bentuk percakapan, seperti contoh pada Gambar 1 pertanyaan mengenai pengurusan surat-surat.



Gambar 1. Halaman Beranda Antarmuka Masyarakat

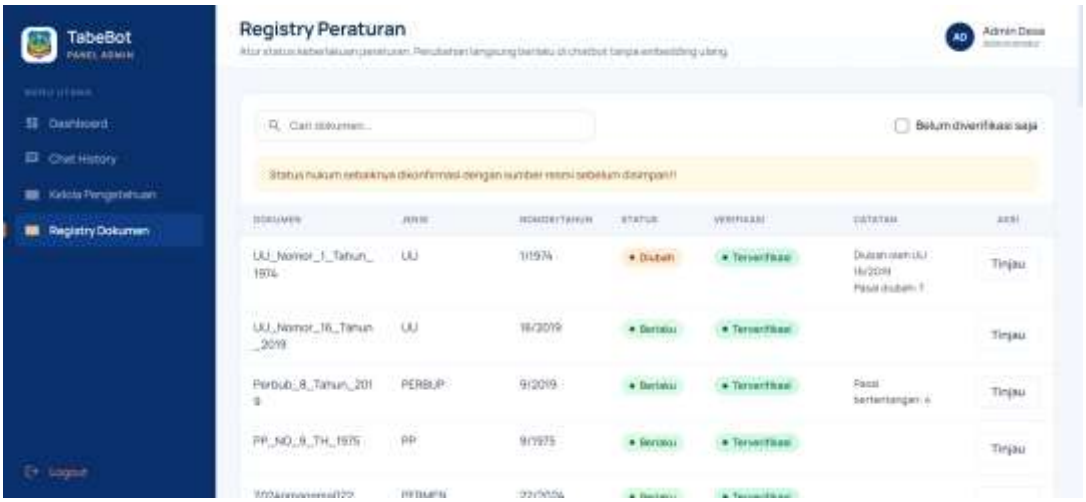
Antarmuka administrator diawali halaman *login* yang memuat kolom email dan kata sandi. Setelah autentikasi berhasil, administrator diarahkan ke *dashboard*. Antarmuka administrator terdiri atas *dashboard*, riwayat percakapan, kelola pengetahuan, dan registry peraturan. *Dashboard* menampilkan jumlah sesi percakapan, jumlah pesan, jumlah pengetahuan yang tersimpan, percakapan terbaru, dan pertanyaan yang paling sering diajukan. Halaman riwayat percakapan memuat identitas sesi, cuplikan pesan terakhir, waktu, dan jumlah pesan, serta dilengkapi fitur pencarian dan halaman detail sesi.

Halaman kelola pengetahuan menampilkan daftar dokumen beserta jumlah halaman dan jumlah *chunk* hasil pemrosesan. Pada halaman ini administrator dapat menambah, mengubah, dan menghapus dokumen PDF. Tampilan halaman kelola pengetahuan ditunjukkan pada Gambar 2.



Gambar 2. Halaman Kelola Pengetahuan

Halaman registry peraturan memuat daftar peraturan beserta jenis, nomor dan tahun, status keberlakuan (Berlaku, Diubah, Dicabut), status verifikasi, dan catatan perubahan yang mencakup peraturan pengubah dan jumlah pasal terdampak seperti yang ditunjukkan pada Gambar 3. Status yang diusulkan sistem secara otomatis ditinjau dan dikonfirmasi oleh administrator sebelum digunakan.

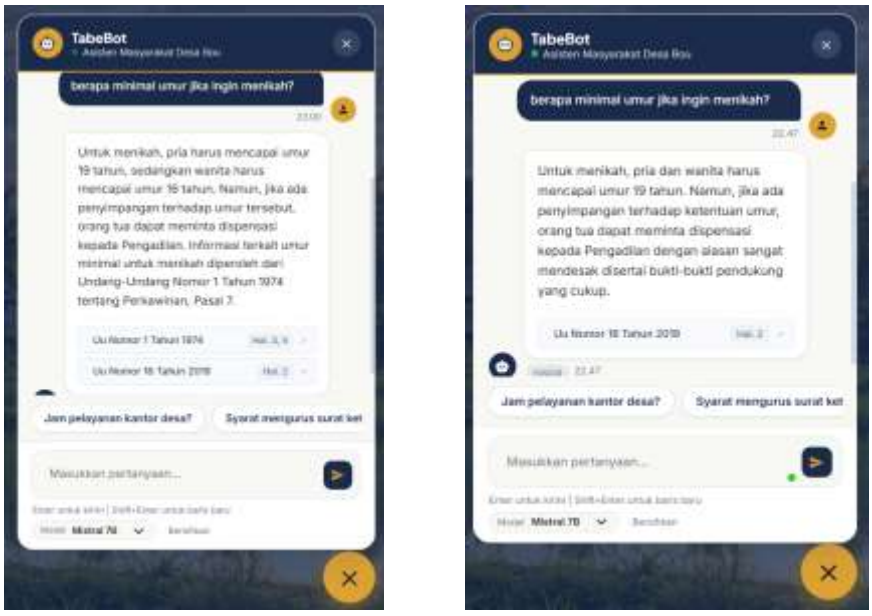


Gambar 3. Halaman Registry Peraturan

3.2 Hasil Registry Peraturan dan *Retrieval*

Pengujian registry peraturan dilakukan untuk mengamati pengaruh informasi status keberlakuan peraturan terhadap proses *retrieval* dan jawaban yang dihasilkan *chatbot*. Pengujian dilakukan dengan membandingkan hasil sistem sebelum dan sesudah penerapan aturan dokumen hukum berbasis metadata menggunakan pertanyaan yang sama. Perbandingan difokuskan pada keberadaan potongan ketentuan lama dalam konteks akhir serta kesesuaian jawaban yang dihasilkan.

Salah satu pengujian dilakukan menggunakan pertanyaan "Berapa minimal umur jika ingin menikah?". Pertanyaan ini dipilih karena basis pengetahuan memuat ketentuan batas usia perkawinan yang mengalami perubahan. Sebelum penerapan Registry Peraturan, sistem menghasilkan jawaban bahwa batas usia minimum perkawinan adalah 19 tahun bagi laki-laki dan 16 tahun bagi perempuan berdasarkan Undang-Undang Nomor 1 Tahun 1974. Jawaban tersebut tidak lagi sesuai dengan ketentuan setelah perubahan melalui Undang-Undang Nomor 16 Tahun 2019, yang menetapkan batas usia minimum 19 tahun bagi laki-laki dan perempuan. Setelah penerapan Registry Peraturan, sistem menghasilkan jawaban yang sesuai dengan ketentuan terbaru, termasuk informasi mengenai dispensasi perkawinan melalui pengadilan. Perbandingan jawaban tersebut ditunjukkan pada Gambar 4.



(a) (b)

Gambar 4. Perbandingan jawaban TabeBot (a)sebelum dan (b)sesudah penerapan Registry Peraturan

Perubahan jawaban tersebut didukung oleh perubahan hasil pemeringkatan dan seleksi konteks. Pada proses *reranking* sebelum penerapan aturan dokumen hukum berbasis metadata, Penjelasan Umum

Undang-Undang Nomor 1 Tahun 1974 menempati peringkat pertama dengan skor 1,71 karena memuat informasi batas usia 16 tahun bagi perempuan. Sementara itu, Pasal 7 Undang-Undang Nomor 16 Tahun 2019 yang memuat ketentuan terbaru berada pada peringkat keenam dengan skor −0,10. Kondisi ini menunjukkan bahwa kemiripan isi dokumen dengan pertanyaan belum cukup untuk membedakan ketentuan yang masih berlaku dan ketentuan yang telah mengalami perubahan.

Setelah informasi status keberlakuan peraturan diterapkan, satu kandidat yang ditandai usang disingkirkan dari proses seleksi. Dari sembilan kandidat yang tersisa, sistem memilih empat potongan dokumen untuk membentuk konteks akhir. Pasal 7 Undang-Undang Nomor 16 Tahun 2019 kemudian menempati peringkat pertama dengan skor akhir 3,25. Dengan demikian, ketentuan terbaru memperoleh prioritas dalam penyusunan konteks yang diberikan kepada model bahasa untuk menghasilkan jawaban.

Pengujian juga dilakukan pada tiga pertanyaan lain yang berkaitan dengan informasi pemerintahan desa. Hasil perbandingan menunjukkan bahwa sebelum penerapan Registry Peraturan, potongan ketentuan lama masih ditemukan dalam konteks akhir pada keempat pertanyaan yang diuji. Setelah penerapan aturan dokumen hukum berbasis metadata, potongan tersebut tidak lagi ditemukan pada konteks akhir. Namun, pengaruhnya terhadap jawaban tidak selalu berupa perubahan dari salah menjadi benar. Ringkasan hasil perbandingan ditunjukkan pada Tabel 1.

Tabel 1. Perbandingan Hasil Pengujian Sebelum dan Sesudah Penerapan Registry Peraturan

Pertanyaan	Potongan ketentuan lama dalam konteks akhir (sebelum/sesudah)	Jawaban sebelum	Jawaban sesudah
Batas usia minimum perkawinan	1/0	Laki-laki 19 tahun dan perempuan 16 tahun	Laki-laki dan perempuan 19 tahun
Masa jabatan kepala desa	1/0	8 tahun	8 tahun, paling banyak 2 kali masa jabatan
Jumlah periode jabatan kepala desa	1/0	Menyebut paling banyak 3 kali sebelum menyebut 2 kali	Paling banyak 2 kali
Masa keanggotaan BPD	1/0	8 tahun	8 tahun

Berdasarkan Tabel 1, penerapan Registry Peraturan menunjukkan perubahan yang konsisten pada keempat pertanyaan, yaitu tidak ditemukannya potongan ketentuan lama dalam konteks akhir setelah penerapan aturan berbasis metadata. Meskipun demikian, dampaknya terhadap jawaban berbeda pada setiap pertanyaan. Pada pertanyaan mengenai batas usia perkawinan, perubahan konteks menghasilkan perbaikan jawaban dari ketentuan yang sudah tidak berlaku menjadi ketentuan terbaru. Pada pertanyaan mengenai jumlah periode jabatan kepala desa, jawaban menjadi lebih tepat karena informasi mengenai batas dua kali masa jabatan disampaikan secara langsung tanpa didahului ketentuan lama.

Sementara itu, pada pertanyaan mengenai masa jabatan kepala desa dan masa keanggotaan BPD, jawaban sebelum dan sesudah penerapan Registry Peraturan tetap menunjukkan nilai yang benar, meskipun potongan ketentuan lama masih terdapat dalam konteks sebelum penerapan. Hal ini menunjukkan bahwa keberadaan informasi lama dalam konteks tidak selalu menyebabkan kesalahan jawaban, karena model masih dapat menghasilkan jawaban yang sesuai berdasarkan informasi lain yang tersedia. Dengan demikian, pengaruh Registry Peraturan tidak hanya diamati dari perubahan jawaban, tetapi juga dari kemampuannya mengurangi masuknya potongan ketentuan yang sudah tidak berlaku ke dalam konteks akhir.

Secara keseluruhan, hasil pengujian menunjukkan bahwa penerapan Registry Peraturan membantu mengarahkan proses *retrieval* dengan mempertimbangkan status keberlakuan dokumen hukum, bukan hanya berdasarkan kemiripan semantik. Mekanisme ini menjadi penting pada basis pengetahuan yang memuat peraturan dengan ketentuan yang telah mengalami perubahan, sehingga potongan dari peraturan lama tidak secara otomatis menjadi rujukan utama bagi model LLM. Namun, hasil ini masih terbatas pada empat pertanyaan yang diuji dengan satu percobaan pada setiap kondisi. Oleh karena itu, temuan tersebut menunjukkan pengaruh penerapan Registry Peraturan pada skenario pengujian yang dilakukan dan belum dapat digeneralisasi untuk seluruh jenis pertanyaan layanan desa.

3.3 Hasil Pengujian *Black Box*

Pengujian *black box* dilakukan terhadap fungsi aplikasi web pada sisi masyarakat dan administrator dengan membandingkan keluaran sistem terhadap hasil yang diharapkan pada setiap skenario. Hasil pengujian ditunjukkan pada Tabel 2.

Tabel 2. Hasil Pengujian *Black Box*

No	Fitur yang diuji	Skenario Pengujian	Hasil yang Diharapkan	Status
1	Input pertanyaan	Pengguna memasukkan pertanyaan	Sistem memproses pertanyaan dan menampilkan jawaban	Berhasil
2	*Login*	Admin memasukkan email dan kata sandi yang valid	Sistem menampilkan halaman *dashboard*	Berhasil
3	*Login*	Admin memasukkan email atau kata sandi yang tidak valid	Sistem menampilkan pesan kesalahan *login*	Berhasil
4	Pencarian riwayat percakapan	Admin memasukkan kata kunci pencarian	Sistem menampilkan riwayat sesuai kata kunci	Berhasil
5	Navigasi halaman riwayat	Admin menekan tombol *next* atau nomor halaman	Sistem menampilkan halaman yang dipilih	Berhasil
6	Detail riwayat percakapan	Admin memilih salah satu sesi percakapan	Sistem menampilkan isi percakapan sesuai sesi	Berhasil
7	Pencarian pengetahuan	Admin memasukkan kata kunci pencarian	Sistem menampilkan dokumen sesuai pencarian	Berhasil
8	Tambah pengetahuan	Admin menekan Tambah tanpa mengisi data	Sistem menampilkan pesan validasi	Berhasil
9	Tambah pengetahuan	Admin mengunggah berkas selain PDF	Sistem menampilkan pesan kesalahan	Berhasil
10	Tambah pengetahuan	Admin mengisi seluruh isian dengan benar	Data tersimpan dan tampil pada tabel dokumen	Berhasil
11	Ubah pengetahuan	Admin menekan Simpan tanpa perubahan	Sistem kembali ke daftar tanpa kesalahan	Berhasil
12	Ubah pengetahuan	Admin mengubah judul atau mengunggah berkas baru, lalu menyimpan	Data berhasil diperbarui	Berhasil
13	Hapus pengetahuan	Admin menghapus dokumen tertentu	Data terhapus dan tidak tampil pada daftar	Berhasil
14	Hapus pengetahuan	Admin membatalkan penghapusan dokumen	Data tidak terhapus	Berhasil
15	Navigasi halaman pengetahuan	Admin menekan tombol *next* atau nomor halaman	Sistem menampilkan halaman yang dipilih	Berhasil
16	Daftar registry peraturan	Admin membuka halaman registry peraturan	Sistem menampilkan daftar peraturan beserta jenis, nomor dan tahun, status, verifikasi, dan catatan	Berhasil
17	Ubah status peraturan	Admin mengubah status suatu peraturan dari Berlaku menjadi Diubah	Status berubah dan label Diubah tampil pada daftar	Berhasil
18	Verifikasi status	Admin memverifikasi status peraturan	Kolom verifikasi menunjukkan status telah diverifikasi	Berhasil
19	*Logout*	Admin menekan tombol *logout*	Sistem kembali ke halaman *login*	Berhasil

3.4 Hasil Evaluasi RAGAS

Basis pengetahuan yang terdiri atas 34 dokumen menghasilkan 6.611 *chunk* setelah melalui proses *parsing* dan *chunking*. Evaluasi dilakukan menggunakan 125 pertanyaan uji beserta jawaban referensi untuk membandingkan kinerja LLaMA 3.1 dan Mistral 7B. Kedua model memperoleh pertanyaan dan konteks hasil *retrieval* yang sama, sehingga perbandingan dapat difokuskan pada kemampuan masing-masing model dalam menghasilkan jawaban berdasarkan informasi yang disediakan. Evaluasi menggunakan Qwen2.5:14B sebagai *LLM judge* dan *intfloat/multilingual-e5-base* sebagai model *embedding*. Hasil evaluasi berdasarkan empat metrik RAGAS ditunjukkan pada Tabel 3.

Tabel 3. Hasil Evaluasi RAGAS

Metrik	Mistral 7B	LLaMA 3.1
Faithfulness	0,82	0,84
Answer relevancy	0,82	0,85

Metrik	Mistral 7B	LLaMA 3.1
Context precision	0,81	0,81
Context recall	0,76	0,76

Berdasarkan Tabel 3, LLaMA 3.1 memperoleh nilai *faithfulness* sebesar 0,84, sedikit lebih tinggi dibandingkan Mistral 7B sebesar 0,82. Hasil tersebut menunjukkan bahwa kedua model memiliki kemampuan yang relatif baik dalam menghasilkan jawaban berdasarkan konteks yang diperoleh dari proses *retrieval*. Meskipun demikian, LLaMA 3.1 menunjukkan kesesuaian yang sedikit lebih tinggi antara informasi dalam jawaban dan konteks yang disediakan. Dengan demikian, LLaMA 3.1 cenderung lebih konsisten dalam mempertahankan keterkaitan jawaban dengan informasi pada basis pengetahuan.

Pada metrik *answer relevancy*, LLaMA 3.1 memperoleh nilai 0,85, sedangkan Mistral 7B memperoleh nilai 0,82. Selisih sebesar 0,03 menunjukkan bahwa LLaMA 3.1 memiliki keunggulan dalam menghasilkan jawaban yang relevan terhadap maksud pertanyaan pengguna. Hasil ini mengindikasikan bahwa LLaMA 3.1 cenderung lebih baik dalam menyampaikan informasi yang sesuai dengan kebutuhan pertanyaan. Sementara itu, nilai Mistral 7B yang juga mencapai 0,82 menunjukkan bahwa model tersebut tetap mampu menghasilkan jawaban dengan tingkat relevansi yang baik, meskipun nilainya sedikit lebih rendah dibandingkan LLaMA 3.1.

Berbeda dengan kedua metrik sebelumnya, *context precision* dan *context recall* memperoleh nilai yang sama pada kedua model, masing-masing sebesar 0,81 dan 0,76. Kesamaan ini sesuai dengan rancangan pengujian, karena proses *retrieval* hanya dilakukan satu kali untuk setiap pertanyaan dan hasil konteks yang diperoleh digunakan bersama oleh kedua model. Nilai *context precision* sebesar 0,81 menunjukkan bahwa sebagian besar konteks yang diambil memiliki relevansi terhadap pertanyaan. Sementara itu, nilai *context recall* sebesar 0,76 menunjukkan bahwa konteks hasil *retrieval* telah mencakup sebagian besar informasi yang dibutuhkan berdasarkan jawaban referensi, meskipun masih terdapat informasi relevan yang belum sepenuhnya terjangkau. Perbedaan nilai kedua metrik tersebut juga menunjukkan bahwa ketepatan konteks yang diambil masih lebih baik dibandingkan kelengkapan cakupan informasi yang berhasil diperoleh.

Secara keseluruhan, LLaMA 3.1 memperoleh nilai lebih tinggi pada dua metrik yang mengevaluasi kualitas jawaban, yaitu *faithfulness* dan *answer relevancy*, sedangkan Mistral 7B memiliki nilai yang sedikit lebih rendah pada kedua metrik tersebut. Adapun kesamaan nilai *context precision* dan *context recall* menunjukkan bahwa kedua model dievaluasi menggunakan kualitas konteks hasil *retrieval* yang setara. Berdasarkan hasil tersebut, LLaMA 3.1 menunjukkan kinerja yang lebih baik dalam pengujian kualitas jawaban pada sistem TabeBot. Oleh karena itu, LLaMA 3.1 dipilih sebagai model bahasa yang digunakan dalam implementasi akhir *chatbot* layanan masyarakat Desa Bou.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan TabeBot sebagai *chatbot* layanan masyarakat Desa Bou berbasis *Retrieval-Augmented Generation* (RAG). Sistem dikembangkan menggunakan metode *Rational Unified Process* (RUP) dengan basis pengetahuan sebanyak 34 dokumen PDF yang menghasilkan 6.611 *chunk*. Arsitektur RAG mengintegrasikan *hybrid retrieval* berbasis BM25 dan *dense vector retrieval*, *Cross-Encoder Reranker*, serta Registry Peraturan untuk mengelola informasi hukum berdasarkan status keberlakuan dan perubahan ketentuan. Hasil pengujian *Black Box* terhadap 19 skenario menunjukkan seluruh fungsi yang diuji berjalan sesuai dengan hasil yang diharapkan. Penerapan Registry Peraturan menunjukkan manfaat dalam menyaring konteks yang memuat ketentuan hukum yang sudah tidak berlaku. Pada empat pertanyaan pengujian, *chunk* yang memuat ketentuan lama berhasil dikeluarkan dari konteks akhir yang diberikan kepada model. Salah satu pengujian memperlihatkan perubahan jawaban mengenai batas usia perkawinan menjadi sesuai dengan ketentuan terbaru, yaitu 19 tahun bagi pria dan wanita. Namun, pengujian ini masih terbatas pada empat pertanyaan dengan satu percobaan untuk setiap kondisi sehingga belum dapat mewakili seluruh jenis pertanyaan terkait peraturan. Berdasarkan evaluasi RAGAS terhadap 125 pertanyaan, LLaMA 3.1 memperoleh nilai *faithfulness* sebesar 0,84 dan *answer relevancy* sebesar 0,85, sedangkan Mistral 7B memperoleh nilai masing-masing sebesar 0,82. Kedua model memperoleh nilai *context precision* sebesar 0,81 dan *context recall* sebesar 0,76 karena menggunakan konteks hasil *retrieval* yang sama. Hasil tersebut menunjukkan bahwa LLaMA 3.1 memiliki kinerja lebih baik pada aspek kesesuaian dan relevansi jawaban dalam pengujian ini sehingga dipilih sebagai model bahasa untuk implementasi akhir TabeBot. Penelitian selanjutnya dapat memperluas variasi pertanyaan untuk menguji Registry Peraturan, mengukur waktu respons secara terkontrol, serta mengevaluasi sistem pada skenario penggunaan yang lebih beragam. TabeBot juga berpotensi dikembangkan dan diintegrasikan dengan situs resmi Desa Bou melalui penambahan dokumen layanan serta pembaruan basis pengetahuan secara berkala agar dapat mendukung penyediaan informasi administrasi dan pemerintahan desa berdasarkan dokumen resmi.

REFERENCES

[1] I. Hadi and I. A. P. S. Widnyani, "Modernisasi dan Digitalisasi Public Servis: Mewujudkan Indonesia Emas Melalui Harmonisasi Sistem Pemerintahan Berbasis Elektronik (SPBE)," *Jurnal Kridatama Sains dan Teknologi*, vol. 6, no. 2, pp. 639–658, 2024.

[2] A. Bora and H. Cuayáhuitl, "Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 4, pp. 2355–2374, Dec. 2024.

[3] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2005.11401

[4] H. de Martim, "An Ontology-Driven Graph RAG for Legal Norms: A Structural, Temporal, and Deterministic Approach," Sep. 2025, doi: 10.3233/FAIA251598.

[5] M. Prior, A. Schultz, and M. Grabmair, "Asking For An Old Friend: Diagnosing and Mitigating Temporal Failure Modes in LLM-based Statutory Question Answering," *Proceedings of International Conference on Artificial Intelligence and Law (ICAIL)*, vol. 1, Sep. 2026, [Online]. Available: https://github.com/TUMLegalTech/temporal_bias.

[6] V. Karpukhin *et al.*, "Dense Passage Retrieval for Open-Domain Question Answering," pp. 6769–6781, Nov. 2020, [Online]. Available: https://github.com/facebookresearch/DPR.

[7] A. Y. Yulestiono, A. P. Sari, and A. Rakhmadi, "Evaluasi Hybrid Retrieval Berbasis Reciprocal Rank Fusion untuk Pencarian Informasi pada Dokumen Peraturan Daerah," *Seminar Nasional Informatika Bela Negara (SANTIKA)*, vol. 6, pp. 148–153, 2026.

[8] M. D. Salman, Rahmaddeni, T. Nasution, and Susanti, "Implementation of Retrieval-Augmented Generation Method on Large Language Model for Development of Campus Service and Information Chatbot," *INOVTEK Polbeng-Seri Informatika*, vol. 11, no. 1, pp. 298–309, Feb. 2026.

[9] A. Grattafiori *et al.*, "The Llama 3 Herd of Models," Nov. 2024, [Online]. Available: http://arxiv.org/abs/2407.21783

[10] A. Q. Jiang *et al.*, "Mistral 7B," Oct. 2023.

[11] S. Elysia and Herianto, "Chatbot Berbasis Retrieval Augmented Generation (RAG) untuk Peningkatan Layanan Informasi Sekolah," *Journal TIFDA (Technology Information and Data Analytic)*, vol. 1, no. 2, pp. 52–58, Dec. 2024.

[12] D. Firdaus, I. Sumardi, and Y. Kulsum, "Integrating Retrieval-Augmented Generation with Large Language Model Mistral 7b for Indonesian Medical Herb," *Jurnal Informatika Sunan Kalijaga*, vol. 9, no. 3, pp. 230–243, Sep. 2024.

[13] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated Evaluation of Retrieval Augmented Generation," Apr. 2025.

[14] Ragas, "Available Metrics," Ragas Documentation. Accessed: Nov. 09, 2026. [Online]. Available: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/

[15] Y. I. Chandra, D. R. Irawati, and M. Riastuti, "Penerapan Model Rational Unified Process (RUP) Dalam Membangun Aplikasi Promosi dan Penjualan Berbasis Web (Studi Kasus: Toko Mie Ayam 'Neng Kene Kiyeh')," *Ikraith Informatika*, vol. 8, no. 3, pp. 107–120, Nov. 2024.