KAMPUS AKADEMIK PUBLISING
Jurnal Ilmiah Penelitian Mahasiswa (JIPM)
Vol.4, No.5 Oktober 2026
e-ISSN: 3025-5465, p-ISSN : 3025-7964, Hal 1454-1465
DOI: https://doi.org/10.61722/jipm.v4i5.3074

RANCANG BANGUN SISTEM RAG UNTUK MANAJEMEN DOKUMENTASI AKREDITASI PROGRAM STUDI

Kyan Dillan Verado[1], Ryan Putranda Kristianto[2]
[1,2]Program Studi Ilmu Informatika, Fakultas Teknik, Universitas Katolik Darma Cendika, Surabaya
*Penulis Korespondensi: [1]kyan.verado@student.ukdc.ac.id, [2]ryan@ukdc.ac.id

Abstract. *Study program accreditation involves a large and complex set of regulatory documents, which makes it difficult for administrators to retrieve accurate information quickly. This study designs and develops a Question-Answering (QA) system based on Retrieval-Augmented Generation (RAG) for managing accreditation documentation of the Informatics study program at Universitas Katolik Darma Cendika (UKDC). The system integrates 84 regulatory documents extracted using PyMuPDF into 4,613 text chunks indexed in the ChromaDB vector database. The RAG architecture employs a fine-tuned BGE-M3 embedding model trained with the sme-embeddings contrastive-learning method (MultipleNegativesRankingLoss) to optimize domain-specific document retrieval, together with GPT-4o-mini as the answer generator, benchmarked against GLM-4.7-Flash. Retrieval evaluation shows a significant improvement after fine-tuning, with Recall@1 rising from 0.8100 to 0.9300, NDCG@1 from 0.8100 to 0.9300, and MRR@10 from 0.8757 to 0.9603. BERTScore evaluation shows the F1-Score of GPT-4o-mini improved from 72.85% to 73.35%, outperforming GLM-4.7-Flash which only reached 63.85%. A usability test using Nielsen's Heuristics conducted by three lecturer evaluators obtained an average score of 4.64 out of 5.00. These results confirm that the developed RAG system is feasible and effective as an administrative aid in the study program accreditation process..*
Keywords: *study program accreditation, Retrieval-Augmented Generation (RAG), fine-tuning, sme-embeddings, GPT-4o-mini, GLM-4.7-Flash*

Abstrak. Proses akreditasi program studi melibatkan dokumen regulasi yang kompleks dan berjumlah besar, sehingga menyulitkan pengelola dalam mencari informasi secara cepat dan tepat. Penelitian ini bertujuan merancang dan membangun sistem Question-Answering (QA) berbasis Retrieval-Augmented Generation (RAG) untuk manajemen dokumentasi akreditasi Program Studi Ilmu Informatika di Universitas Katolik Darma Cendika (UKDC). Sistem mengintegrasikan 84 dokumen regulasi yang diekstraksi menggunakan PyMuPDF menjadi 4.613 potongan teks (chunk) dan diindeks ke dalam basis data vektor ChromaDB. Arsitektur RAG memanfaatkan model embedding BGE-M3 yang di-fine-tuning dengan metode sme-embeddings berbasis contrastive learning (MultipleNegativesRankingLoss) untuk mengoptimalkan pencarian dokumen spesifik domain, serta model bahasa besar (LLM) GPT-4o-mini sebagai generator jawaban akhir, dengan GLM-4.7-Flash sebagai model pembanding. Evaluasi modul retrieval menunjukkan peningkatan signifikan setelah fine-tuning, dengan Recall@1 naik dari 0,8100 menjadi 0,9300, NDCG@1 dari 0,8100 menjadi 0,9300, dan MRR@10 dari 0,8757 menjadi 0,9603. Evaluasi kualitas jawaban menggunakan BERTScore menunjukkan F1-Score GPT-4o-mini meningkat dari 72,85% menjadi 73,35%, unggul dibandingkan GLM-4.7-Flash yang hanya mencapai 63,85%. Pengujian usability dengan Nielsen's Heuristics oleh tiga dosen penguji memperoleh rata-rata skor 4,64 dari skala 5,00. Hasil ini mengonfirmasi bahwa sistem RAG yang dikembangkan layak dan efektif digunakan sebagai alat bantu administratif dalam proses akreditasi program studi..
Kata kunci: akreditasi program studi, Retrieval-Augmented Generation (RAG), fine-tuning, sme-embeddings, GPT-4o-mini, GLM-4.7-Flash.

1. LATAR BELAKANG

Akreditasi merupakan instrumen krusial dalam sistem penjaminan mutu pendidikan tinggi karena berfungsi mengevaluasi program studi berdasarkan standar dan panduan yang telah ditetapkan (Putri et al., 2026). Proses ini tidak hanya memastikan akuntabilitas institusi di mata publik, tetapi juga menjadi motor penggerak bagi perbaikan kualitas

Naskah Masuk: 06 Agustus 2026; Revisi: 07 Agustus 2026; Diterima: 08 Agustus 2026; ; Terbit: 20 Agustus 2026

pendidikan secara berkelanjutan (Noviyanti & Salabi, 2026). Dalam praktiknya, proses akreditasi sering kali melibatkan dokumen regulasi yang sangat kompleks, teknis, dan berjumlah banyak, sehingga sulit diinterpretasikan dengan cepat oleh pengelola akreditasi program studi. Kondisi ini menimbulkan beban administratif yang tinggi akibat banyaknya pertanyaan berulang yang diajukan kepada staf akreditasi (Putri et al., 2026). Permasalahan serupa juga dialami oleh Universitas Katolik Darma Cendika (UKDC) dalam melakukan manajemen dokumentasi akreditasi program studi, sehingga muncul urgensi akan adanya sebuah aplikasi yang dapat memfasilitasi pencarian aturan atau pedoman secara cepat tanpa harus menelusuri dokumen-dokumen regulasi secara manual.

Untuk mengatasi tantangan tersebut, implementasi sistem Question-Answering (QA) berbasis kecerdasan buatan mulai dikembangkan untuk membantu menavigasi informasi akreditasi secara otomatis (Putri et al., 2026). Penelitian-penelitian terbaru menunjukkan efektivitas arsitektur Retrieval-Augmented Generation (RAG) dalam menyediakan layanan informasi akreditasi yang andal, dengan cara menarik konten otoritatif dari dokumen regulasi kemudian menghasilkan penjelasan yang sesuai konteks (Putri et al., 2026). Arsitektur RAG dinilai lebih unggul dibandingkan model bahasa besar (Large Language Model/LLM) konvensional karena mampu memitigasi risiko halusinasi informasi dengan membatasi jawaban hanya pada dokumen referensi yang valid (Fajri & Mandala, 2025).

Meskipun sistem RAG telah menunjukkan performa yang menjanjikan, efektivitas jawaban yang dihasilkan sangat bergantung pada akurasi modul pencarian data (retrieval) dalam menemukan dokumen yang paling relevan (Putri et al., 2026). Pada domain yang sangat spesifik dan teknis seperti akreditasi, model embedding standar sering kali kurang memiliki pengetahuan domain yang cukup mendalam untuk membedakan istilah-istilah regulasi yang kompleks, seperti IAPS, LED, dan LKPS (Fajri & Mandala, 2025). Penggunaan model embedding generik tanpa optimasi berisiko menurunkan akurasi pencarian ketika dihadapkan pada kueri yang bersifat implisit atau menggunakan istilah yang sangat teknis (Fajri & Mandala, 2025). Oleh karena itu diperlukan optimasi lebih lanjut melalui teknik fine-tuning pada model embedding untuk meningkatkan relevansi pencarian dokumen spesifik domain (Fajri & Mandala, 2025). Penerapan fine-tuning pada model embedding BGE-M3 dengan metode sme-embeddings sebagaimana dilaporkan pada repositori publik terbukti mampu meningkatkan akurasi pencarian dari 50% menjadi 65% (Jc7k, 2026), sementara pemilihan model LLM GPT-4o-mini didasarkan pada performanya yang lebih unggul dibandingkan model lain pada tugas pemahaman teks, dengan capaian F1-Score sebesar 0,9894 (Pilicita & Barra, 2025).

Berdasarkan uraian tersebut, penelitian ini bertujuan mengembangkan sistem QA akreditasi kampus yang menggabungkan kerangka kerja RAG dengan optimasi fine-tuning model embedding BGE-M3 menggunakan metode sme-embeddings, serta membandingkan performa jawaban antara model LLM GPT-4o-mini dan GLM-4.7-Flash. Rumusan masalah dalam penelitian ini terdiri atas dua hal: (1) bagaimana merancang dan membangun sistem QA berbasis RAG yang mampu mengekstraksi dan menyajikan informasi secara akurat dari dokumen akreditasi program studi, dan (2) seberapa efektif jawaban yang dihasilkan dari implementasi fine-tuning model embedding BGE-M3 dengan metode sme-embeddings menggunakan model LLM GPT-4o-mini dan GLM-4.7-Flash. Penelitian ini dibatasi pada pemrosesan dokumen resmi akreditasi Program Studi Ilmu Informatika UKDC, dengan evaluasi menggunakan metrik Recall, NDCG, MRR, BERTScore, dan Nielsen's Heuristics, serta memanfaatkan

ChromaDB, PyMuPDF, FastAPI, OpenAI API, Zhipu AI API, Google Colab, dan Gradio sebagai perangkat pendukung. Penelitian ini diharapkan dapat digunakan untuk membantu pekerjaan akreditasi yang lebih presisi dalam mendukung penjaminan mutu di perguruan tinggi, sekaligus menambah literatur mengenai penerapan fine-tuning model embedding untuk dokumen berbahasa Indonesia formal.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif, di mana data yang diperoleh dari tim penjamin mutu akreditasi Program Studi Ilmu Informatika Universitas Katolik Darma Cendika (UKDC) digunakan untuk merancang dan mengevaluasi aplikasi RAG manajemen dokumentasi akreditasi. Penelitian dilaksanakan melalui tahapan identifikasi masalah, studi literatur, pengumpulan dan pra-pemrosesan data, fine-tuning model embedding, pengembangan sistem RAG dan antarmuka aplikasi, hingga evaluasi dan analisis hasil.

Pengumpulan dan Pra-Pemrosesan Data

Data penelitian dikumpulkan dari dokumen-dokumen resmi akreditasi Program Studi Ilmu Informatika UKDC, dengan total 84 dokumen PDF yang terdiri atas Rencana Pembelajaran Semester (RPS), Laporan Evaluasi Diri (LED), Laporan Kinerja Program Studi (LKPS), Instrumen Akreditasi Program Studi (IAPS), modul praktikum, dokumen roadmap penelitian, kurikulum, matriks penilaian, laporan masa studi, serta dokumen pendukung akreditasi lainnya. Seluruh dokumen diekstraksi menggunakan pustaka PyMuPDF, kemudian dipotong (chunking) menggunakan pendekatan berbasis struktur dokumen dengan target panjang 600 karakter per chunk yang dikonfigurasi melalui variabel CHUNK_TARGET_SIZE. Proses ini menghasilkan total 4.613 potongan teks yang selanjutnya diubah menjadi representasi vektor menggunakan model embedding BGE-M3 dan disimpan ke dalam basis data vektor ChromaDB.

Tingkat kesamaan semantik antara vektor kueri pengguna dan vektor chunk diukur menggunakan cosine distance, dengan nilai berkisar antara 0 (sangat mirip) hingga 2 (sangat berbeda). Sebuah chunk hanya dianggap relevan apabila nilai cosine distance terendah dari hasil pencarian tidak melebihi ambang batas (threshold) yang ditetapkan sebesar 1,55. Selain basis data referensi, disusun pula dataset pelatihan fine-tuning berupa pasangan pertanyaan-jawaban (question-answering pairs) yang dihasilkan menggunakan GPT-4o-mini berdasarkan konten yang tersimpan di ChromaDB. Dataset ini terdiri atas 900 triplet untuk pelatihan (training) dan 100 triplet untuk validasi (validation), dengan masing-masing triplet berisi satu query, satu konteks positif (dokumen relevan), dan empat konteks negatif (dokumen tidak relevan yang berasal dari file berbeda). Contoh struktur triplet yang digunakan ditunjukkan pada Tabel 1.

Tabel 1. Contoh Dataset Triplet untuk Fine-Tuning

Komponen	Contoh
Query	*Apa saja mata kuliah wajib yang diambil pada semester 5 Prodi Ilmu Informatika?*
Positive	*[Kurikulum.pdf] ... Semester 5: Rekayasa Perangkat Lunak (3 SKS), Pemrograman Web (3 SKS), Basis Data (3 SKS) ...*
Negative 1	*[RPS_Pemrograman_Mobile.pdf] ... Capaian pembelajaran mata kuliah Pemrograman Mobile ...*

Komponen	Contoh
Negative 2	*[LED_2024.pdf] ... Visi dan Misi Program Studi Ilmu Informatika ...*
Negative 3	*[IAPS_2024.pdf] ... Standar Mutu Proses Pembelajaran ...*
Negative 4	*[Roadmap_Penelitian.pdf] ... Rencana Penelitian Tahun 2025-2029 ...*

Pemilihan konteks negatif berasal dari file yang berbeda dengan file asal konteks positif, dengan tujuan memastikan model belajar membedakan relevansi antar dokumen berdasarkan konten dan bukan sekadar kemiripan permukaan teks.

Fine-Tuning Model Embedding BGE-M3

Proses fine-tuning terhadap model embedding BGE-M3 dilakukan menggunakan metode sme-embeddings dengan fungsi loss MultipleNegativesRankingLoss (MNRL). Pelatihan dijalankan pada lingkungan Google Colab menggunakan GPU T4 dengan 16 GB VRAM, yang merupakan konfigurasi minimum yang dibutuhkan untuk batch size 2 pada sequence length 512. Konfigurasi pelatihan yang digunakan ditunjukkan pada Tabel 2.

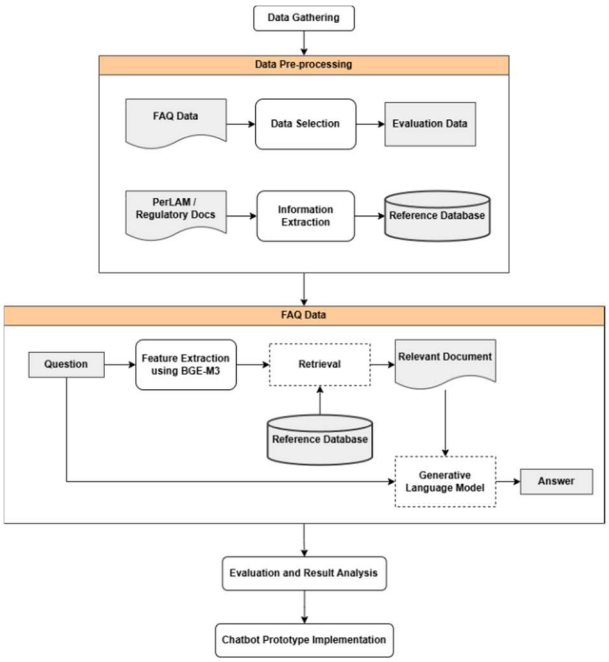
Tabel 2. Konfigurasi Fine-Tuning Model Embedding BGE-M3

Parameter	Nilai
GPU	NVIDIA T4 (16 GB VRAM) via Google Colab
Epoch	3
Batch Size	2
Learning Rate	2e-5
Max Sequence Length	512

Model hasil fine-tuning disimpan ke Google Drive untuk selanjutnya diunduh dan di-deploy ke server aplikasi. Evaluasi performa model dilakukan dengan membandingkan kinerja model dasar (baseline) dan model fine-tuned menggunakan metrik Recall@K, NDCG@K, dan MRR@K terhadap 100 query validasi melalui pencarian pada ChromaDB.

Perancangan Arsitektur Sistem RAG

Sistem RAG untuk manajemen dokumentasi akreditasi dirancang melalui tiga tahapan utama, yaitu (1) pengumpulan dan pra-pemrosesan data, (2) perancangan dan implementasi model QA, serta (3) evaluasi dan pengembangan prototipe, sebagaimana diilustrasikan pada Gambar 1.



Gambar 1. Arsitektur Alur Kerja Sistem RAG Akreditasi

Pada tahap perancangan dan implementasi, alur sistem dimulai ketika pengguna memasukkan pertanyaan terkait akreditasi ke dalam aplikasi. Teks pertanyaan dikonversi menjadi representasi vektor numerik (embedding) menggunakan model BGE-M3 yang telah di-fine-tuning, kemudian dicocokkan dengan seluruh vektor dokumen di dalam basis data referensi menggunakan pencarian kemiripan (similarity search) pada ChromaDB. Hasil pencarian berupa dokumen-dokumen regulasi yang paling relevan dengan pertanyaan pengguna kemudian digabungkan dengan pertanyaan asli dan dikirimkan ke model bahasa generatif, yaitu GPT-4o-mini (dengan GLM-4.7-Flash sebagai pembanding), untuk disintesis menjadi jawaban akhir yang akurat, mudah dipahami, dan berlandaskan pada dokumen resmi. Orkestrasi seluruh alur pemrosesan bahasa alami pada sisi backend diimplementasikan menggunakan framework FastAPI, yang berfungsi sebagai jembatan antara kueri pengguna dan jalur pemrosesan RAG. Antarmuka pengguna (frontend) dikembangkan menggunakan pustaka Gradio untuk memfasilitasi interaksi yang mudah diakses oleh pengurus akreditasi, termasuk fitur riwayat percakapan, manajemen dokumen, serta unggah dan hapus berkas PDF.

Selain alur pemrosesan kueri, rancangan antarmuka pengguna (user interface) turut disusun dalam bentuk mockup high-fidelity sebelum implementasi, mencakup tahapan memasukkan pertanyaan, konversi pertanyaan menjadi representasi vektor menggunakan BGE-M3, hingga penyajian jawaban akhir beserta dokumen sumber yang relevan. Rancangan ini menjadi acuan bagi pengembangan antarmuka aktual menggunakan Gradio, dengan mempertimbangkan aspek kebergunaan (usability) yang akan diuji pada tahap evaluasi.

Desain Evaluasi Sistem

Evaluasi sistem dilakukan dalam tiga aspek. Pertama, evaluasi modul retrieval menggunakan metrik Recall@K, NDCG@K, dan MRR@K untuk membandingkan performa model embedding baseline dan fine-tuned dalam menemukan dokumen relevan dari ChromaDB berdasarkan 100 query validasi. Kedua, evaluasi kualitas jawaban sistem

RAG menggunakan metrik BERTScore, dengan membandingkan jawaban kandidat yang dihasilkan pipeline RAG (retrieval TOP_K=6 chunk, dilanjutkan generation menggunakan GPT-4o-mini dan GLM-4.7-Flash) terhadap jawaban referensi pada 100 pasang pertanyaan validasi yang tersimpan pada berkas validation_triplets.json. Prompt sistem dirancang agar model menjawab hanya berdasarkan konteks yang diberikan tanpa menggunakan pengetahuan di luar konteks tersebut. Ketiga, evaluasi usability antarmuka prototipe menggunakan Nielsen's Heuristics yang dilakukan bersama tiga dosen penguji dari UKDC, dengan sepuluh pernyataan yang dinilai menggunakan skala Likert 1 sampai 5 setelah penguji melakukan uji coba langsung berupa bertanya, mengunduh, dan mengunggah dokumen pada sistem.

3. HASIL DAN PEMBAHASAN

Hasil Fine-Tuning Model Embedding BGE-M3

Hasil validasi model embedding BGE-M3 sebelum dan sesudah fine-tuning terhadap 100 triplet validasi ditunjukkan pada Tabel 3.

Tabel 3. Hasil Validasi Fine-Tuning Model Embedding BGE-M3

K	Metrik	Baseline	Fine-Tuned	Peningkatan
1	Recall@1	0,8100	0,9300	+14,8%
1	NDCG@1	0,8100	0,9300	+14,8%
3	Recall@3	0,9300	0,9900	+6,5%
3	NDCG@3	0,8805	0,9665	+9,8%
5	Recall@5	0,9700	1,0000	+3,1%
5	NDCG@5	0,8973	0,9704	+8,2%
10	Recall@10	0,9900	1,0000	+1,0%
10	NDCG@10	0,9039	0,9704	+7,4%
10	MRR@10	0,8757	0,9603	+9,7%

Tabel 3 menunjukkan bahwa model fine-tuned berhasil meningkatkan seluruh metrik retrieval pada setiap nilai K yang diuji. Recall@1 meningkat dari 0,8100 menjadi 0,9300, atau setara peningkatan sebesar 14,8%, yang menandakan bahwa dokumen paling relevan lebih sering muncul pada posisi teratas hasil pencarian. Recall@3 dan Recall@5 masing-masing meningkat dari 0,9300 menjadi 0,9900 dan dari 0,9700 menjadi 1,0000, menunjukkan bahwa hampir seluruh dokumen relevan berhasil ditemukan dalam lima kandidat teratas. NDCG mengalami peningkatan konsisten pada seluruh nilai K, dengan NDCG@1 naik dari 0,8100 menjadi 0,9300, NDCG@3 dari 0,8805 menjadi 0,9665, dan NDCG@5 dari 0,8973 menjadi 0,9704, yang mengindikasikan bahwa kualitas peringkat dokumen relevan turut membaik, bukan sekadar cakupan pencariannya. Adapun MRR@10 meningkat dari 0,8757 menjadi 0,9603, atau sebesar 9,7%, yang menegaskan bahwa dokumen relevan pertama secara rata-rata ditemukan pada posisi yang lebih dekat dengan peringkat teratas.

Peningkatan ini mengonfirmasi bahwa fine-tuning model embedding BGE-M3 dengan metode sme-embeddings sangat krusial dalam mengatasi kompleksitas bahasa pada regulasi akreditasi. Model yang telah di-fine-tuning tidak lagi sekadar memproses teks secara leksikal umum, melainkan menjadi lebih optimal terhadap terminologi

akademik yang spesifik, seperti IAPS, LED, dan LKPS. Dengan menempatkan dokumen relevan pada peringkat teratas, sistem RAG secara efektif memitigasi noise atau informasi tidak relevan sebelum diteruskan ke model generatif, sehingga berdampak langsung pada kualitas jawaban akhir yang dihasilkan.

Evaluasi Kualitas Jawaban dengan BERTScore

Evaluasi BERTScore dilakukan terhadap 100 pasang pertanyaan validasi seputar akreditasi Program Studi Ilmu Informatika, dengan membandingkan jawaban kandidat yang dihasilkan pipeline RAG (menggunakan embedding baseline maupun fine-tuned) terhadap jawaban referensi. Hasil evaluasi untuk kedua model LLM ditunjukkan pada Tabel 4.

Tabel 4. Hasil Evaluasi BERTScore pada Model GPT-4o-mini dan GLM-4.7-Flash

Metrik	Baseline GPT-4o-mini	Fine-Tuned GPT-4o-mini	Baseline GLM-4.7-Flash	Fine-Tuned GLM-4.7-Flash
Precision	77,88%	78,33%	68,81%	68,85%
Recall	68,63%	69,15%	59,01%	59,84%
F1-Score	72,85%	73,35%	63,24%	63,85%

Tabel 4 menunjukkan bahwa penggunaan embedding fine-tuned memberikan peningkatan pada seluruh metrik BERTScore untuk kedua model LLM. Pada GPT-4o-mini, precision meningkat dari 77,88% menjadi 78,33% (naik 0,45 poin persentase), recall meningkat dari 68,63% menjadi 69,15% (naik 0,52 poin persentase), dan F1-Score meningkat dari 72,85% menjadi 73,35% (naik 0,50 poin persentase). Pada GLM-4.7-Flash, precision meningkat dari 68,81% menjadi 68,85%, recall dari 59,01% menjadi 59,84%, dan F1-Score dari 63,24% menjadi 63,85%. Peningkatan pada kedua model LLM ini mengonfirmasi bahwa optimalisasi komponen retriever melalui fine-tuning BGE-M3 secara langsung mendukung kualitas teks jawaban yang dihasilkan, karena chunk yang lebih relevan memberikan konteks yang lebih tepat bagi model generatif untuk memparafrase dan menyusun jawaban.

Selain itu, Tabel 4 memperlihatkan bahwa GPT-4o-mini secara konsisten memperoleh hasil evaluasi yang lebih tinggi dibandingkan GLM-4.7-Flash pada seluruh metrik dan kedua konfigurasi embedding. Dengan embedding baseline, F1-Score GPT-4o-mini mencapai 72,85% sedangkan GLM-4.7-Flash hanya 63,24%, atau selisih 9,61 poin persentase. Dengan embedding fine-tuned, F1-Score GPT-4o-mini mencapai 73,35% sedangkan GLM-4.7-Flash sebesar 63,85%, dengan selisih yang relatif tetap yaitu 9,50 poin persentase. Hasil ini menunjukkan bahwa GPT-4o-mini menghasilkan jawaban yang secara semantik lebih sesuai dengan jawaban referensi dibandingkan GLM-4.7-Flash pada konteks dokumen akreditasi berbahasa Indonesia, sekalipun kedua model menerima konteks retrieval yang identik.

Evaluasi Usability dengan Nielsen's Heuristics

Pengujian usability dilakukan bersama tiga dosen penguji Universitas Katolik Darma Cendika, yang diberikan waktu untuk melakukan uji coba terhadap sistem RAG berupa bertanya, mengunduh, dan mengunggah dokumen, kemudian mengisi forum penilaian berisi sepuluh pernyataan pada skala Likert 1 hingga 5. Hasil rata-rata skor pada setiap pernyataan ditunjukkan pada Tabel 5.

Tabel 5. Hasil Evaluasi Usability dengan Nielsen's Heuristics

No	Pernyataan Heuristik	Rata-Rata Skor
1	Saat mengajukan pertanyaan, pengguna dengan cepat mengetahui bahwa jawaban sedang diproses oleh chatbot (visibility of system status).	5,0
2	Chatbot dapat memahami bahasa Indonesia dari pertanyaan pengguna dan jawaban yang diberikan.	4,6
3	Pengguna dapat mengajukan pertanyaan baru, menghentikan proses jawaban, dan menghapus riwayat obrolan dengan mudah (user control and freedom).	4,6
4	Struktur penempatan kalimat dari jawaban yang diberikan oleh chatbot mudah dibaca.	4,6
5	Tautan untuk mengakses sumber informasi memudahkan pengguna memeriksa keaslian informasi.	5,0
6	Tanpa instruksi tambahan, pengguna memahami cara mengunduh dokumen referensi dari jawaban yang diberikan.	4,3
7	Chatbot dapat memahami dua gaya bahasa pertanyaan berbeda yang memiliki konteks yang sama.	4,3
8	Antarmuka chatbot terlihat bersih dan tidak berantakan (aesthetic and minimalist design).	5,0
9	Jika pertanyaan berada di luar topik akreditasi dan akademik, chatbot dapat memberikan pesan kesalahan yang informatif.	4,0
10	Opsi pertanyaan yang ditampilkan pada halaman awal chatbot memudahkan pengguna mencari informasi dengan cepat.	5,0

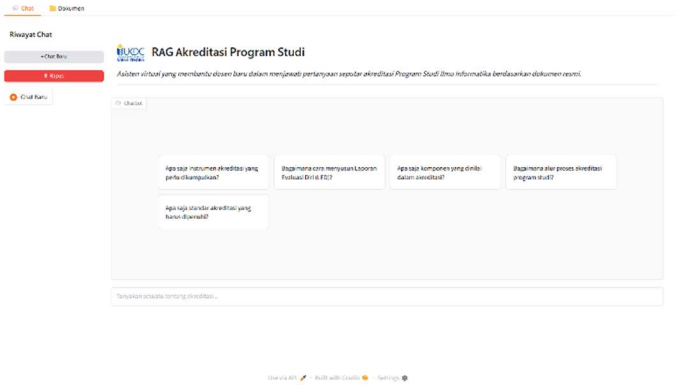
Hasil evaluasi pada Tabel 5 menunjukkan bahwa sistem memperoleh rata-rata skor keseluruhan sebesar 4,64 dari skala 5,00. Empat aspek memperoleh skor tertinggi (5,0), yaitu indikator status pemrosesan, ketersediaan tautan sumber informasi, kebersihan antarmuka, serta opsi pertanyaan pada halaman awal. Adapun aspek dengan skor terendah adalah pemahaman terhadap gaya bahasa pertanyaan yang berbeda dan kemudahan mengunduh dokumen (4,3), serta penanganan pertanyaan di luar topik akreditasi (4,0). Meskipun demikian, seluruh aspek tetap berada di atas ambang 4,0, sehingga secara umum sistem RAG yang dikembangkan telah memenuhi standar kelayakan dan kenyamanan pengguna.

Para penguji turut memberikan sejumlah masukan kualitatif, di antaranya agar sistem RAG difokuskan sebagai alat bantu penyiapan kebutuhan informasi dokumen akreditasi dan akademik, bukan sebagai alat penghitung nilai akreditasi. Penguji juga menyarankan peningkatan rekognisi terhadap pertanyaan pengguna yang kurang detail, memiliki kekurangan kata, atau mengandung kesalahan penulisan, agar sistem tetap dapat memberikan jawaban yang relevan. Masukan lain terkait fitur unggah dokumen adalah agar dokumen yang bersifat umum cukup diunggah oleh satu program studi apabila

sistem diperluas untuk seluruh program studi di UKDC, sementara dokumen yang spesifik pada satu program studi tetap dipisahkan penyimpanannya agar tidak tersebar ke program studi lain.

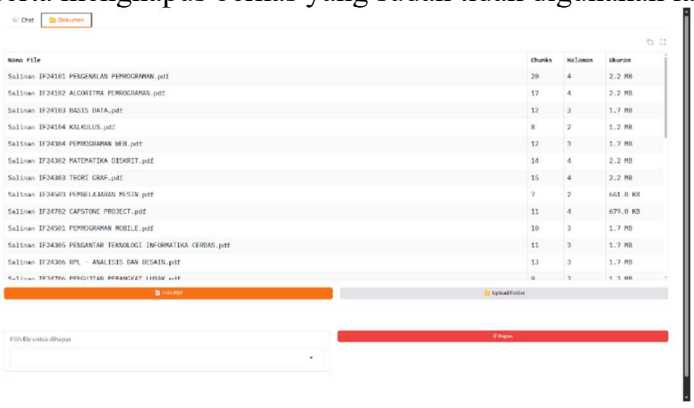
Implementasi Antarmuka Aplikasi

Prototipe aplikasi RAG akreditasi dikembangkan menggunakan pustaka Gradio dengan dua halaman utama, yaitu halaman percakapan (chat) dan halaman manajemen dokumentasi. Halaman beranda, sebagaimana ditunjukkan pada Gambar 2, menampilkan logo universitas, deskripsi kegunaan chatbot untuk membantu dosen dan pengurus akreditasi, serta lima contoh pertanyaan yang dapat digunakan pengguna sebagai titik awal interaksi. Sebelah kiri halaman menyediakan sidebar untuk melihat riwayat percakapan, memulai percakapan baru, menghapus riwayat, dan berpindah ke halaman manajemen dokumentasi.



Gambar 2. Halaman Beranda Aplikasi RAG Akreditasi

Halaman manajemen dokumentasi, sebagaimana ditunjukkan pada Gambar 3, menampilkan tabel daftar berkas PDF yang telah diindeks beserta informasi nama berkas, jumlah potongan teks (chunk), jumlah halaman, dan ukuran berkas. Pada halaman ini, pengguna dapat mengunggah satu berkas, beberapa berkas PDF sekaligus, atau seluruh folder dokumen, serta menghapus berkas yang sudah tidak digunakan lagi.



Gambar 3. Halaman Manajemen Dokumentasi

Pembahasan Umum

Secara keseluruhan, hasil pengujian yang telah diuraikan menunjukkan bahwa implementasi arsitektur RAG pada dokumen akreditasi memiliki tingkat kinerja yang optimal, baik dari sisi teknis maupun kebergunaan antarmuka. Penerapan fine-tuning pada model embedding BGE-M3 dengan metode sme-embeddings terbukti krusial dalam

mengatasi kompleksitas bahasa pada regulasi akreditasi, sebagaimana ditunjukkan oleh peningkatan Recall@1 menjadi 0,9300 dan MRR@10 menjadi 0,9603. Peningkatan performa retrieval ini kemudian berdampak positif pada kualitas jawaban akhir, yang dikonfirmasi oleh kenaikan F1-Score BERTScore dari 72,85% menjadi 73,35% pada model GPT-4o-mini, sehingga menegaskan bahwa keberhasilan modul retriever dalam menyediakan chunk yang relevan berkontribusi langsung terhadap optimalnya hasil jawaban LLM.

Perbandingan antara kedua model LLM menunjukkan bahwa GPT-4o-mini menghasilkan jawaban yang secara semantik lebih sesuai dengan jawaban referensi dibandingkan GLM-4.7-Flash, dengan selisih F1-Score sekitar 9,5 hingga 9,6 poin persentase pada kedua konfigurasi embedding. Meskipun GLM-4.7-Flash menawarkan keunggulan berupa penggunaan tanpa biaya token, konsistensi keunggulan GPT-4o-mini pada seluruh metrik precision, recall, dan F1-Score menunjukkan bahwa pemilihan model generator tetap perlu mempertimbangkan trade-off antara biaya operasional dan akurasi semantik jawaban, khususnya untuk domain dengan konsekuensi akurasi yang tinggi seperti dokumentasi akreditasi. Adapun hasil pengujian usability dengan skor rata-rata 4,64 dari skala 5,00 mengonfirmasi bahwa antarmuka sistem RAG telah memenuhi standar kelayakan dan kenyamanan pengguna, sehingga secara keseluruhan sistem yang dikembangkan layak digunakan sebagai instrumen bantu administratif dalam mendukung proses akreditasi program studi.

Implikasi Praktis dan Keterbatasan Penelitian

Dari sisi implikasi praktis, sistem RAG yang dikembangkan berpotensi mengurangi beban administratif tim penjamin mutu dalam menjawab pertanyaan berulang seputar regulasi akreditasi, sekaligus mempercepat proses onboarding dosen atau staf baru yang belum terbiasa dengan struktur dokumen akreditasi. Ketersediaan fitur unggah dan pengelolaan dokumen pada halaman manajemen dokumentasi juga memungkinkan basis pengetahuan sistem diperbarui secara mandiri tanpa memerlukan intervensi teknis, sehingga sistem tetap relevan seiring perubahan regulasi akreditasi dari waktu ke waktu.

Meskipun demikian, penelitian ini memiliki sejumlah keterbatasan. Pertama, proses fine-tuning dilakukan pada GPU T4 dengan kapasitas 16 GB VRAM yang membatasi ukuran batch size dan panjang maksimum sequence yang dapat digunakan, sehingga berpotensi memengaruhi stabilitas dan cakupan konteks yang dipelajari model selama pelatihan. Kedua, evaluasi kualitas jawaban dan usability dilakukan dengan jumlah data validasi dan jumlah evaluator yang relatif terbatas, yaitu 100 pasang pertanyaan validasi dan tiga dosen penguji, sehingga generalisasi hasil pada populasi pengguna yang lebih luas atau pada dokumen program studi lain masih memerlukan pengujian lebih lanjut. Ketiga, sebagaimana disampaikan oleh penguji, sistem saat ini masih memiliki keterbatasan dalam mengenali pertanyaan yang kurang detail atau mengandung kesalahan penulisan, serta belum sepenuhnya optimal dalam menangani pertanyaan di luar topik akreditasi dan akademik.

4. KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan keseluruhan proses perancangan, pengujian, dan analisis sistem Retrieval-Augmented Generation (RAG) untuk manajemen dokumentasi akreditasi Program Studi Ilmu Informatika di Universitas Katolik Darma Cendika, dapat ditarik kesimpulan sebagai berikut.

1. Sistem Question-Answering (QA) berbasis RAG telah berhasil dirancang dan dibangun melalui integrasi 84 dokumen regulasi beserta dokumen operasional akreditasi, meliputi LED, LKPS, IAPS, dan RPS. Proses ekstraksi dan chunking dilaksanakan menggunakan PyMuPDF dengan target panjang 600 karakter per chunk, yang selanjutnya diindeks ke dalam basis data vektor ChromaDB. Orkestrasi pemrosesan bahasa alami diimplementasikan menggunakan FastAPI pada sisi backend, sedangkan antarmuka pengguna dikembangkan menggunakan Gradio. Berdasarkan pengujian usability dengan Nielsen's Heuristics, sistem meraih rata-rata skor kepuasan sebesar 4,64 dari skala 5,00, yang membuktikan kelayakan dan efektivitasnya sebagai instrumen bantu administratif.
2. Respons yang dihasilkan dari integrasi model embedding BGE-M3 yang di-fine-tuning dengan metode sme-embeddings dan model LLM GPT-4o-mini terbukti sangat efektif dan berakurasi tinggi secara semantik. Fine-tuning berhasil meningkatkan performa retrieval secara signifikan, dengan Recall@1 sebesar 0,9300, NDCG@1 sebesar 0,9300, dan MRR@10 sebesar 0,9603. Evaluasi kualitas jawaban menggunakan BERTScore menunjukkan bahwa fine-tuning embedding BGE-M3 efektif meningkatkan performa jawaban GPT-4o-mini, dengan Precision 78,33%, Recall 69,15%, dan F1-Score 73,35%, sedangkan GLM-4.7-Flash memperoleh Precision 68,85%, Recall 59,84%, dan F1-Score 63,85%. Dengan demikian, GPT-4o-mini menghasilkan jawaban yang lebih sesuai dengan jawaban referensi secara semantik dibandingkan GLM-4.7-Flash.

Saran

Beberapa saran dapat diberikan untuk pengembangan aplikasi RAG akreditasi selanjutnya. Pertama, penelitian lanjutan disarankan menggunakan GPU dengan kapasitas VRAM yang lebih besar dibandingkan T4 GPU pada Google Colab, sehingga memungkinkan penggunaan batch size yang lebih besar untuk proses pelatihan yang lebih stabil dan optimal, sekaligus memungkinkan penggunaan max sequence length yang lebih panjang agar model dapat memproses dokumen dengan konteks yang lebih luas tanpa harus memotong informasi penting. Kedua, pengembangan selanjutnya dapat menambahkan fitur kustomisasi contoh pertanyaan pada halaman beranda melalui basis data SQL, sehingga pengurus akreditasi dapat menyesuaikan contoh pertanyaan sesuai kebutuhan aktual tanpa mengubah kode sumber aplikasi.

DAFTAR REFERENSI

AI, Z. (2026). GLM-4.7-Flash. Hugging Face. https://huggingface.co/zai-org/GLM-4.7-Flash

Bovas, A. P., Joy, M., Santo, N. C., Borde, S. R., & Deshpande, V. V. (2025). Navi: RAG-Powered LLM Chatbot for Academic Institutions. International Journal for Research in Applied Science and Engineering Technology, 13(11), 2339–2346. https://doi.org/10.22214/ijraset.2025.75655

Cui, J., Ning, M., Li, Z., Chen, B., Yan, Y., Li, H., Ling, B., Tian, Y., & Yuan, L. (2024). Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model. 1–11. http://arxiv.org/abs/2306.16092

Elkiran, H., & Rasheed, J. (2025). EvaRAG: Evaluating Advanced RAG Techniques With Indexing and Distance Metrics. IEEE Access, 13(December), 215724–215747. https://doi.org/10.1109/ACCESS.2025.3646665

Fajri, A., & Mandala, R. (2025). Optimizing Retrieval-Augmented Generation for Domain-Specific Knowledge Systems through Fine-Tuning and Prompt Engineering. 5(2), 344–350.

Forootani, A., Aliabadi, D. E., & Thraen, D. (2025). Bio-Eng-LMM AI Assist chatbot: A Comprehensive Tool for Research and Education. http://arxiv.org/abs/2409.07110

Husain, M. L., Wibisono, Y., & Anisyah, A. (2025). Development of an Academic Services Chatbot Based on Retrieval-Augmented Generation (RAG). Brilliance: Research of Artificial Intelligence, 5(2), 727–735. https://jurnal.itscience.org/index.php/brilliance/article/view/6719

Jc7k. (2026). sme-embeddings. https://github.com/jc7k/sme-embeddings

Mansour, Y. E. I., & Ibrahim, A. (2025). AI-Driven Framework for Academic Program Evaluation and Accreditation Standards. The Barcelona Conference on Education 2025: Official Conference Proceedings, 393–403. https://doi.org/10.22492/issn.2435-9467.2025.32

Moreno-Cediel, A., Garcia-Lopez, E., Garcia-Cabot, A., & De-Fitero-Dominguez, D. (2025). Optimising retrieval performance in RAG systems: A new growing window semantic chunking strategy to address weak semantic boundaries. Knowledge-Based Systems, 331(November), 114896. https://doi.org/10.1016/j.knosys.2025.114896

Noviyanti, & Salabi, A. S. (2026). Akreditasi dalam Perubahan Global: Isu Klasik, Dinamika Regulasi dan Teknologi, serta Strategi Adaptif. 9, 2880–2887.

Pilicita, A., & Barra, E. (2025). LLMs in Education: Evaluation GPT and BERT Models in Student Comment Classification. Multimodal Technologies and Interaction, 9(5). https://doi.org/10.3390/mti9050044

Putri, R. A., Nurramdhani, H. I., & Hartati, S. (2026). A Retrieval-Augmented Generation Framework for a Study Program Accreditation Question-Answering System. 16(2), 33375–33381.

Rotaru, O., Orhei, C., & Vasiu, R. (2026). Hybrid Usability Evaluation of an Automotive REM Tool: Human and LLM-Based Heuristic Assessment of IBM Doors Next. Applied Sciences, 16(2), 723. https://doi.org/10.3390/app16020723

Swacha, J., & Gracel, M. (2025). Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. Applied Sciences (Switzerland), 15(8). https://doi.org/10.3390/app15084234

Tamm, Y. M., Damdinov, R., & Vasilev, A. (2021). Quality metrics in recommender systems: Do We calculate metrics consistently? In RecSys 2021 - 15th ACM Conference on Recommender Systems (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3460231.3478848

Tang, G., Yousuf, O., & Jin, Z. (2024). Improving BERTScore for Machine Translation Evaluation Through Contrastive Learning. IEEE Access, 12, 77739–77749. https://doi.org/10.1109/ACCESS.2024.3406993

Yu, R., Wang, T., Liu, R., & Wang, Y. (2026). SF-RAG: Structure-Fidelity Retrieval-Augmented Generation for Academic Question Answering (Vol. 1, Issue 1). arXiv. http://arxiv.org/abs/2602.13647