Vol. 11, No. 2, Oktober 2026, Page. 566-577
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Implementasi Metode Hybrid Retrieval dan Token-Oriented Object Notation pada Chatbot Akademik Berbasis Large Language Model

Bayu Winata[*1], Defry Hamdhana[2], Rizki Suwanda[3], Fadlisyah[4], Rini Meiyanti[5]

[1,2,3,4,5]Teknik Informatika, Fakultas Teknik, Universitas Malikussaleh, Indonesia
Email: [1]bayu.220170103@mhs.unimal.ac.id, [2]defryhamdhana@unimal.ac.id,
[3]rizkisuwanda@unimal.ac.id , [4]fadlisyah@unimal.ac.id , [5]rinimeiyanti@unimal.ac.id

Abstrak

Layanan informasi akademik di perguruan tinggi masih sering bersifat konvensional, mengharuskan mahasiswa datang ke kampus atau menghubungi pihak universitas secara manual. Penelitian ini mengembangkan *chatbot* akademik berbasis *Large Language Model* (LLM) dengan *hybrid retrieval* yang menggabungkan data tidak terstruktur dari dokumen akademik dan data terstruktur dari basis data relasional, serta mengimplementasikan *Token-Oriented Object Notation* (TOON) untuk efisiensi token. Model LLM yang digunakan adalah *meta-llama/llama-4-scout-17b-16e-instruct* melalui Groq, sedangkan model *embedding* yang digunakan adalah *paraphrase-multilingual-MiniLM-L12-v2*. Sistem dibangun menggunakan arsitektur *Retrieval Augmented Generation* (RAG) dengan ChromaDB dan *Text-to-SQL*, diintegrasikan melalui *Intent Router* berbasis *prompt-based classification*. Pengujian menggunakan metrik ROUGE, *Execution Accuracy*, dan *Token Reduction Rate*. Hasil pengujian menunjukkan rata-rata F1-Score sebesar 79,13% pada ROUGE-1 69,01% pada ROUGE-2, dan 78,50% pada ROUGE-L untuk jalur RAG, serta *Execution Accuracy* 100% pada jalur data terstruktur. TOON berhasil menghemat token rata-rata 24,60% dibandingkan JSON. *Chatbot* berbasis LLM dengan *hybrid retrieval* dan TOON mampu memberikan jawaban cepat, tepat, kontekstual, dan efisien dalam penggunaan token.

Kata kunci: *Chatbot, Hybrid Retrieval, Large Language Model*, RAG, TOON

Abstract

Academic information services in higher education are often still conventional, requiring students to visit campus or contact staff manually. This research develops an academic chatbot based on Large Language Model (LLM) by implementing hybrid retrieval that combines unstructured data from academic documents and structured data from relational databases, as well as implementing Token-Oriented Object Notation (TOON) to improve token efficiency. The LLM model used is meta-llama/llama-4-scout-17b-16e-instruct accessed through Groq, while the embedding model used is paraphrase-multilingual-MiniLM-L12-v2. The system was built using Retrieval Augmented Generation (RAG) architecture with ChromaDB and Text-to-SQL, integrated through Intent Router based on prompt-based classification. Testing used ROUGE metrics, Execution Accuracy, and Token Reduction Rate. Results showed average F1-Score of 79,13% on ROUGE-1, 6,.01% on ROUGE-2, and 78,50% on ROUGE-L for RAG path, and Execution Accuracy of 100% on structured data path. TOON saved an average of 24,60% tokens compared to JSON. The LLM-based chatbot with hybrid retrieval and TOON provides fast, accurate, contextual, and token-efficient answers.

Keywords: *Chatbot, Hybrid Retrieval, Large Language Model, RAG, TOON*

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)

1. PENDAHULUAN

Layanan informasi akademik di perguruan tinggi merupakan komponen vital dalam mendukung kelancaran proses pendidikan. Namun, berdasarkan observasi di Program Studi Magister Teknologi Informasi Universitas Malikussaleh, layanan informasi akademik masih didominasi oleh sistem konvensional yang mengharuskan mahasiswa datang langsung ke kampus atau menghubungi pihak universitas secara manual. Informasi akademik yang tersedia masih tersebar di berbagai halaman website statis tanpa fitur tanya jawab interaktif. Akibatnya, mahasiswa masih harus mencari informasi secara manual yang dinilai tidak efisien dan memakan waktu [1]. Kondisi serupa juga ditemukan pada

berbagai perguruan tinggi lain di Indonesia, di mana sistem informasi akademik yang berjalan secara konvensional mengharuskan mahasiswa menghubungi pihak universitas untuk mendapatkan informasi, sehingga proses tersebut berpotensi menghasilkan ketidaktepatan informasi [2][3].

Perkembangan kecerdasan buatan, khususnya *Natural Language Processing* (NLP) dan *Large Language Model* (LLM) membuka peluang pengembangan *chatbot* cerdas yang mampu memahami dan merespons pertanyaan pengguna secara alami [4]. LLM seperti *meta-llama/llama-4-scout-17b-16e-instruct* telah menunjukkan kemampuan luar biasa dalam menghasilkan teks yang menyerupai bahasa manusia. Untuk meningkatkan kualitas jawaban, digunakan *Retrieval-Augmented Generation* (RAG), yaitu teknik yang menggabungkan model bahasa dengan mekanisme pengambilan informasi dari sumber eksternal, sehingga sistem dapat menghasilkan jawaban yang lebih akurat dan kontekstual [5][6] Namun, implementasi LLM di lingkungan akademik menghadapi dua tantangan utama. Pertama, data akademik bersifat heterogen, terdiri dari data tidak terstruktur (dokumen PDF, panduan akademik) dan data terstruktur (basis data relasional). Kedua, biaya penggunaan token pada LLM komersial relatif mahal, sehingga efisiensi penggunaan token menjadi faktor penting dalam pengembangan sistem [7].

Penelitian terdahulu telah mengembangkan *chatbot* berbasis RAG di berbagai ruang lingkup. Pujiono dkk. [8] mengimplementasikan RAG dan *vector database* pada *chatbot* di lembaga layanan publik, dengan hasil terbaik menggunakan GPT-4 yang memperoleh rata-rata *cosine similarity score* sebesar 0,404. Pangestu dkk. [9] mengembangkan *chatbot* konsultasi akademik menggunakan RAG dengan LLM *LLaMA 3.1-8B-Instant*, memperoleh rata-rata F1-Score sebesar 53,36%. Albert & Voutama [10] mengembangkan *chatbot* berbasis PDF menggunakan Local RAG dan Ollama, dengan hasil ROUGE-L sebesar 0,85. Basallamah dkk. [11] mengembangkan *chatbot hybrid* RAG dan Text-to-SQL untuk informasi kesehatan, dengan akurasi Text-to-SQL mencapai 93,33% dan akurasi Router 85%. Kachhwal dkk. [12] mengembangkan sistem pencarian hybrid multi-modal yang memadukan RAG dan Text2SQL untuk informasi semantik. Meskipun penelitian-penelitian tersebut menunjukkan hasil yang menjanjikan, masih terdapat celah penelitian, yaitu belum adanya *chatbot* akademik yang menggabungkan *hybrid retrieval* (RAG + Text-to-SQL) dengan optimasi efisiensi token menggunakan format representasi data yang ringkas.

Penelitian ini mengusulkan pengembangan *chatbot* akademik berbasis LLM dengan metode *hybrid retrieval* dan *Token-Oriented Object Notation* (TOON). *Hybrid retrieval* menggabungkan dua jalur pemrosesan, yaitu jalur RAG untuk data tidak terstruktur (dokumen akademik) dan jalur Text-to-SQL untuk data terstruktur (basis data MySQL). Kedua jalur tersebut diintegrasikan melalui *Intent Router* berbasis *prompt-based classification* yang menentukan sumber data yang paling sesuai dengan pertanyaan pengguna. Selain itu, penelitian ini mengimplementasikan TOON, yaitu format serialisasi data terstruktur yang dirancang khusus untuk meminimalkan penggunaan token pada input LLM dibandingkan dengan JSON [13]. TOON relevan digunakan sebagai format representasi data pada sistem berbasis LLM, khususnya ketika efisiensi token menjadi perhatian utama dalam pertukaran data antara sistem dan model.

Meskipun penelitian-penelitian tersebut menunjukkan hasil yang menjanjikan, terdapat celah penelitian yang belum terjawab. Pertama, chatbot akademik yang ada umumnya hanya menangani satu jenis data (dokumen atau basis data), padahal kebutuhan informasi akademik bersifat heterogen sehingga memerlukan hybrid retrieval yang menggabungkan RAG dan Text-to-SQL dalam satu sistem. Kedua, penelitian hybrid retrieval sebelumnya belum mempertimbangkan efisiensi token, padahal biaya token merupakan kendala utama pada LLM komersial. Ketiga, belum ada penelitian yang mengintegrasikan hybrid retrieval dengan format representasi data ringkas seperti TOON untuk menekan konsumsi token tanpa mengorbankan kualitas jawaban. Kombinasi ketiga aspek inilah yang menjadi kebaruan penelitian ini.

Sistem dibangun menggunakan arsitektur *Retrieval-Augmented Generation* dengan model LLM *meta-llama/llama-4-scout-17b-16e-instruct* yang diakses melalui platform Groq. Model embedding yang digunakan adalah *paraphrase-multilingual-MiniLM-L12-v2* untuk mengubah teks menjadi representasi vektor. Data tidak terstruktur berupa tiga dokumen akademik (Buku Panduan Akademik, Buku Panduan Penulisan Proposal dan Tesis, serta Profil Program Studi) diproses melalui tahapan *chunking* dan *embedding*, kemudian disimpan ke dalam *vector database* ChromaDB. Data terstruktur berupa informasi dosen, kalender akademik, mata kuliah, dan struktur organisasi program studi disimpan dalam basis data MySQL. *Intent Router* berbasis *prompt-based classification* menentukan

jalur pemrosesan, yaitu jalur RAG untuk pertanyaan yang membutuhkan informasi dari dokumen dan jalur SQL+TOON untuk pertanyaan yang membutuhkan data terstruktur. Pada jalur data terstruktur, hasil query dalam format JSON dikonversi ke format TOON untuk meningkatkan efisiensi token. Evaluasi sistem dilakukan menggunakan metrik ROUGE-1, ROUGE-2, dan ROUGE-L untuk jalur RAG, *Execution Accuracy* (EX) untuk jalur data terstruktur, serta *Token Reduction Rate* untuk mengukur efisiensi TOON dibandingkan JSON [14][15].

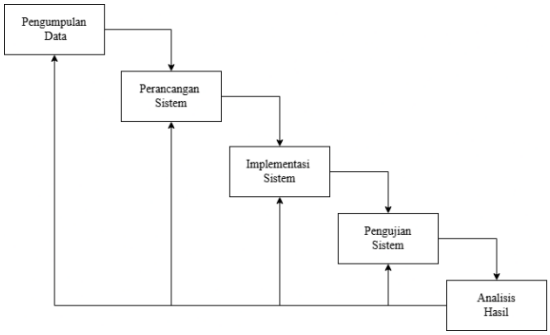
2. METODE PENELITIAN

2.1 Tempat dan Waktu Penelitian

Penelitian ini dilaksanakan di Program Studi Magister Teknologi Informasi, Universitas Malikussaleh yang berlokasi di Kota Lhokseumawe. Waktu pelaksanaan penelitian dimulai dari tahap perencanaan pada bulan April 2026 hingga selesai.

2.2 Skema Penelitian

Penelitian ini dilaksanakan melalui serangkaian tahapan yang tersusun secara sistematis dan berurutan. Tahapan pengembangan sistem mengacu pada metode Waterfall, yaitu metode pengembangan perangkat lunak yang dilakukan secara bertahap dari satu proses ke proses berikutnya. Skema penelitian terdiri dari lima tahapan utama sebagai gambar 1 berikut :



Gambar 1. Skema Penelitian

Pemetaan setiap tahap Waterfall ke dalam aktivitas penelitian pada gambar 1 adalah sebagai berikut:

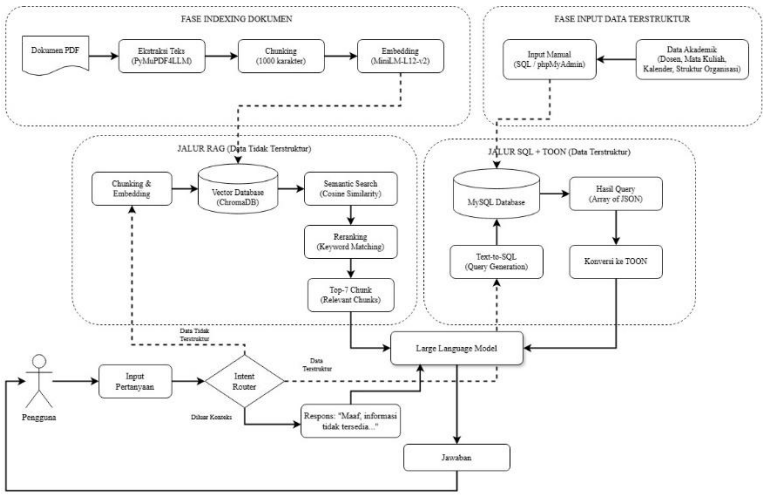
1. Analisis Kebutuhan. Mengidentifikasi permasalahan layanan informasi akademik dan merumuskan kebutuhan sistem, meliputi kemampuan *hybrid retrieval* (RAG dan Text-to-SQL), efisiensi token menggunakan TOON, serta integrasi dengan LLM.
2. Desain Sistem. Merancang arsitektur sistem, alur pemrosesan *intent router*, basis data relasional (MySQL) untuk data terstruktur, dan *vector database* (ChromaDB) untuk data tidak terstruktur, serta menentukan model LLM dan *embedding* yang digunakan.
3. Implementasi. Menerjemahkan desain ke dalam kode program menggunakan Python, ChromaDB, dan MySQL, termasuk fungsi konversi JSON ke TOON.
4. Pengujian. Mengevaluasi kinerja sistem menggunakan metrik ROUGE, *Execution Accuracy*, *Token Reduction Rate*, serta akurasi, presisi, *recall*, dan F1-Score untuk *Intent Router*.
5. Pemeliharaan. Melakukan perbaikan berdasarkan hasil evaluasi, seperti penambahan jumlah *chunk*, perbaikan parsing PDF, dan pengembangan fitur tambahan.

2.3 Proses Sistem

Alur kerja proses sistem *chatbot* akademik ini diawali saat pengguna memberikan *input* pertanyaan yang kemudian dianalisis oleh *Intent Router* untuk menentukan jalur pemrosesan berdasarkan jenis data yang dibutuhkan. Apabila pertanyaan membutuhkan data tidak terstruktur, rute akan diarahkan ke Jalur RAG, di mana sistem akan melakukan *Semantic Search* (*Cosine Similarity*) pada *Vector Database* (ChromaDB) yang berisi data dari proses *chunking* dan *embedding*. Hasil pencarian tersebut kemudian disaring kembali melalui tahap *Reranking* (*Keyword Matching*) untuk mengekstrak *Top-7*

Vol. 11, No. 2, Oktober 2026, Page. 566-577
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Chunk yang paling relevan. Sebaliknya, jika pertanyaan membutuhkan data terstruktur, *Intent Router* akan mengarahkannya ke Jalur SQL + TOON, di mana modul *Text-to-SQL* (*Query Generation*) akan menarik data dari *MySQL Database* untuk menghasilkan *Array of JSON*, yang kemudian dikonversi ke dalam format TOON. Pada tahap akhir, konteks informasi yang didapatkan dari salah satu jalur tersebut beserta pertanyaan asli pengguna akan diteruskan bersama-sama ke *Large Language Model* (LLM) untuk disintesis menjadi sebuah jawaban akhir yang kemudian dikembalikan kepada pengguna sesuai dengan gambar 2 berikut :



Gambar 2. Diagram Proses Sistem

2.4 Metrik Evaluasi

Evaluasi model dalam penelitian ini dilakukan dengan menggunakan beberapa metrik sebagai berikut.

a. ROUGE

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) adalah metrik evaluasi otomatis yang mengukur kemiripan teks hasil sistem dengan teks referensi. Penelitian ini menggunakan ROUGE-1 dan ROUGE-2 (berbasis tumpang tindih *n-gram*) serta ROUGE-L (berbasis *Longest Common Subsequence*), yang masing-masing dihitung menggunakan *precision*, *recall*, dan F1-Score sebagaimana ditunjukkan pada persamaan (1), (2), dan (3).

$$\text{ROUGE-N} = \frac{\text{Jumlah n-gram yang cocok}}{\text{Jumlah n-gram pada referensi}} \quad (1)$$

$$\text{ROUGE-L} = \frac{\text{LCS}}{\text{Panjang teks referensi}} \quad (2)$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

b. *Execution Accuracy* (EX)

Execution Accuracy (EX) adalah metrik yang mengukur persentase query SQL hasil LLM yang menghasilkan data identik dengan *ground truth* dari basis data. Metrik ini digunakan pada jalur data terstruktur karena jawaban yang diharapkan bersifat pasti dan terukur.

$$\text{EX} = \frac{\text{Jumlah query benar}}{\text{Jumlah total pertanyaan uji}} \times 100\% \quad (4)$$

c. Token Reduction Rate

Token Reduction Rate (TRR) mengukur persentase penghematan token ketika representasi data dikonversi dari format JSON ke format TOON. Semakin tinggi nilai TRR, semakin efisien format TOON dalam mengurangi konsumsi token pada komunikasi antara sistem dan LLM.

$$TRR = \frac{Token_{JSON} - Token_{TOON}}{Token_{JSON}} \times 100\% \quad (5)$$

2.5 Protokol Pengujian dan Dataset

Pemilihan 60 sampel pertanyaan tidak dilakukan secara acak, melainkan secara purposif dengan kriteria: (a) mewakili kategori layanan akademik utama (dosen, kalender akademik, mata kuliah, struktur organisasi, prosedur/aturan akademik, dan profil prodi), (b) mencakup variasi panjang dan struktur kalimat (pendek–panjang, pertanyaan tunggal–majemuk) serta (c) mencakup tiga tingkat kesulitan (*factual*, *procedural*, *multi-entity*). Dengan demikian, 60 sampel tersebut diupayakan cukup bervariasi untuk merepresentasikan kebutuhan informasi akademik.

Pemisahan sampel ke jalur terstruktur dan tidak terstruktur didasarkan pada sifat jawaban yang dibutuhkan pertanyaan yang jawabannya berupa fakta terukur dan tersimpan dalam basis data (misalnya nama dosen, SKS mata kuliah, tanggal kalender akademik) dikelompokkan ke jalur SQL, sedangkan pertanyaan yang jawabannya berupa narasi, prosedur, atau aturan yang tertulis dalam dokumen akademik dikelompokkan ke jalur RAG. Pemisahan ini ditetapkan pada tahap penyusunan ground truth dan diverifikasi oleh validator, berikut contohnya pada tabel 1 dan 2 :

Tabel 1. Distribusi Tingkat Kesulitan Dataset Uji

Kategori	Factual	Procedural	Multi-entity	Total
RAG	12	12	6	30
SQL	15	—	15	30
Total	27	12	21	60

Tabel 2. Contoh Pertanyaan uji per jalur

Jalur	Contoh Pertanyaan	Tingkat Kesulitan
RAG	Berapa jumlah kata maksimal dalam penulisan abstrak tesis?	Factual
RAG	Bagaimana prosedur pengajuan judul tesis?	Procedural
SQL	Siapa Dekan Fakultas Teknik?	Factual
SQL	Apa saja mata kuliah semester 1 beserta kode dan SKS-nya?	Multi-entity

2.6 Detail Implementasi dan Pengendalian Variabel

Implementasi sistem menggunakan parameter yang ditetapkan pada tabel 3 agar eksperimen dapat direplikasi

Tabel 3. Parameter Implementasi Sistem

Komponen	Parameter	Nilai/Konfigurasi
Chunking	Ukuran chunk / overlap	1000 / 200 karakter
Embedding	Model	paraphrase-multilingual-MiniLM-L12-v2
Retrieval	Metode / top-k	Cosine similarity / 7
Reranking	Metode	Keyword matching
LLM	Model	meta-llama/llama-4-scout-17b-16e-instruct
LLM	Temperature / max_tokens	0.1 / 1024
Vector DB	Database	ChromaDB
Basis data	DBMS	MySQL

Reranking menghitung skor kecocokan kata kunci tiap chunk (jumlah kata kunci cocok dibagi jumlah kata kunci pertanyaan) lalu memilih 7 chunk teratas, prompt Intent Router dan Text-to-SQL bersifat statis dengan LLM diberi daftar kategori jalur beserta skema tabel untuk mengembalikan label jalur atau query SQL konversi JSON ke TOON menghilangkan tanda kutip, kurung kurawal, dan koma dan pengendalian variabel dilakukan dengan menggunakan 30 pertanyaan identik, model, temperature,

Vol. 11, No. 2, Oktober 2026, Page. 566-577
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

serta konteks data yang sama pada pengujian JSON vs TOON, sehingga satu-satunya variabel yang diubah adalah format representasi dan perbedaan token maupun ROUGE dapat diatribusikan pada format yang diuji.

3. HASIL DAN PEMBAHASAN

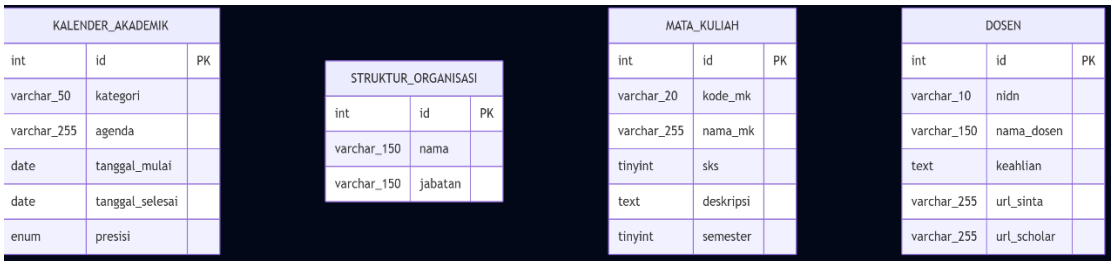
3.1. Persiapan Data

Data tidak terstruktur diperoleh dari beberapa dokumen akademik resmi yang dikumpulkan melalui situs web kampus. Proses pengumpulan data dilakukan dengan menyalin konten dari halaman-halaman tersebut dan menyusunnya kembali ke dalam format PDF agar kompatibel untuk diproses oleh sistem.

Tabel 4. Data Tidak Terstruktur

No	Nama Dokumen	Keterangan
1	Buku Panduan Akademik	Memuat informasi akademik Universitas
2	Buku Panduan Penulisan Proposal dan Tesis	Memuat aturan dan format penulisan proposal serta tesis
3	Profil Program Studi	Memuat visi, misi, kontak, lokasi, dan tujuan program studi

Data terstruktur pada tabel 4 diperoleh dari sumber resmi, yaitu situs web kampus serta beberapa sumber lain yang relevan. Data yang dikumpulkan mencakup informasi dosen, kalender akademik, mata kuliah, dan struktur organisasi program studi. Data tersebut kemudian disusun ke dalam basis data relasional menggunakan MySQL agar dapat diakses oleh sistem melalui mekanisme query pada jalur data terstruktur sesuai dengan gambar 3 berikut :



Gambar 3. Data Terstruktur

3.2. Implementasi Sistem

a. Pemrosesan Data Tidak Terstruktur

Pada jalur ini, dokumen akademik dalam format PDF diekstraksi dan dipecah menjadi potongan-potongan teks kecil (*chunking*) dengan ukuran 1000 karakter dan *overlap* 200 karakter untuk menjaga kontinuitas informasi. Setiap *chunk* kemudian diubah menjadi representasi vektor (*embedding*) menggunakan model *paraphrase-multilingual-MiniLM-L12-v2* dan disimpan ke dalam *vector database* ChromaDB. Ketika pengguna mengajukan pertanyaan, sistem mengubah pertanyaan tersebut ke dalam vektor dan mencari *chunk* dengan kemiripan makna tertinggi menggunakan *cosine similarity* dengan jumlah *top-k* = 7, kemudian dilakukan *reranking* berbasis *keyword matching* untuk menyaring hasil paling relevan sebelum dikirim sebagai konteks ke LLM, berikut adalah contohnya pada tabel 5 :

Tabel 5. Contoh Potongan Dokumen Hasil Chunking

ID Chunk	Sumber Dokumen	Isi Chunk
Buku Panduan Akademik_chunk_0	Buku Panduan Akademik	[Sumber: Buku Panduan Akademik | Topik: **BAB I KETENTUAN UMUM ** Pasal 1 Dalam Peraturan Akademik ini yang dimaksud dengan: (1) Peraturan Akademik adalah semua....

Vol. 11, No. 2, Oktober 2026, Page. 566-577
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

b. Jalur Data Terstruktur dan Konversi ke TOON

Apabila hasil klasifikasi menentukan bahwa pertanyaan harus dijawab melalui jalur data terstruktur, sistem akan memproses pertanyaan tersebut dengan pendekatan *Text-to-SQL*, yaitu LLM secara dinamis menghasilkan perintah *query* SQL berdasarkan pertanyaan pengguna dan tabel basis data yang telah ditentukan pada tahap klasifikasi sebelumnya, sesuai denga kode pada gambar 4 :

```
if routing["sumber"] == "database":
    data, sql_query = tanya_mysql(pertanyaan)

    if data:
        konteks_toon = _encode_ke_toon(data, pertanyaan)
```

Gambar 4. Eksekusi Text-to-SQL dan Konversi Data TOON

Query SQL yang dihasilkan kemudian dieksekusi terhadap basis data MySQL, dan hasilnya ditangkap dalam bentuk *array of JSON objects*. Hasil dalam format JSON tersebut selanjutnya dikonversi ke dalam format TOON menggunakan fungsi konversi khusus yang telah dirancang seperti pada tabel 6 :

Tabel 6. Contoh Konversi Data dari Format JSON ke TOON

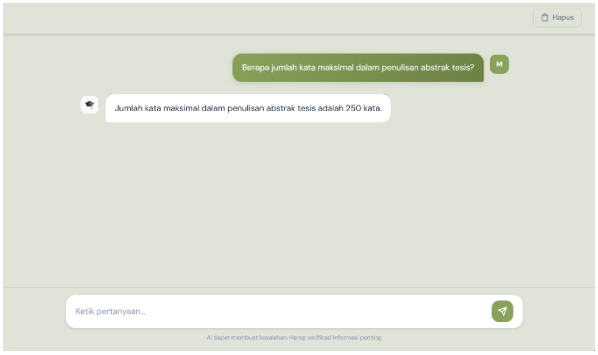
Format	**Contoh Data**
JSON (Input)	`[{"id":1,"kode_mk":"MTI113","nama_mk":"Aset dan Infrastruktur Teknologi Informasi","sks":3,"deskripsi":"Mata kuliah Aset dan Infrastruktur...","semester":1},{"id":2,"kode_mk":"MTI123","nama_mk":"Manajemen Strategis Teknologi Informasi","sks":3,"deskripsi":"...","semester":1}]`

TOON (Output)	`mata_kuliah[4]{id,kode_mk,nama_mk,sks,deskripsi,semester}:`
	`1,MTI113,Aset dan Infrastruktur Teknologi Informasi,3,"...",1`
	`2,MTI123,Manajemen Strategis Teknologi Informasi,3,"...",1`
	`3,MTI133,Manajemen Data dan Algoritma,3,"...",1`
	`4,MTI143,Tata Kelola Teknologi Informasi,3,"...",1`

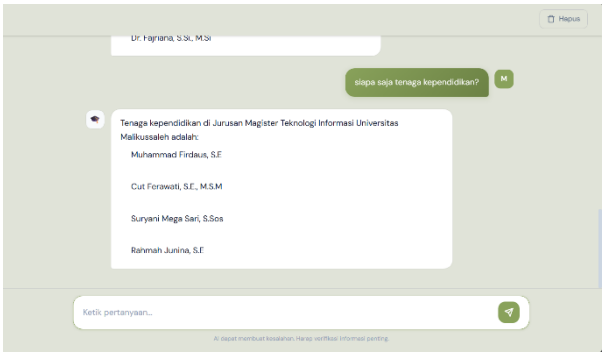
Format TOON menghilangkan tanda kutip, kurung kurawal, dan karakter struktural JSON yang tidak diperlukan, sehingga menghasilkan representasi data yang lebih ringkas.

c. Tampilan Antarmuka *Chatbot*

Antarmuka pengguna sistem *chatbot* akademik dirancang berbasis *web* yang responsif untuk memastikan pengguna dapat berinteraksi dengan *chatbot* secara mudah, baik melalui perangkat komputer maupun *smartphone* .



Gambar 5. Respons Data Tidak Terstruktur



Gambar 6. Respons Data Terstruktur

Vol. 11, No. 2, Oktober 2026, Page. 566-577
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Tampilan antarmuka sistem pada Gambar 5 menunjukkan respons terhadap pertanyaan data tidak terstruktur, di mana *Intent Router* mengklasifikasikan kueri ke jalur RAG dan sistem menampilkan jawaban dari dokumen akademik. Sementara itu, Gambar 6 memperlihatkan penanganan pertanyaan data terstruktur, di mana *router* mengarahkan kueri ke jalur *Text-to-SQL* untuk mengambil informasi dari basis data. Kedua skenario membuktikan bahwa *router* mampu mengarahkan pertanyaan ke sumber data yang tepat sesuai dengan konteks pertanyaan.

3.3 Hasil Pengujian Sistem

a. Hasil Evaluasi Jalur Data Terstruktur

Evaluasi pada jalur data terstruktur dilakukan untuk mengukur dua aspek, yaitu ketepatan hasil *query* SQL yang dihasilkan secara otomatis oleh LLM (*Execution Accuracy*) dan efisiensi penggunaan token pada hasil konversi data ke format TOON dibandingkan format JSON sesuai dengan tabel 7 dan 8 :

Tabel 7. Hasil Evaluasi *Execution Accuracy* (EX)

Kategori	Jumlah	Persentase
Jumlah pertanyaan uji	30	100%
Jawaban benar (skor 1)	30	100%
Jawaban salah (skor 0)	0	0%
Execution Accuracy (EX)		100%

Tabel 8. Hasil Evaluasi Jalur Data Terstruktur

Pertanyaan	Jawaban Referensi	Jawaban LLM	EX Score
Siapa Dekan Fakultas Teknik?	Dekan Fakultas Teknik adalah Dr. Muhammad Daud, S.T., M.T.	Dekan Fakultas Teknik adalah Dr. Muhammad Daud, S.T., M.T.	1
Siapa Koordinator Program Studi Magister Teknologi Informasi?	Koordinator Program Studi Magister Teknologi Informasi adalah Dr. Eng. Defry Hamdhana, S.T., M.Kom	Koordinator Program Studi Magister Teknologi Informasi adalah: 1. Dr. Eng. Defry Hamdhana, S.T., M.Kom	1

Sistem *chatbot* pada jalur data terstruktur memperoleh nilai *Execution Accuracy* sebesar 100%, yang menunjukkan bahwa seluruh *query* SQL yang dihasilkan secara otomatis oleh LLM berhasil menghasilkan data yang identik dengan data acuan pada setiap pertanyaan uji yang diberikan. Berikut ini adalah tabel 9 hasil dari efisiensi token format JSON vs TOON :

Tabel 9. Hasil Efisiensi Token Format JSON vs TOON pada 30 Pertanyaan Uji

Metrik Statistik	Token JSON	Token TOON	Hemat Token	Hemat Persen
Nilai Minimum	241	221	20	8,30%
Nilai Maksimum	663	367	296	44,65%
Rata-rata	258,40	192,40	66	24,60%

b. Hasil Evaluasi Jalur Data Tidak Terstruktur

Pengujian dilakukan dengan membandingkan jawaban yang dihasilkan oleh sistem terhadap jawaban referensi menggunakan metrik ROUGE, yang mencakup tiga varian, yaitu ROUGE-1, ROUGE-2, dan ROUGE-L sesuai dengan tabel 10 dan tabel 11 :

Tabel 10. Hasil Statistik Hasil Evaluasi ROUGE pada 30 Pertanyaan Uji Jalur RAG

Metrik	ROUGE-1 (F1)	ROUGE-2 (F1)	ROUGE-L (F1)
Nilai Minimum	36,84%	27,78%	36,84%
Nilai Maksimum	100%	100%	100%

Rata-rata	79,13%	69,01%	78,50%

Tabel 11. Contoh Hasil Evaluasi Jalur Data Tidak Terstruktur

Pertanyaan	**Jawaban Referensi**	**Jawaban LLM**	**ROUGE -1**	**ROUGE -2**	**ROUGE -L**
Apa website resmi Jurusan/Program Studi MTI Unimal?	Website resmi MTI Unimal adalah https://mti.unimal.ac.id	Website resmi Jurusan/Program Studi Magister Teknologi Informasi Unimal adalah https://mti.unimal.ac.id.	72,0%	60,87%	72,0%
Berapa jumlah kata maksimal dalam penulisan abstrak tesis?	Jumlah kata maksimal dalam penulisan abstrak tesis adalah 250 kata	Jumlah kata maksimal dalam penulisan abstrak tesis adalah 250 kata. Abstrak juga harus disertai kata kunci maksimal 5 kata.	68,97%	66,67%	68,97%

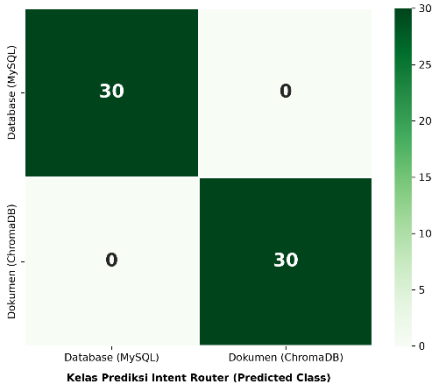
c. Hasil Evaluasi Kinerja *Intent Router*

Intent Router bertugas mengklasifikasikan pertanyaan pengguna ke jalur RAG (data tidak terstruktur) atau SQL+TOON (data terstruktur). Pengujian dilakukan terhadap 60 sampel pertanyaan (30 RAG dan 30 SQL).

Tabel 12. Hasil Evaluasi Kinerja Intent Router pada 60 Sampel Pertanyaan Uji

Metrik Evaluasi	**Hasil Capaian**	**Keterangan**
Akurasi	100% (60/60)	Seluruh pertanyaan diklasifikasikan dengan benar
Presisi	100% (30/30)	Tidak ada *FP* (pertanyaan dokumen salah ke SQL)
Recall	100% (30/30)	Tidak ada FN (pertanyaan SQL terlewat)
F1-Score	100%	Keseimbangan sempurna presisi dan *recall*

Berdasarkan Tabel 12, *router* memperoleh akurasi, presisi, recall, dan F1-Score semuanya sebesar 100% . Nilai presisi 100% menunjukkan tidak ada pertanyaan dokumen yang salah diarahkan ke basis data (*FP = 0*), sedangkan recall 100% menandakan seluruh pertanyaan basis data berhasil diarahkan ke jalur yang tepat (*FN = 0).*



Gambar 7. *Confusion Matrix* Klasifikasi Intent Router

Confusion matrix pada gambar 7 memperkuat temuan tersebut, di mana seluruh 60 sampel berada pada sel diagonal (prediksi sesuai aktual) tanpa ada kesalahan klasifikasi lintas jalur.

Kinerja *router* yang sempurna menjadi fondasi penting karena kesalahan klasifikasi akan berdampak langsung pada kualitas jawaban akhir yang dihasilkan sistem.

d. Hasil Evaluasi Dampak Format TOON terhadap Kualitas Jawaban LLM

Pengujian pada sub-bab ini menggunakan 30 pertanyaan uji yang sama dengan jalur data terstruktur, sehingga perbandingan JSON dan TOON dilakukan pada kondisi data yang identik.

Tabel 13. Perbandingan Kualitas Jawaban dan Efisiensi Token Format JSON vs TOON

Metrik	Format JSON	Format TOON	Selisih
Rata-rata Token	258,40	192,40	24,60%
ROUGE-1 F1 (%)	74,22	74,59	+0,37%
ROUGE-2 F1 (%)	56,90	58,22	+1,32%
ROUGE-L F1 (%)	69,86	71,14	+1,28%

Hasil pada tabel 13 menunjukkan bahwa format TOON menghemat konsumsi token rata-rata sebesar 24,60% dan menghasilkan kualitas jawaban yang setara atau sedikit lebih tinggi dibandingkan JSON pada seluruh metrik ROUGE. Namun, selisih skor ROUGE yang teramati relatif kecil sehingga belum dapat diklaim sebagai peningkatan kualitas yang signifikan. Klaim utama yang didukung bukti empiris adalah bahwa TOON tidak menurunkan kualitas jawaban sekaligus mengurangi konsumsi token, sehingga format TOON layak digunakan sebagai representasi data yang lebih efisien pada sistem berbasis LLM.

Hasil penghematan token TOON sejalan dengan temuan Matveev [13] yang melaporkan efisiensi TOON dibandingkan JSON, sekaligus memperluas penerapannya pada konteks chatbot akademik berbasis hybrid retrieval yang belum diteliti sebelumnya. Dibandingkan Pangestu dkk. [9] yang memperoleh F1-Score 53,36% pada chatbot akademik berbasis RAG, sistem ini memperoleh F1-Score ROUGE-1 79,13%, meskipun dengan cakupan data dan konfigurasi model yang berbeda sehingga perbandingan bersifat indikatif.

3.4 Analisis Kegagalan

Meskipun sistem menunjukkan kinerja yang baik secara keseluruhan, terdapat beberapa pertanyaan uji pada jalur RAG yang memperoleh skor ROUGE di bawah rata-rata. Berdasarkan tabel 14, pertanyaan dengan skor terendah (ROUGE-1 = 36,84%) adalah:

Tabel 14. Contoh Hasil Evaluasi ROUGE pada Jalur RAG

Aspek	Detail
Pertanyaan	Aplikasi referensi apa yang wajib digunakan dalam penulisan sitasi tesis?
Jawaban Referensi	Aplikasi referensi yang wajib digunakan adalah Mendeley Desktop atau Mendeley Reference.
Jawaban Sistem	Berdasarkan Buku Panduan Penulisan Proposal dan Tesis, aplikasi referensi yang wajib digunakan dalam penulisan sitasi tesis adalah Mendeley dengan style IEEE (Institute of Electrical and Electronics Engineering).

Penyebab Skor Rendah:

1. Perbedaan Struktur Kalimat. Jawaban sistem lebih panjang dan menambahkan informasi "style IEEE" yang tidak ada di referensi, menyebabkan banyak *n-gram* tidak cocok.
2. Kata Kunci Tidak Tepat. Referensi menyebut "Mendeley Desktop" dan "Mendeley Reference", sedangkan sistem hanya menyebut "Mendeley", sehingga mengurangi kecocokan *n-gram*.
3. Informasi Tambahan. Sistem menambahkan konteks "Berdasarkan Buku Panduan..." dan "style IEEE" yang tidak diminta, sehingga menurunkan skor ROUGE meskipun jawaban secara faktual benar.

4. KESIMPULAN

Penelitian ini mengembangkan *chatbot* akademik berbasis LLM dengan *hybrid retrieval* (RAG dan

SQL+TOON) pada Program Studi Magister Teknologi Informasi Universitas Malikussaleh. Sistem dirancang untuk mengintegrasikan data tidak terstruktur dan terstruktur, serta memberikan jawaban cepat dan kontekstual. Berdasarkan hasil pengujian, sistem memperoleh rata-rata F1-Score ROUGE-1 79,13%, ROUGE-2 69,01%, ROUGE-L 78,50%, dan *Execution Accuracy* 100%. TOON menghemat token rata-rata 24,60% tanpa menurunkan kualitas jawaban; selisih ROUGE yang kecil belum dapat diklaim sebagai peningkatan signifikan. Namun demikian, penelitian ini memiliki keterbatasan yang perlu disampaikan secara eksplisit. Jumlah sampel data uji yang digunakan terbatas, yaitu 30 pertanyaan untuk jalur RAG, 30 pertanyaan untuk jalur data terstruktur, dan 60 sampel untuk Intent Router, serta hanya mencakup tiga dokumen akademik dan satu program studi. Oleh karena itu, hasil eksperimen,baik dari segi performa hybrid retrieval maupun persentase efisiensi token format TOON masih bersifat indikatif dan belum dapat digeneralisasi untuk skala operasional yang lebih luas. Generalisasi memerlukan pengujian bertahap pada dataset yang lebih besar, cakupan dokumen dan program studi yang lebih beragam, serta validasi dengan pengguna akhir. Saran penelitian selanjutnya meliputi perluasan cakupan data, *fine-tuning* LLM, uji coba berbagai model, pengujian UAT/SUS, serta evaluasi berbasis RAGAS.

5. DAFTAR PUSTAKA

[1] B. Prasojo, M. Huda, I. N. Khasanah, and E. Wahyuningsih, "Aplikasi Chatbot Berbasis Telegram Untuk Layanan Informasi dan Akademik Kampus Universitas Ma'arif Nahdlatul Ulama Kebumen," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, pp. 893–902, Apr. 2024, doi: 10.23960/jitet.v12i2.4013.

[2] P. B. Wintoro, H. Hermawan, M. A. Muda, and Y. Mulyani, "Implementasi Long Short-Term Memory pada Chatbot Informasi Akademik Teknik Informatika Unila," *EXPERT: Jurnal Manajemen Sistem Informasi dan Teknologi*, vol. 12, no. 1, pp. 68–75, Jun. 2022, doi: 10.36448/expert.v12i1.2593.

[3] Herman, A. Yulianto, and S. Lim, "Analisa Aplikasi Chatbot Untuk Pertanyaan Umum Terkait Universitas di Universitas XYZ," *Jurnal Pendidikan Teknologi Informasi (JUKANTI)*, vol. 6(2), pp. 415–437, 2023, doi: 10.37792/jukanti.v6i2.1069.

[4] A. S. Susanto and L. Lisana, "Chat AI LLM (Large Language Model) di Universitas: Tinjauan Sistematis terhadap Variabel Pendorong Adopsi, Kerangka Teoritis, dan Tren Global," *JURNAL LOCUS: Penelitian & Pengabdian*, vol. 4 (10), pp. 9380–9398, 2025, doi: 10.58344/locus.v4i10.4324.

[5] I. K. R. Arthana, N. Gunantara, M. Sudarma, and M. Sukarsa, "A Systematic Literature Review of Retrieval-Augmented Generation Implementation for Enhancing Large Language Models in Education," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 15, no. 1, pp. 91–109, Mar. 2026, doi: 10.23887/janapati.v15i1.112281.

[6] S. Elysia and Herianto, "Chatbot Berbasis Retrieval Augmented Generation (RAG) untuk Peningkatan Layanan Informasi Sekolah," *Journal Technology Information and Data Analytic*, vol. 1(2), pp. 52–58, 2024, doi: 10.70491/tifda.v1i2.52.

[7] A. T. U. BR. Lubis, N. S. Harahap, S. Agustian, M. Irsyad, and I. Afrianty, "Question Answering System pada Chatbot Telegram Menggunakan Large Language Models (LLM) dan Langchain (Studi Kasus UU Kesehatan)," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 3, pp. 955–964, May 2024, doi: 10.57152/malcom.v4i3.1378.

[8] I. Pujiono, I. M. Agtyaputra, and Y. Ruldeviyani, "Implementing Retrieval-Augmented Generation And Vector Databases for Chatbots in Public Services Agencies Context," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 10, no. 1, pp. 216–223, Aug. 2024, doi: 10.33480/jitk.v10i1.5572.

[9] A. Pangestu, D. Hamdhana, and R. Suwanda, "Student Academic Consultation Chatbot Using Meta AI Large Language Models and Retrieval-Augmented Generation," *Journal of Artificial Intelligence and Software Engineering*, vol. 6, no. 1, pp. 26–36, 2026, doi: 10.30811/jaise.v6i1.8875.

[10] G. D. Albert and A. Voutama, "Pengembangan Chatbot Berbasis PDF Menggunakan Local Retrieval-Augmented Generation (RAG) Dan Ollama," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 2, pp. 937–944, Apr. 2025, doi: 10.23960/jitet.v13i2.6361.

[11] M. A. Basallamah, L. S. Riza, and A. Anisyah, "Pengembangan Chatbot Informasi Kesehatan Ibu dan Anak Jawa Barat Berbasis Hybrid RAG dan TextToSQL," *Jurnal Komputer Teknologi Informasi Sistem Informasi (JUKTISI)*, vol. 4, no. 3, pp. 1895–1902, Jan. 2026, doi: 10.62712/juktisi.v4i3.783.

[12] A. Kachhwal, S. Gupta, A. Jangid, and A. Sharma, "A Multi-Modal Hybrid Search System using RAG and Text2SQL for Semantic Information Retrieval," *Anshul Kachhwal*, vol. 13, no. 2, pp. 1–9, 2025, Accessed: Apr. 22, 2026. [Online]. Available: https://www.ijset.in/wp-content/uploads/IJSET_V13_issue2_450.pdf

[13] I. Matveev, "Token-Oriented Object Notation vs JSON: A Benchmark of Plain and Constrained Decoding Generation," pp. 1–9, Feb. 2026, doi: 10.48550/arXiv.2603.03306.

[14] F. F. Kartamanah, A. R. Atmadja, and I. Budiman, "Analyzing PEGASUS Model Performance with ROUGE on Indonesian News Summarization," *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, vol. 9(1), pp. 31–42, Jan. 2025, doi: 10.33395/sinkron.v9i1.14303.

[15] R. Suwanda, S. F. Anshari, and W. Ningsih, "Information System for Operational Goods Management at the Career Guidance and Entrepreneurship Center Malikussaleh University," *Jurnal Inotera*, vol. 9, no. 1, pp. 164–169, May 2024, doi: 10.31572/inotera.vol9.iss1.2024.id324.