Prosiding Seminar Nasional Indonesia e-ISSN: 3026-5169
Vol. 3 No. 3 2026, hal., 103-115

RANCANG BANGUN QUESTION ANSWERING SYSTEM LAYANAN INFORMASI AKADEMIK BERBASIS RAG

Muhammad Surya Adhi Setiawan
Universitas Pembangunan Nasional "Veteran" Jawa Timur
Correspondence author email: 21082010167@student.upnjatim.ac.id

Abstract

The complexity of academic information in the New Student Admissions (PPMB) process often overwhelms conventional helpdesk services. This study aims to design and build a prototype Question Answering (QA) System based on Retrieval-Augmented Generation (RAG) that can automate information services accurately. The system is built using a three-layer architecture: Presentation Layer (Gradio UI), Application Layer (Python/LangChain), and Data Layer (ChromaDB). A key focus of the development is the data pipeline strategy, specifically handling "Indivisible Information Units" in PDF tables by setting a dynamic chunking limit of 3000 tokens. The prototype features a Knowledge Base Manager for dynamic document updates and a multilingual Chat Interface. Testing demonstrates the system's ability to process heterogeneous data from 30 sources and successfully retrieve specific procedural information, such as the "Golden Ticket" requirements, with high precision. The system is deployed using a reasoning model engine to ensure logical answer synthesis.

Keywords: *Information Retrieval*, *Prototype*, *RAG Architecture*, *Software Engineering*, *System Implementation*.

Abstrak

Kompleksitas informasi akademik pada proses Penerimaan Peserta Mahasiswa Baru (PPMB) sering kali membebani layanan *helpdesk* konvensional. Penelitian ini bertujuan untuk merancang dan membangun prototipe *Question Answering* (QA) *System* berbasis *Retrieval-Augmented Generation* (RAG) yang mampu mengotomatisasi layanan informasi secara akurat. Sistem dibangun menggunakan arsitektur tiga lapis (*three-tier architecture*): *Presentation Layer* (Antarmuka Gradio), *Application Layer* (Python/LangChain), dan *Data Layer* (ChromaDB). Fokus utama pengembangan adalah pada strategi *pipeline* data, khususnya penanganan "Unit Informasi Tak Terpisahkan" (*Indivisible Information Units*) pada tabel PDF dengan menetapkan batas *chunking* dinamis sebesar 3000 token. Prototipe dilengkapi dengan fitur *Knowledge Base Manager* untuk pembaruan dokumen secara dinamis dan *Chat Interface* multibahasa. Pengujian fungsional menunjukkan kemampuan sistem dalam memproses data heterogen dari 30 sumber dan berhasil melakukan *retrieval* informasi prosedural spesifik, seperti persyaratan "Golden Ticket", dengan presisi tinggi. Sistem di-*deploy* menggunakan mesin *reasoning model* untuk menjamin sintesis jawaban yang logis.

Kata Kunci : Arsitektur RAG, Implementasi Sistem, *Information Retrieval*, Prototipe, Rekayasa Perangkat Lunak.

PENDAHULUAN

Transparansi dan aksesibilitas informasi merupakan pilar utama dalam tata kelola institusi pendidikan tinggi modern (Jongbloed dkk., 2018). Di Indonesia, kewajiban ini dipertegas melalui Undang-Undang Nomor 14 Tahun 2008 tentang Keterbukaan Informasi Publik, yang memandatkan setiap badan publik, termasuk Perguruan Tinggi Negeri (PTN), untuk menyediakan informasi yang akurat, benar, dan tidak menyesatkan (UU KIP, 2008). Dalam ekosistem universitas, periode Penerimaan Peserta Mahasiswa Baru (PPMB) merupakan fase dengan intensitas pertukaran informasi tertinggi. Universitas Pembangunan Nasional "Veteran" Jawa Timur (UPN "Veteran" Jatim), sebagai institusi dengan ribuan pendaftar setiap tahunnya, menghadapi tantangan operasional yang signifikan dalam mendiseminasikan informasi yang kompleks dan dinamis (upnhumas, 2025). Informasi tersebut mencakup rincian jalur seleksi (SNBP, SNBT, Mandiri), persyaratan spesifik program studi, daya tampung, hingga struktur Biaya Kuliah Tunggal (BKT) dan Uang Kuliah Tunggal (UKT) yang bervariasi (ppmb.upnjatim, 2025).

Ekosistem layanan informasi perguruan tinggi saat ini masih sangat bergantung pada arsitektur konvensional yang tidak lagi relevan dengan perilaku digital pengguna. Informasi disebarkan melalui dokumen statis (PDF) di situs web atau melalui layanan *helpdesk* yang dioperasikan oleh staf manusia. Pendekatan ini memiliki kelemahan inheren yang mendasar. Dokumen PDF akademik, seperti Surat Keputusan Rektor tentang penetapan UKT, sering kali berisi tabel-tabel panjang yang sulit dibaca pada perangkat seluler (Noyes, 2019; Tensmeyer dkk., 2023). Di sisi lain, layanan *helpdesk* manual dibatasi oleh kapasitas biologis staf yang rentan terhadap kelelahan kognitif (*cognitive fatigue*) akibat beban kerja mental yang tinggi, yang secara langsung menurunkan performa dan kecepatan respons. Ketika terjadi lonjakan volume pertanyaan (*seasonal spikes*) selama periode penerimaan, terjadi *bottleneck* komunikasi yang berakibat pada keterlambatan respons (Mahdavi dkk., 2024). Latensi respons yang timbul dari *bottleneck* ini bukan sekadar kendala teknis, melainkan pemicu psikologis utama yang memperburuk kecemasan dan ketidakpastian calon mahasiswa, yang secara empiris berkorelasi negatif dengan keputusan pendaftaran mereka (Aunul dkk., 2022; Bauer-Wolf, 2023). Oleh karena itu, diperlukan sebuah solusi teknologi yang mampu mengotomatisasi layanan ini tanpa mengorbankan akurasi.

Evolusi teknologi *Artificial Intelligence* (AI), khususnya dalam bidang *Natural Language Processing* (NLP), telah menghadirkan *Large Language Models* (LLM) berbasis *transformer* seperti GPT-4 dan Gemini. Model-model ini memiliki kemampuan linguistik yang luar biasa untuk memahami dan menghasilkan teks layaknya manusia (Yin dkk., 2024). Namun, penggunaan LLM secara langsung (*direct generation*) untuk layanan informasi akademik mengandung risiko fatal. LLM menderita fenomena "halusinasi", dimana model dapat mengarang fakta dengan sangat meyakinkan ketika tidak memiliki akses ke data yang benar (Ni dkk., 2025). Misalnya, sebuah LLM mungkin dengan

percaya diri menyebutkan nominal UKT yang salah karena data pelatihannya sudah kadaluwarsa. Dalam konteks administrasi publik dan akademik, kesalahan informasi sekecil apa pun tidak dapat ditoleransi karena menyangkut hak dan kewajiban finansial mahasiswa (Pulkundwar dkk., 2025).

Untuk mengatasi kelemahan tersebut, arsitektur *Retrieval-Augmented Generation* (RAG) lebih efektif dibandingkan *fine-tuning* dalam menangani data yang bersifat dinamis (Lewis dkk., 2020). RAG bekerja dengan mekanisme melakukan pencarian (*retrieval*) dokumen relevan dari basis pengetahuan internal (*knowledge base*) terlebih dahulu sebelum menjawab pertanyaan. Dokumen yang ditemukan kemudian dijadikan konteks bagi LLM untuk menyusun jawaban (Guu dkk., 2020). Dengan demikian, jawaban sistem selalu *grounded* pada fakta yang valid (Gupta dkk., 2024). Namun, implementasi RAG pada domain akademik tidak sesederhana menghubungkan basis data dengan LLM. Terdapat tantangan rekayasa (*engineering challenges*) yang spesifik terkait karakteristik data akademik.

Masalah utama dalam dokumen akademik adalah adanya *large indivisible information units*. Dokumen akademik tidak hanya berisi paragraf naratif, tetapi juga tabel data yang sangat padat dan panjang. Sebagai contoh, tabel kuota penerimaan mahasiswa baru per program studi atau matriks biaya UKT sering kali membentang hingga beberapa halaman. Teknik *chunking* (pemotongan dokumen) konvensional yang membagi teks berdasarkan jumlah karakter tetap (misalnya per 500 karakter) sering kali merusak struktur tabel ini. Pemotongan sembarangan dapat memisahkan *header* kolom dari baris data isinya, sehingga ketika potongan tersebut diambil oleh mesin pencari, konteks datanya hilang. Jika *retriever* gagal menyajikan tabel yang utuh, maka LLM tercanggih sekalipun tidak akan mampu menjawab pertanyaan komparatif atau kuantitatif dengan benar.

Selain masalah struktur data, tantangan lainnya adalah kemampuan penalaran sistem. Pertanyaan calon mahasiswa sering kali bersifat prosedural dan kondisional, seperti "Apakah saya bisa mendaftar jalur mandiri prestasi jika sertifikat saya tingkat kabupaten?". Menjawab pertanyaan ini memerlukan lebih dari sekadar pencarian kata kunci; diperlukan penalaran logis (*reasoning*) untuk menghubungkan aturan dalam dokumen dengan kondisi pengguna (Oche dkk., 2025). Generasi terbaru LLM, yang dikenal sebagai *reasoning models* seperti OpenAI o *series* atau Gemini *thinking series*, menawarkan kemampuan *Chain-of-Thought* (CoT) yang menjanjikan untuk menangani kompleksitas ini (Kane, 2025; OpenAI, 2024). Namun, integrasi model jenis ini ke dalam arsitektur RAG untuk layanan publik masih jarang dieksplorasi (Firdaus dkk., 2024; Tohir dkk., 2024).

Penelitian ini bertujuan untuk mengisi kesenjangan teknis tersebut dengan merancang dan membangun sebuah prototipe *question answering system* berbasis RAG yang secara khusus direkayasa untuk menangani kompleksitas data akademik PPMB UPN "Veteran" Jatim. Fokus penelitian ini diletakkan pada aspek rekayasa sistem

(*system engineering*) yang meliputi perancangan *pipeline* data yang mampu menangani tabel PDF kompleks melalui strategi *semantic markdown chunking*, implementasi arsitektur sistem modular yang memisahkan lapisan presentasi, aplikasi, dan data, serta pemanfaatan *reasoning models* untuk meningkatkan akurasi logika jawaban. Luaran dari penelitian ini adalah perangkat lunak yang fungsional dan teruji, yang dapat mendemonstrasikan kelayakan teknis transformasi layanan informasi akademik dari manual menjadi otomatis dan cerdas.

METODE PENELITIAN

Penelitian ini menggunakan paradigma *software engineering* dengan model pengembangan *prototype*. Pendekatan ini dipilih untuk memungkinkan iterasi cepat dalam menguji efektivitas arsitektur sistem (Pressman & Maxim, 2019). Tahapan penelitian dirancang secara sistematis, mulai dari analisis data, perancangan arsitektur, implementasi kode, hingga *deployment* sistem.

Analisis Karakteristik Data dan Sumber Informasi

Langkah fundamental dalam membangun sistem RAG adalah memahami data yang akan dikelola. Berdasarkan analisis terhadap kebutuhan informasi calon mahasiswa baru, penelitian ini mengidentifikasi dan mengumpulkan 30 sumber data primer yang merepresentasikan *ground truth* informasi PPMB UPN "Veteran" Jatim. Sumber data ini dikategorikan menjadi tiga jenis format yang memiliki tantangan pemrosesan berbeda:

1. **Dokumen PDF:** Mencakup dokumen legal seperti Surat Keputusan Rektor, Pedoman Pendaftaran KIP Kuliah, dan Panduan Portofolio. Karakteristik teknis dokumen ini adalah *multi-column layout*, mengandung tabel dengan *merged-cells*, dan sering kali memiliki *header/footer* berulang. Dokumen jenis ini adalah yang paling sulit diproses oleh mesin karena struktur visualnya tidak linier dengan struktur teksnya.
2. **Halaman Web:** Informasi profil dan fasilitas yang diambil langsung dari situs ppmb.upnjatim.ac.id. Tantangan utamanya adalah membersihkan elemen navigasi, iklan, dan *script* agar hanya konten relevan yang tersisa.
3. **Data API:** Data statistik seperti keketatan persaingan dan daya tampung yang diambil dari pangkalan data pendidikan tinggi. Data ini berformat JSON dan sangat terstruktur, namun perlu dinarasikan agar dapat dipahami oleh model bahasa.

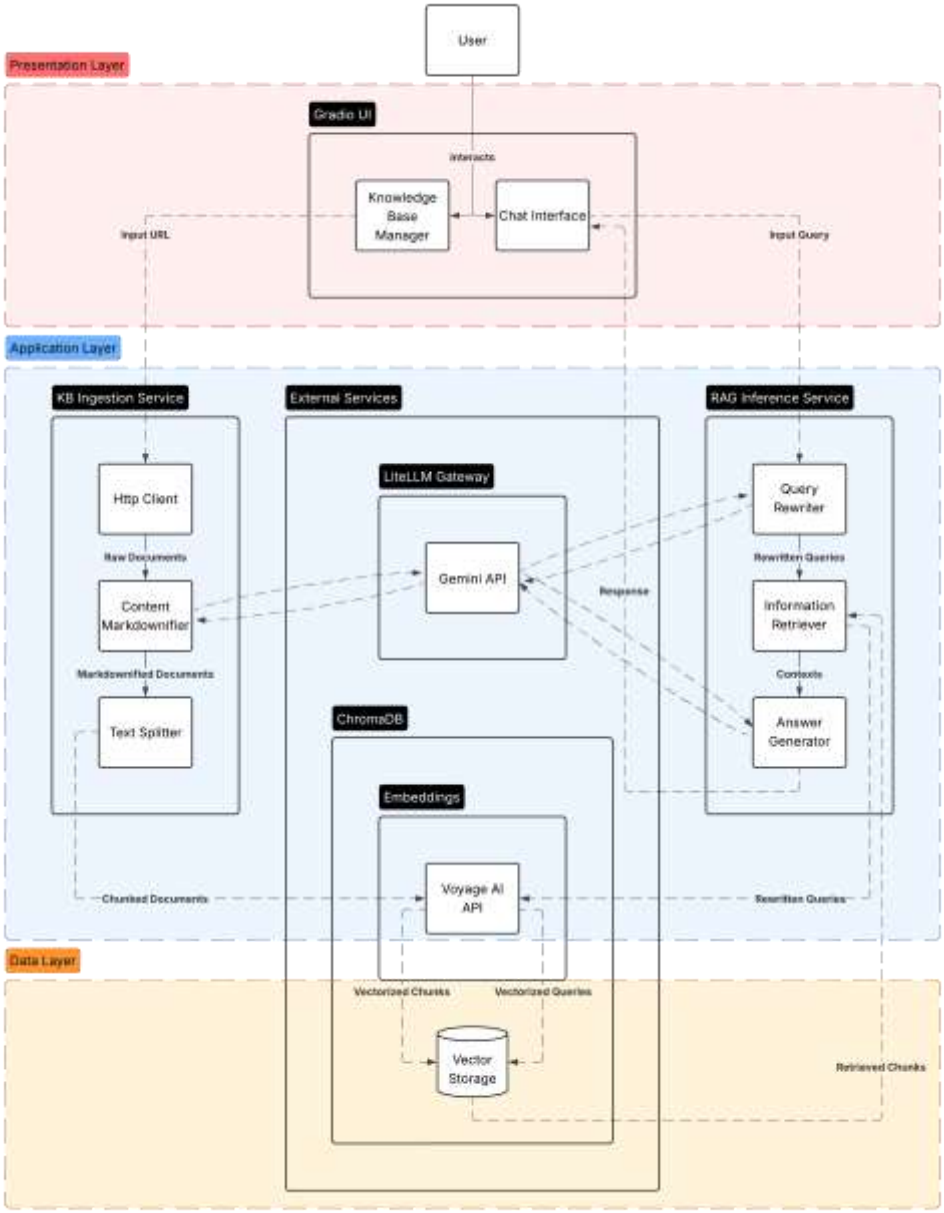
Data Engineering Pipeline

Untuk mengubah data mentah yang heterogen tersebut menjadi basis pengetahuan yang dapat dicari (*searchable knowledge base*), dikembangkan sebuah *pipeline* data otomatis menggunakan bahasa pemrograman Python. *Pipeline* ini terdiri dari tiga tahap krusial:

1. **Intelligent Acquisition:** Dikembangkan modul fetch_content yang menggunakan pustaka httpx. Modul ini dilengkapi dengan logika penanganan URL khusus. Mengingat banyak dokumen akademik disimpan di Google Drive, sistem menanamkan fungsi identify_drive_url yang menggunakan *regular expression* (RegEx) untuk mendeteksi pola URL Google Drive dan mengubahnya dari mode pratinjau (*preview*) menjadi mode unduhan langsung (*direct download*). Tanpa logika ini, sistem hanya akan mengunduh halaman HTML pembungkus Google Drive, bukan isi dokumen PDF-nya.
2. **Markdown Conversion:** Ini adalah inovasi teknis utama dalam pra-pemrosesan. Metode ekstraksi teks tradisional (seperti PyPDF2) sering menghancurkan struktur tabel. Penelitian ini menggunakan pendekatan *vision-based parsing* dengan memanfaatkan API gemini-2.5-flash-lite. Dokumen PDF dikirim ke model sebagai *input* visual, dan model diinstruksikan melalui *system prompt* khusus untuk menulis ulang konten tersebut ke dalam format Markdown. Format Markdown dipilih karena efisiensi tokennya dan kemampuannya merepresentasikan hierarki dokumen (Header 1 #, Header 2 ##) serta struktur tabel (|...|) yang dapat dipahami oleh mesin (Donghun Shin dkk., 2024).
3. **Semantic Chunking:** Setelah menjadi Markdown, dokumen harus dipecah (*chunking*). Untuk mengatasi masalah *indivisible information unit* pada tabel besar, diterapkan *threshold chunking* sebesar 3000 token. Ambang batas ini ditetapkan berdasarkan analisis empiris terhadap tabel terbesar dalam korpus data (Tabel Tarif UKT) yang berukuran sekitar 2.500 token. Pemecahan dilakukan menggunakan MarkdownHeaderTextSplitter, yang membagi dokumen berdasarkan *header* logisnya, bukan sekadar jumlah karakter.

Three-Tier Architecture

Sistem dibangun dengan desain arsitektur modular untuk memastikan skalabilitas dan kemudahan pemeliharaan (*maintainability*). Arsitektur ini terdiri dari:



Gambar 1. Arsitektur Sistem

1. **Presentation Layer:** Dibangun menggunakan kerangka kerja Gradio. Lapisan ini menyediakan antarmuka pengguna berbasis web yang responsif. Terdapat dua modul antarmuka yaitu Knowledge Base Manager yang merupakan antarmuka CRUD (*Create*, *Read*, *Update*, *Delete*) yang memungkinkan administrator sistem (*non-programmer*) untuk menambah atau menghapus dokumen referensi cukup dengan memasukkan URL. Sedangkan antarmuka Chat Interface merupakan

antarmuka percakapan untuk pengguna akhir yang mendukung sitasi sumber terkait.

2. **Application Layer:** Merupakan otak dari sistem yang ditulis dengan Python dan LangChain. Lapisan ini terdiri dari Query Rewriter sebagai modul yang menggunakan LLM untuk memperbaiki pertanyaan pengguna yang ambigu atau informal sebelum dilakukan pencarian, Retriever Engine sebagai modul yang melakukan pencarian vektor, dan Answer Generator sebagai modul yang menyusun jawaban akhir menggunakan *reasoning model*.
3. **Data Layer:** Menggunakan ChromaDB sebagai basis data vektor (*vector store*). Lapisan ini menyimpan *embeddings* (representasi vektor) dari dokumen yang dihasilkan oleh model Voyage-3-large. Pemilihan VoyageAI didasarkan pada peringkatnya yang tinggi pada tugas *retrieval* dalam Massive Multilingual Text Embedding Benchmark (MTEB), yang krusial untuk menangani dokumen berbahasa Indonesia.

Implementasi Mesin Inferensi

Sebagai inti kecerdasan sistem, penelitian ini mengimplementasikan model gemini-2.5-flash. Model ini dipilih karena termasuk dalam kategori *reasoning models* yang memiliki kemampuan *internal thinking process*. Berbeda dengan LLM standar yang memprediksi kata demi kata secara langsung, model ini melakukan evaluasi logika terhadap konteks yang diberikan sebelum menghasilkan *output* (Kane, 2025). Hal ini sangat vital untuk menjawab pertanyaan prosedural akademik yang sering kali memiliki syarat dan ketentuan bertingkat.

HASIL DAN PEMBAHASAN

Bagian ini memaparkan hasil implementasi teknis, evaluasi kinerja pipeline data, serta validasi fungsional sistem melalui studi kasus dan pengujian kuantitatif.

Analisis Kinerja Pipeline Data dan Integritas Pengetahuan

Keberhasilan fundamental dari sistem yang dibangun terletak pada kualitas *knowledge base* yang dihasilkan. Proses transformasi data dari PDF ke Markdown melalui pendekatan *vision-based parsing* terbukti efektif dalam menjaga integritas struktur informasi. Sebagai studi kasus, dokumen "Informasi Umum SNPMB 2025" yang memuat tabel kuota penerimaan SNBP, SNBT, Mandiri berhasil diproses dengan sempurna.

Pada metode konvensional, tabel yang kompleks sering kali terfragmentasi, menyebabkan hilangnya konteks antara *header* dan isi tabel. Dengan menerapkan batas *threshold* sebesar 3000 token, sistem mampu memastikan unit informasi tak terpisahkan (*indivisible information unit*), seperti Tabel Tarif UKT atau Kuota Penerimaan, tetap berada dalam satu blok *chunk*.

Berdasarkan pengujian terhadap 30 sumber data primer, dihasilkan total 85 *chunks* yang padat informasi. Hal ini membuktikan bahwa *strategi semantic markdown chunking* meminimalkan risiko kegagalan *retrieval* yang disebabkan oleh hilangnya konteks struktural pada dokumen akademik yang padat tabel.

Penelusuran Alur RAG: Studi Kasus "Golden Ticket"

Untuk memvalidasi efektivitas arsitektur sistem secara menyeluruh, dilakukan penelusuran (*tracing*) terhadap pemrosesan sebuah *query* kompleks. Pertanyaan uji yang digunakan adalah: "haloo kak untuk dapetin golden ticket nya tuh gimana ya? TO apa prestasi apa cuma dateng aja ya kak?". Pertanyaan ini sengaja dipilih karena mengandung bahasa informal, singkatan ambigu, dan miskonsepsi pengguna.

1. **Tahap Query Rewriting**

   Modul *rewriter* menerima *input* mentah tersebut dan berhasil memetakan maksud pengguna dengan menulis ulang pertanyaan menjadi tiga variasi formal seperti berikut:

   Tabel 1. Hasil Query Rewriting

ID	Rewritten Query
0	haloo kak untuk dapetin golden ticket nya tuh gimana ya? TO apa prestasi apa cuma dateng aja ya kak?
1	Bagaimana cara mendapatkan golden ticket di UPN "Veteran" Jawa Timur?
2	Apa saja persyaratan untuk mendapatkan golden ticket di UPN "Veteran" Jawa Timur?
3	Informasi mengenai golden ticket UPN "Veteran" Jawa Timur jalur try out atau prestasi.

   Proses ini memastikan bahwa meskipun pengguna menggunakan istilah ambigu seperti "TO" (Try Out) atau kata tidak baku "dapetin", mesin pencari tetap dapat memahami maksud pengguna. Tanpa langkah ini, pencarian tidak akan menghasilkan dokumen relevan.

2. **Tahap Retrieval**

   Sistem melakukan pencarian vektor pada ChromaDB menggunakan *embedding model* dari VoyageAI. Hasil *trace* menunjukkan sistem berhasil mengambil 10 dokumen dengan relevansi tertinggi.

   Tabel 2. Hasil Retrieval

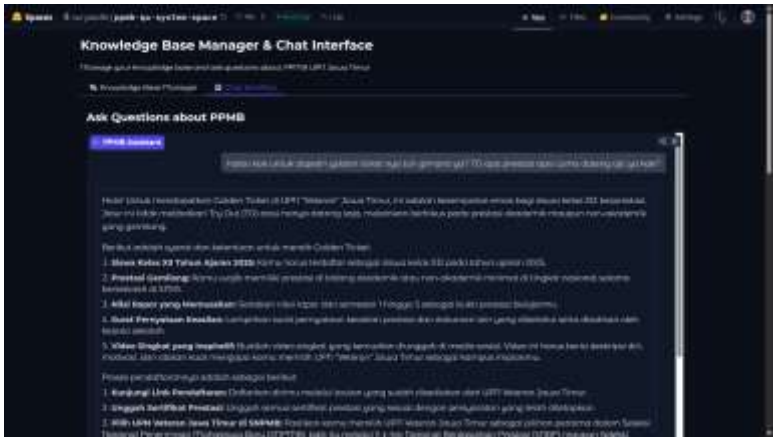
ID	Retrieved Chunks
0	Golden Ticket UPN Jatim 2025-0
1	Seleksi Mandiri Prestasi UPN Jatim 2025-0

2	Informasi SNBP 2025-0
3	Informasi Umum SNPMB 2025-3
4	Seleksi Mandiri Rapor UPN Jatim 2025-0
5	Mandiri UPN Jatim 2025-0
6	Seleksi Mandiri CBT UPN Jatim 2025-2
7	PPMB UPN Jatim 2025-0
8	Seleksi Mandiri UTBK UPN Jatim 2025-0
9	SNBT UPN Jatim 2025-1

Dokumen dengan skor relevansi tertinggi adalah "Golden Ticket UPN Jatim 2025" dan "Seleksi Mandiri Prestasi". Menariknya, sistem juga mengambil dokumen tentang "SNBP". Hal ini menunjukkan bahwa *embedding model* VoyageAI mampu menangkap hubungan semantik bahwa "Golden Ticket" adalah bagian dari mekanisme seleksi prestasi, meskipun kata kuncinya tidak persis sama. Mekanisme deduplikasi di lapisan aplikasi juga bekerja dengan baik, menyaring duplikasi sehingga hanya *chunks* unik yang diteruskan ke tahap selanjutnya.

3. Tahap Generation

Model gemini-2.5-flash menerima *prompt* yang berisi pertanyaan asli dan potongan dokumen yang ditemukan. Hasil jawaban menunjukkan proses penalaran kuat yang dapat dilihat pada Gambar 2.

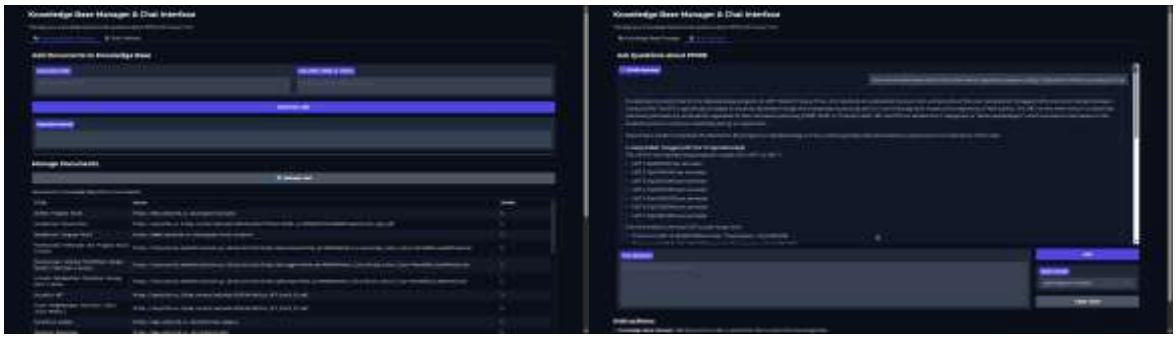


Gambar 2. Hasil Jawaban Sistem

Sistem secara eksplisit mengoreksi asumsi pengguna: " Jalur ini tidak melibatkan Try Out (TO) atau hanya datang saja, melainkan berfokus pada prestasi akademik maupun non-akademik yang gemilang." Sistem kemudian merinci persyaratan: Siswa Kelas XII, Prestasi tingkat Nasional/Internasional, dan sebagainya.

Jawaban diakhiri dengan daftar referensi sumber. Kemampuan sistem untuk melakukan negasi dan klarifikasi merupakan indikator bahwa *reasoning model* bekerja dengan baik.

Evaluasi Antarmuka dan Manajemen Sistem



Gambar 3. Antarmuka Sistem

Prototipe sistem yang di-*deploy* pada *presentation layer* menunjukkan kegunaan (*usability*) yang tinggi untuk operasional kampus. Fitur *Knowledge Base Manager* berhasil mendemonstrasikan alur kerja "*Zero-Code Update*". Dalam simulasi, ketika URL dokumen baru dimasukkan, sistem membutuhkan waktu rata-rata 15-30 detik untuk mengunduh, mengonversi, dan mengindeks dokumen tersebut. Setelah proses selesai, informasi baru tersebut langsung dapat ditanyakan melalui *Chat Interface*. Fitur ini memecahkan masalah klasik perangkat lunak pesanan yang sering kali sulit dipelihara kontennya oleh staf non-teknis. Dengan fitur ini, staf admisi UPN "Veteran" Jatim dapat secara mandiri menjaga kemutakhiran informasi sistem.

Selain itu, kapabilitas multibahasa sistem terbukti berjalan lancar. Meskipun basis data dokumen sepenuhnya dalam Bahasa Indonesia, sistem mampu melayani pertanyaan dalam Bahasa Inggris dan Bahasa Jawa dialek Suroboyoan. Pada pengujian dengan *input* Bahasa Jawa: "Syarat e opo ae cak?", sistem memahami konteks dari riwayat percakapan sebelumnya, mencari dokumen Bahasa Indonesia, dan menjawab kembali dalam Bahasa Jawa. Fleksibilitas ini sangat krusial untuk melayani demografi pendaftar yang beragam secara kultural.

Analisis Kualitas Jawaban

Untuk mengukur efektivitas sistem dalam menyintesis jawaban, dilakukan evaluasi kuantitatif menggunakan tiga metrik utama: *Faithfulness*, *Factual Correctness*, dan *Semantic Similarity*. Hasil pengujian pada 1059 sampel pertanyaan lintas bahasa disajikan pada Tabel 3.

Tabel 3. Hasil Evaluasi Performa Generasi Jawaban Lintas Bahasa

Bahasa	Faithfulness	Factual Correctness	Semantic Similarity

English	0.9404	0.6520	0.9230
Indonesia	0.9250	0.6261	0.9255
Suroboyoan	0.8923	0.6040	0.8962

Skor *Faithfulness* yang tinggi di seluruh bahasa membuktikan bahwa arsitektur RAG yang diimplementasikan sangat efektif dalam mencegah fenomena halusinasi. Skor 0,9404 pada Bahasa Inggris dan 0,9250 pada Bahasa Indonesia menunjukkan bahwa hampir seluruh jawaban yang dihasilkan oleh *reasoning models* didasarkan secara ketat pada potongan dokumen yang diambil dari ChromaDB, bukan dari pengetahuan internal model yang mungkin sudah usang.

Skor *Factual Correctness* berada di rentang 0,60 hingga 0,65. Meskipun terlihat lebih rendah dibanding metrik lainnya, angka ini mencerminkan kompleksitas data akademik yang memerlukan pembacaan angka yang sangat presisi. Rendahnya skor ini pada dialek Suroboyoan (0,6040) dipengaruhi oleh variasi terminologi non-formal yang terkadang membuat pemetaan logika pada tabel PDF menjadi lebih menantang bagi model. Namun, skor ini tetap jauh lebih unggul dibandingkan penggunaan LLM murni tanpa basis pengetahuan (RAG).

Skor *Semantic Similarity* yang stabil di angka 0,89-0,92 menunjukkan bahwa meskipun pengguna bertanya menggunakan bahasa yang berbeda, sistem mampu menangkap esensi pertanyaan dan memberikan jawaban yang secara makna konsisten dengan informasi resmi. Hal ini memvalidasi penggunaan *embedding model* VoyageAI dan gemini-2.5-flash dalam mempertahankan pemahaman kontekstual yang kuat, yang sangat krusial untuk melayani demografi calon mahasiswa yang beragam secara kultural.

KESIMPULAN

Penelitian ini telah berhasil merancang dan membangun sebuah prototipe *question answering system* layanan informasi akademik yang fungsional. Keberhasilan sistem ini tidak terlepas dari pendekatan rekayasa perangkat lunak yang holistik, mulai dari penanganan data di hulu hingga presentasi informasi di hilir.

Kesimpulan teknis utama dari penelitian ini adalah bahwa kualitas sistem RAG sangat bergantung pada strategi rekayasa data (*data engineering*). Penerapan *semantic markdown chunking* dengan batas 3000 token terbukti menjadi solusi efektif untuk menjinakkan kompleksitas dokumen akademik yang kaya tabel. Tanpa strategi ini, integritas informasi akan terfragmentasi, menyebabkan kegagalan sistemik pada tahap *retrieval*. Selain itu, arsitektur modular tiga lapis yang dilengkapi dengan *knowledge base manager* memberikan fleksibilitas operasional yang tinggi, memungkinkan sistem untuk beradaptasi dengan dinamika informasi PPMB yang cepat berubah tanpa memerlukan intervensi kode ulang.

Penggunaan *reasoning models* sebagai mesin inferensi juga terbukti memberikan nilai tambah yang signifikan dalam hal akurasi logika dan kemampuan sintesis jawaban, menjadikan sistem ini lebih dari sekadar mesin pencari, melainkan asisten cerdas yang mampu memberikan konsultasi prosedural. Prototipe yang dihasilkan siap untuk dikembangkan lebih lanjut menjadi sistem produksi, memberikan kontribusi nyata bagi peningkatan kualitas layanan publik di lingkungan pendidikan tinggi.

DAFTAR PUSTAKA

Aunul, S., Handayani, F., & Riswandi, R. (2022). Uncertainty Reduction of First-Year College Students in Virtual Class. *CHANNEL: Jurnal Komunikasi*, *10*(1), 21–26. https://doi.org/10.12928/channel.v10i1.22088

Bauer-Wolf, J. (2023, Agustus 25). *Over half of students rank college applications as their most stressful academic experience, survey finds | Higher Ed Dive*. https://www.highereddive.com/news/over-half-of-students-rank-college-applications-as-their-most-stressful-aca/691808/

Donghun Shin, Xigui Li, Li, H., Shaojie Shi, Kaitao Chen, & Daocheng Fu. (2024). *Prompt Engineering and Format on LLMs in the Financial Domain*. https://doi.org/10.13140/RG.2.2.17057.11365

Firdaus, D., Sumardi, I., & Kulsum, Y. (2024). Integrating Retrieval-Augmented Generation with Large Language Model Mistral 7b for Indonesian Medical Herb. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *9*(3), Article 3. https://doi.org/10.14421/jiska.2024.9.3.230-243

Gupta, S., Ranjan, R., & Singh, S. N. (2024). *A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions* (arXiv:2410.12837). arXiv. https://doi.org/10.48550/arXiv.2410.12837

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning, ICML'20*, *119*, 3929–3938. https://dl.acm.org/doi/10.5555/3524938.3525306

Jongbloed, B., Vossensteyn, H., van Vught, F., & Westerheijden, D. F. (2018). Transparency in Higher Education: The Emergence of a New Perspective on Higher Education Governance. Dalam A. Curaj, L. Deca, & R. Pricopie (Ed.), *European Higher Education Area: The Impact of Past and Future Policies* (hlm. 441–454). Springer International Publishing. https://doi.org/10.1007/978-3-319-77407-7_27

Kane, P. (2025, Februari 5). *Access the latest 2.0 experimental models in the Gemini app.* Google. https://blog.google/feed/gemini-app-experimental-models/

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Mahdavi, N., Tapak, L., Darvishi, E., Doosti-Irani, A., & Shafiee Motlagh, M. (2024). Unraveling the interplay between mental workload, occupational fatigue, physiological responses and cognitive performance in office workers. *Scientific Reports*, *14*, 17866. https://doi.org/10.1038/s41598-024-68889-4

Ni, B., Liu, Z., Wang, L., Lei, Y., Zhao, Y., Cheng, X., Zeng, Q., Dong, L., Xia, Y., Kenthapadi, K., Rossi, R., Dernoncourt, F., Tanjim, M. M., Ahmed, N., Liu, X., Fan, W., Blasch, E., Wang, Y., Jiang, M., & Derr, T. (2025). *Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey* (arXiv:2502.06872). arXiv. https://doi.org/10.48550/arXiv.2502.06872

Noyes, D. (2019). *Examining the Usability of Content in Canvas: HTML vs. PDF*.

Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). *A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions* (arXiv:2507.18910). arXiv. https://doi.org/10.48550/arXiv.2507.18910

OpenAI. (2024, September 12). *Learning to reason with LLMs*. https://openai.com/index/learning-to-reason-with-llms/

ppmb.upnjatim. (2025). *Pusat Penerimaan Mahasiswa Baru*. https://ppmb.upnjatim.ac.id/

Pressman, R. S., & Maxim, B. R. (2019). *Software Engineering: A Practitioner's Approach*. McGraw-Hill Education.

Pulkundwar, P., Dhanawade, V., Yadav, R., Sonkar, M., Asurlekar, M., & Rathod, S. (2025). *A Concise Review of Hallucinations in LLMs and their Mitigation* (arXiv:2512.02527). arXiv. https://doi.org/10.48550/arXiv.2512.02527

Tensmeyer, C., Bylinski, Z., Cai, T., Miller, D., Nenkova, A., Niklaus, A., & Wallace, S. (2023). Web Table Formatting Affects Readability on Mobile Devices. *Proceedings of the ACM Web Conference 2023, WWW '23*, 1334–1344. https://doi.org/10.1145/3543507.3583506

Tohir, H., Merlina, N., & Haris, M. (2024). Utilizing Retrieval-Augmented Generation in Large Language Models to Enhance Indonesian Language NLP. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *10*(2), 352–360. https://doi.org/10.33480/jitk.v10i2.5916

Undang-Undang Nomor 14 Tahun 2008 tentang Keterbukaan Informasi Publik (2008). Tambahan Lembaran Negara Nomor 4846

upnhumas. (2025, Agustus 16). UPN Veteran Jawa Timur Sambut 6.662 Mahasiswa Baru, Resmikan PKKMB 2025 di Menara Wimaya Twin Tower. *UPN "Veteran" Jawa Timur*. https://upnjatim.ac.id/2025/08/16/upn-veteran-jawa-timur-sambut-6-662-mahasiswa-baru-resmikan-pkkmb-2025-di-menara-wimaya-twin-tower/

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. *National Science Review*, *11*(12), nwae403. https://doi.org/10.1093/nsr/nwae403