Skema Pendanaan: UTAMA

LAPORAN PENELITIAN

PENGEMBANGAN CHATBOT CERDAS BERBASIS AGENTIC AI DENGAN PENDEKATAN RETRIEVAL-AUGMENTED GENERATION (RAG) UNTUK PENINGKATAN PEMAHAMAN KONTEKSTUAL DAN AUTONOMI TINDAKAN

TIM PENELITI

Ketua : Dr. Mohammad Syafrullah, M.Kom, M.Sc (960008)
Anggota : Anwar Rifa'I, S.Pd., M.Pd. (190018)
Riskiana Wulan, S.Kom., M.Kom. (150056)

FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS BUDI LUHUR

Februari 2026

HALAMAN PENGESAHAN
LAPORAN PENELITIAN

Judul Penelitian	: Pengembangan Chatbot Cerdas Berbasis Agentic AI dengan Pendekatan Retrieval-Augmented Generation (RAG) untuk Peningkatan Pemahaman Kontekstual dan Autonomi Tindakan
Bidang Penelitian	: Teknik Informatika
Ketua Peneliti	
a. Nama Lengkap	: Dr. Mohammad Syafrullah, M.Kom, M.Sc
b. NIP/NIDN/ID-SINTA	: 960008/0316097401/5984532
c. Jabatan Fungsional	: Lektor
d. Program Studi	: Teknik Informatika
e. Nomor HP	: 08119119762
f. Alamat e-mail	: muhammad.syafrullah@budiluhur.ac.id
Anggota Peneliti (1)	
a. Nama Lengkap	: Anwar Rifa'i, M.Pd
b. NIP/NIDN/ID-SINTA	: 190018/0303129401/ 6689661
Anggota Peneliti (2)	
a. Nama Lengkap	: Riskiana Wulan, S.Kom., M.Kom
b. NIP/NIDN/ID-SINTA	: 150056/0313048901/6888430
Mahasiswa (1)	
a. Nama Lengkap	: Almas Ikbar Fahriza
b. NIM	: 2211501693
Lama Penelitian	: 6 bulan
Biaya Penelitian	
a. Sumber Universitas Budi Luhur	: Rp. 15.000.000,00
b. Sumber lain (sebutkan jika ada)	: Rp. 0

Jakarta, 26 Februari 2026

Mengetahui,
Dekan/Kepala Pusat Studi

(Dr. Ir. Achmad Solichin, S.Kom., M.T.I.)
NIP. 050023

Ketua Pelaksana

(Dr. Mohammad Syafrullah, M.Kom, M.Sc)
NIP. 960008

Menyetujui,
Direktur Riset dan Pengabdian Kepada Masyarakat

(Prof. Dr. Ir. Prudensius Maring, M.A.)
NIP 190043

No. Registrasi : 009 01 LPJ 02 26
Tanggal : 26 02 26 Paraf :

RINGKASAN

Penelitian ini merespons keterbatasan fundamental pada chatbot berbasis *Large Language Model* (LLM) saat ini yang cenderung pasif dan rentan terhadap halusinasi pengetahuan. Urgensi penelitian difokuskan pada pengembangan arsitektur hibrida yang mensinergikan pendekatan *Retrieval-Augmented Generation* (RAG) untuk menjamin validitas informasi dengan paradigma *Agentic AI* untuk memberikan kapabilitas eksekusi tindakan secara otonom. Tujuan utama riset ini adalah merancang dan membangun prototipe chatbot cerdas yang mampu memahami konteks dokumen institusi secara mendalam serta bertindak otonom dalam menyelesaikan tugas layanan kampus. Metode yang diterapkan adalah pengembangan prototipe (*prototype development*) yang meliputi tahapan perancangan arsitektur, implementasi sistem, dan evaluasi kinerja. Hingga saat ini, penelitian telah berhasil merealisasikan arsitektur "Lingkaran Persepsi-Aksi" (*Perception-Action Loop*) menggunakan platform orkestrasi n8n, integrasi model Ollama, serta basis data vektor Pinecone yang terbukti mampu melakukan pencarian semantik dengan skor relevansi mencapai 0,9989. Hasil pengujian menunjukkan bahwa pendekatan *Direct Natural Language Query* lebih fleksibel dalam menangani variasi bahasa pengguna dibandingkan sistem berbasis perintah fungsi kaku. Penelitian ini juga melakukan studi komparatif antara model lokal (Ollama) dan *cloud* (OpenAI), di mana ditemukan bahwa model *cloud* unggul dalam kecepatan respons (~200-600ms), sedangkan model lokal menjanjikan privasi data yang lebih terjamin dan keberlanjutan biaya. Target luaran penelitian ini adalah publikasi pada jurnal nasional terakreditasi Sinta 2 yang saat ini berstatus draf, serta Hak Kekayaan Intelektual (HKI) berupa Hak Cipta program komputer yang telah berstatus *granted*.

Kata Kunci: Chatbot Cerdas; Agentic AI; Retrieval-Augmented Generation (RAG); Pemahaman Kontekstual; Autonomi Tindakan.

PRAKATA

Puji syukur kehadirat Tuhan Yang Maha Esa atas segala rahmat dan hidayah-Nya, sehingga penulis dapat menyelesaikan laporan akhir penelitian yang berjudul **"Pengembangan Chatbot Cerdas Berbasis Agentic AI dengan Pendekatan Retrieval-Augmented Generation (RAG) untuk Peningkatan Pemahaman Kontekstual dan Autonomi Tindakan"**. Laporan ini disusun sebagai bentuk pertanggungjawaban ilmiah atas pengembangan arsitektur hibrida yang mengintegrasikan teknologi LLM, sistem retrival dokumen, dan kapabilitas eksekusi tindakan otonom guna memitigasi halusinasi informasi pada layanan digital kampus.

Penelitian ini berhasil merealisasikan prototipe fungsional dengan arsitektur *"Perception-Action Loop"* yang diharapkan dapat menjadi referensi strategis dalam implementasi kecerdasan buatan yang lebih faktual dan proaktif di lingkungan akademis. Terselesaikannya karya ini tidak lepas dari bimbingan serta dukungan berbagai pihak, oleh karena itu penulis menyampaikan terima kasih yang tulus kepada:

1. **Bapak Prof. Dr. Agus Setyo Budi, M.Sc.**, selaku Rektor Universitas Budi Luhur, atas dukungan penuh dan fasilitas yang diberikan untuk menunjang produktivitas riset di lingkungan universitas.
2. **Bapak Dr. Ir. Achmad Solichin, S.Kom., M.T.I.**, selaku Dekan Fakultas Teknologi Informasi Universitas Budi Luhur, atas arahan dan motivasi dalam pengembangan inovasi di bidang teknologi informasi.
3. **Bapak Prof. Dr. Ir. Prudensius Maring, M.A.**, selaku Kepala Direktorat Riset dan Pengabdian Kepada Masyarakat Universitas Budi Luhur, atas koordinasi dan bantuan administratif yang memungkinkan penelitian ini terlaksana dengan baik.
4. **Rekan-rekan di AI Center Universitas Budi Luhur**, atas ruang diskusi dan kolaborasi teknis selama proses pengembangan sistem.

Penulis menyadari bahwa laporan ini masih memiliki ruang untuk penyempurnaan, terutama dalam hal optimalisasi latensi sistem hibrida dan pengembangan ke arah *Knowledge Graph*. Kritik dan saran yang membangun sangat diharapkan demi perbaikan di masa mendatang. Semoga laporan ini memberikan manfaat nyata bagi perkembangan teknologi *Artificial Intelligence* di Indonesia.

Jakarta, Februari 2026

Penulis,

DAFTAR ISI

HALAMAN SAMPUL i
HALAMAN PENGESAHAN ii
RINGKASAN iii
PRAKATA iv
DAFTAR ISI v
DAFTAR TABEL vi
DAFTAR GAMBAR vii
DAFTAR LAMPIRAN viii
BAB I PENDAHULUAN 1
1. Latar Belakang dan Rumusan Masalah 1
2. Pendekatan Pemecahan Masalah 2
3. *State of the art* dan Kebaruan 3
4. Peta jalan (road map) penelitian 5 tahun 3
BAB II METODE 5
BAB III HASIL PELAKSANAAN PENELITIAN 9
BAB IV KESIMPULAN DAN SARAN 15
5.1 Kesimpulan 15
5.2 Saran 15
DAFTAR PUSTAKA 17
LAMPIRAN 19

DAFTAR TABEL

Tabel 1. Road Map Penelitian 3
Tabel 2 Perbandingan Kinerja Prototipe 12

DAFTAR GAMBAR

Gambar 1. Diagram Alur Penelitian 5
Gambar 2. Implementasi Pipeline Agentic-RAG pada Platform n8n.....9
Gambar 3. Visualisasi Penyimpanan Vektor pada Pinecone Dashboard.....10
Gambar 4 Diagram Sekuensial Aliran Data (Sequence Diagram) 11

DAFTAR LAMPIRAN

Lampiran 1. Realisasi Penggunaan Anggaran 19
Lampiran 2. Surat Perjanjian Kontrak Penelitian 20
Lampiran 3. Catatan Harian.....23
Lampiran 4. Artikel Ilmiah (draft).....24

BAB I
PENDAHULUAN

1. Latar Belakang dan Rumusan Masalah

Era digital kontemporer ditandai oleh kemajuan pesat dalam kecerdasan buatan (AI), khususnya pada Large Language Models (LLMs) yang telah merevolusi kemampuan agen percakapan atau chatbot (1) . Kemampuan LLMs untuk memahami dan menghasilkan teks yang menyerupai manusia telah membuka potensi aplikasi yang luas, mulai dari asisten pribadi hingga layanan pelanggan yang kompleks (2). Namun, di balik kemampuannya yang impresif, LLMs mewarisi dua keterbatasan fundamental yang menghambat efektivitasnya dalam skenario dunia nyata (3).

Keterbatasan pertama adalah ketergantungan pada pengetahuan internal yang statis. Pengetahuan sebuah LLM terbatas pada data yang digunakan saat proses pelatihannya, sehingga model ini tidak memiliki akses ke informasi yang terjadi setelahnya (4). Keterbatasan ini sering kali menyebabkan model memberikan jawaban yang usang atau, lebih buruk lagi, menghasilkan informasi yang salah (dikenal sebagai "halusinasi") ketika dihadapkan pada topik yang tidak diketahuinya (5). Keterbatasan kedua adalah sifatnya yang pasif. Chatbot konvensional, bahkan yang berbasis LLM canggih sekalipun, pada dasarnya adalah sistem reaktif yang hanya dapat merespons perintah pengguna (6). Mereka tidak memiliki kemampuan untuk mengambil inisiatif atau melakukan tindakan nyata, seperti menjadwalkan rapat, melakukan pemesanan, atau berinteraksi dengan perangkat lunak lain untuk menyelesaikan sebuah tugas.

Untuk mengatasi masalah pengetahuan statis, komunitas riset memperkenalkan pendekatan Retrieval-Augmented Generation (RAG) (7). RAG memperkaya LLM dengan kemampuan untuk mengambil informasi relevan dari sumber data eksternal secara real-time sebelum menghasilkan respons, sehingga secara signifikan meningkatkan akurasi dan faktualitas jawaban (8). Sementara itu, untuk mengatasi sifat pasif, lahir paradigma Agentic AI, di mana LLM diberdayakan untuk menggunakan "alat" (tools) digital, membuat perencanaan, dan mengeksekusi serangkaian tindakan secara otonom untuk mencapai tujuan yang diberikan (9).

Meskipun kedua pendekatan ini menunjukkan hasil yang menjanjikan secara terpisah, terdapat celah riset yang signifikan dalam integrasi sinergis keduanya. Sistem RAG saat ini cenderung hanya berfokus pada penyediaan informasi yang akurat (menjadi "ensiklopedia" yang lebih baik), sementara sistem Agentic AI seringkali bertindak berdasarkan pengetahuannya yang terbatas dan rentan melakukan kesalahan perencanaan. Penelitian ini berangkat dari premis bahwa chatbot yang benar-benar cerdas dan fungsional harus mampu menggabungkan keduanya: memahami dunia secara akurat melalui data eksternal (persepsi) dan bertindak di dunia digital secara otonom (aksi).

Berdasarkan latar belakang tersebut, maka rumusan permasalahan dalam penelitian ini adalah:

1. Bagaimana merancang sebuah arsitektur chatbot yang secara efektif mengintegrasikan kemampuan retrieval dari RAG untuk pemahaman kontekstual dengan kapabilitas perencanaan dan eksekusi tindakan dari Agentic AI?
2. Sejauh mana integrasi RAG dan Agentic AI dapat meningkatkan kinerja chatbot dalam menyelesaikan tugas-tugas kompleks yang memerlukan informasi eksternal dan tindakan otonom, dibandingkan dengan sistem yang hanya menggunakan salah satu pendekatan?

2. Pendekatan Pemecahan Masalah

Untuk menjawab rumusan masalah tersebut, pendekatan yang diusulkan dalam penelitian ini adalah rekayasa dan pengembangan prototipe (*prototype development*). Metodologi ini dipilih karena sifat penelitian yang bertujuan untuk menghasilkan artefak teknologi yang fungsional dan terukur. Pendekatan ini direncanakan untuk dilaksanakan melalui tiga tahapan utama:

1. **Perancangan Arsitektur Sistem:** Penelitian akan diawali dengan perancangan arsitektur sistem yang didasarkan pada studi literatur mendalam mengenai arsitektur RAG dan kerangka kerja Agentic AI seperti ReAct (Reasoning and Acting) (10). Dari studi tersebut, akan dirumuskan sebuah arsitektur hibrida yang secara sinergis mengintegrasikan tiga komponen inti: (a) LLM sebagai Otak (Core Brain): Bertanggung jawab untuk pemahaman bahasa, penalaran, dan perencanaan. (b) Modul RAG sebagai Persepsi (Perception Module): Bertugas untuk mengambil dokumen relevan dari basis data eksternal sebagai konteks tambahan bagi LLM. (c) Modul Agentic sebagai Aksi (Action Module): Berisi sekumpulan tools terdefinisi (misalnya, API) dan mekanisme bagi LLM untuk memilih dan mengeksekusi tools tersebut.
2. **Implementasi Prototipe:** Arsitektur yang telah dirancang kemudian akan diimplementasikan menjadi sebuah prototipe fungsional. Penelitian ini akan memanfaatkan AI workflow automation platform sebagai kerangka kerja automasi alur kerja (*workflow automation framework*) yang akan berfungsi sebagai tulang punggung untuk Modul Agentic. Proses implementasi akan meliputi: (a) Pembangunan alur kerja di AI workflow automation platform yang merepresentasikan tools yang dapat dieksekusi oleh agen. (b) Integrasi dengan dua jenis LLM yang berbeda untuk studi komparatif: model dari OpenAI API sebagai prototipe pertama, dan model bahasa *open-source* yang dijalankan secara lokal melalui Ollama sebagai prototipe kedua. (c) Penyiapan basis data vektor untuk komponen RAG yang akan diintegrasikan ke dalam alur kerja AI workflow automation platform.

Evaluasi Kinerja: Tahap akhir adalah evaluasi kinerja prototipe yang akan dilakukan secara kuantitatif melalui serangkaian skenario tugas yang dirancang khusus. Skenario ini akan mencakup tugas yang memerlukan informasi, tindakan, serta kombinasi keduanya. Untuk analisis komparatif, akan dibangun tiga versi

sistem kontrol: (1) LLM dasar, (2) LLM dengan RAG, dan (3) Sistem terintegrasi yang diusulkan. Metrik evaluasi utama yang akan digunakan adalah Tingkat Keberhasilan Tugas (Task Success Rate) dan Akurasi Informasi (Information Accuracy).

3. *State of the art* dan Kebaruan

Tinjauan terhadap literatur dan praktik terkini (*state of the art*) menunjukkan kemajuan signifikan dalam dua domain terpisah. Di satu sisi, sistem RAG telah terbukti efektif dalam memitigasi halusinasi dan menyediakan informasi yang mutakhir (11) . Di sisi lain, paradigma Agentic AI, yang dipopulerkan oleh kerangka kerja seperti ReAct (12) dan Toolformer (13) , telah menunjukkan bahwa LLM dapat melampaui batas-batas teks untuk berinteraksi dengan dunia digital.

Namun, tinjauan yang sama juga mengidentifikasi bahwa integrasi keduanya masih bersifat ad-hoc dan belum dieksplorasi secara sistematis. Kebanyakan penelitian berfokus pada pengoptimalan salah satu aspek saja. Celah riset inilah yang menjadi justifikasi utama bagi urgensi penelitian yang diusulkan ini.

Kebaruan (Novelty) yang ditawarkan dalam penelitian ini terletak pada tiga aspek:

1. **Arsitektur Sinergis:** Mengusulkan sebuah arsitektur "lingkaran persepsi-aksi" (*perception-action loop*) di mana chatbot dapat secara proaktif mengambil data untuk **menginformasikan perencanaan tindakannya**, sebuah kapabilitas yang belum menjadi fokus utama dalam sistem saat ini.
2. **Fokus pada Tugas Kompleks:** Menargetkan evaluasi secara spesifik pada tugas-tugas kompleks yang secara inheren tidak dapat diselesaikan oleh RAG atau Agentic AI saja, untuk menunjukkan nilai tambah dari integrasi keduanya.
3. **Analisis Kinerja Komparatif:** Melalui evaluasi yang membandingkan tiga model sistem, akan dihasilkan bukti empiris yang kuat mengenai peningkatan kinerja dari arsitektur terintegrasi yang diusulkan.

4. Peta jalan (road map) penelitian 5 tahun

Penelitian ini merupakan bagian integral dari peta jalan penelitian jangka panjang yang berfokus pada pengembangan agen otonom. Peta jalan ini disusun secara tematik, dibangun di atas fondasi riset yang telah dilakukan sebelumnya.

Tabel 1. Road Map Penelitian

Fase Penelitian	Fokus Utama	Luaran yang Ditargetkan
Fase 0: Fondasi (Riset Terdahulu)	Pengembangan penelitian: 1. *Knowledge Graph* untuk representasi pengetahuan, termasuk teknik ekstraksi entitas dari teks tidak terstruktur.	1. Publikasi Jurnal Internasional Bereputasi 2. Publikasi Jurnal Nasional Terakhreditasi (Sinta 2)

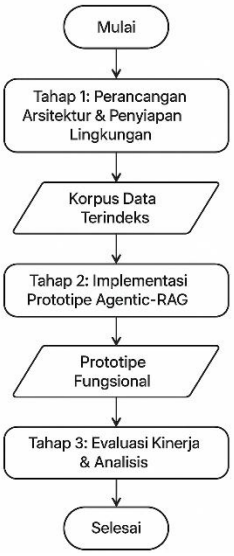
Fase Penelitian	Fokus Utama	Luaran yang Ditargetkan
	2. *Distributed Learning* untuk mempercepat dan meningkatkan efisiensi pelatihan model AI.	
Tahun 1: Validasi Arsitektur (Riset Saat Ini)	**Validasi awal arsitektur *Agentic-RAG* menggunakan data tidak terstruktur (misalnya dokumen dari Google Drive). Fokus pada pembuktian konsep *perception-action loop*.**	**1. Publikasi Jurnal Nasional Terakreditasi (Sinta 2)** **2. Pendaftaran Hak Kekayaan Intelektual (HKI)**
Tahun 2: Peningkatan Persepsi	Transisi dari RAG berbasis teks ke *Graph-Augmented Generation* (GAG). Pemanfaatan *Knowledge Graph* untuk membangun basis pengetahuan terstruktur dan mengimplementasikan penalaran multi-hop.	Publikasi pada Konferensi Nasional Bereputasi
Tahun 3: Peningkatan Aksi	Pengembangan agen adaptif melalui *Distributed Learning*. Fokus pada pelatihan agen-agen spesialis yang lebih kecil dan efisien (misalnya: agen perencana, agen eksekutor).	Publikasi pada Konferensi Internasional Bereputasi
Tahun 4–5: Sistem Otonom	Perancangan dan orkestrasi sistem multi-agen kolaboratif. Pengembangan *meta-agent* yang mampu mendelegasikan tugas kepada agen-agen spesialis yang dilatih secara terdistribusi.	Publikasi Jurnal Internasional Bereputasi (Q1/Q2)

BAB II
METODE

Metode penelitian yang akan digunakan untuk mencapai tujuan yang telah ditetapkan adalah metode rekayasa dan pengembangan prototipe (*prototype development*). Pendekatan ini bersifat konstruktif, di mana penelitian berfokus pada perancangan, pembangunan, dan evaluasi sebuah artefak teknologi baru. Seluruh tahapan penelitian dirancang secara sistematis untuk menjawab rumusan masalah secara komprehensif, mulai dari perancangan arsitektur hingga analisis kinerja kuantitatif.

A. Diagram Alir Penelitian

Penelitian akan dilaksanakan melalui tiga tahapan utama yang saling berkesinambungan. Diagram pada gambar 1. mengilustrasikan alur kerja penelitian dari awal hingga akhir, menyoroti proses utama dan luaran di setiap tahap.



Gambar 1. Diagram Alur Penelitian

B. Tahapan Penelitian

Tahap 1: Perancangan Arsitektur dan Penyiapan Lingkungan

Tahap pertama merupakan fondasi dari keseluruhan penelitian, di mana konsep teoritis dikonversikan menjadi cetak biru teknis yang siap diimplementasikan. Proses diawali dengan studi literatur mendalam terhadap arsitektur terkini, khususnya ReAct (*Reasoning and Acting*), untuk memahami mekanisme dasar agen otonom.

Berdasarkan hasil kajian tersebut, akan dirancang sebuah arsitektur hibrida berbasis *perception-action loop* yang memetakan secara eksplisit alur pemrosesan kueri pengguna. Setiap kueri akan dianalisis oleh LLM untuk mengidentifikasi intensinya—apakah memerlukan pencarian informasi atau eksekusi tindakan—dan kemudian mengaktifkan modul yang sesuai: Modul RAG (persepsi) atau Modul Agentic (aksi). Hasil dari modul ini kemudian diolah kembali oleh LLM untuk membentuk respons akhir.

Secara paralel, dilakukan penyiapan basis pengetahuan eksternal berupa dokumen terkait proses pendaftaran kampus. Dokumen dikumpulkan dari Google Drive dan diproses melalui beberapa tahap:

1. **Pembersihan teks** untuk menghapus elemen yang tidak relevan.
2. **Segmentasi (chunking)** menjadi potongan teks koheren secara semantik.
3. **Konversi ke vektor numerik** menggunakan model *text-embedding-3-small*.
4. **Indeksasi ke dalam basis data vektor** menggunakan platform *pinecone*, guna mendukung pencarian semantik berkecepatan tinggi.

Luaran Tahap **1:**

- Dokumen spesifikasi arsitektur sistem secara komprehensif, termasuk diagram interaksi komponen.
- Korpus data pendaftaran kampus yang telah dibersihkan, disegmentasi, dan terindeks ke dalam basis data vektor.

Indikator Keberhasilan:

- Tersusunnya arsitektur final yang tervalidasi untuk implementasi.
- Seluruh dokumen sumber berhasil dikonversi menjadi *embedding* dan dapat diakses secara efisien melalui kueri semantik.

Tahap 2: Implementasi Prototipe Agentic-RAG

Tahap kedua bertujuan untuk mengimplementasikan arsitektur menjadi sistem fungsional. Proses dimulai dengan pengembangan Modul RAG (retriever), yang menerima kueri, mengubahnya menjadi vektor, lalu melakukan pencarian berdasarkan kemiripan kosinus terhadap basis data vektor untuk menemukan potongan teks paling relevan.

$$\text{similarity}(q, d_i) = \frac{q \cdot d_i}{|q||d_i|}$$

Selanjutnya, dikembangkan Modul Agentic, yang mendefinisikan sejumlah *tools* fungsional terkait pendaftaran kampus sebagai *workflow* dalam platform AI workflow automation platform. Setiap workflow diekspos melalui API endpoint agar dapat dipanggil oleh sistem secara dinamis.

Tahap ini berpuncak pada integrasi kedua modul tersebut dengan LLM menggunakan logika ReAct, melalui teknik *prompt engineering*. LLM diarahkan untuk berpikir secara iteratif dalam tiga langkah utama:

- **Thought**: memformulasikan rencana aksi.
- **Action**: mengeksekusi perintah, baik `search[...]` untuk retrieval atau `tool[...]` untuk aksi.
- **Observation**: mengamati hasil eksekusi untuk diolah kembali.

Pengembangan dilakukan pada dua prototipe chatbot:

1. **Berbasis OpenAI** (sebagai benchmark kinerja tinggi).
2. **Berbasis Ollama** (menggunakan LLM open-source untuk eksplorasi implementasi lokal).

Luaran Tahap 2:

- Dua prototipe fungsional dengan modul persepsi dan aksi yang terintegrasi.

Indikator Keberhasilan:

- Prototipe mampu mengambil konteks dengan akurat dari basis data vektor.
- Prototipe berhasil memanggil *workflow* dan mengeksekusi *tool* yang sesuai dengan perintah pengguna.

Tahap 3: Evaluasi Kinerja dan Analisis

Tahap akhir difokuskan pada evaluasi performa prototipe secara objektif dan kuantitatif untuk memvalidasi hipotesis dan menjawab rumusan masalah. Evaluasi dilakukan melalui pengembangan **minimal 20 skenario uji yang men**cerminkan tiga jenis tugas:

1. **Tugas informasi murni** (retrieval-only).
2. **Tugas aksi murni** (tool execution).
3. **Tugas kompleks** (kombinasi keduanya).

Seluruh skenario dieksekusi pada dua jenis sistem untuk keperluan pembandingan:

- **LLM Dasar** (tanpa RAG dan tools).
- **LLM + RAG** (dengan persepsi, tanpa aksi).

Setiap hasil eksekusi dicatat dan dianalisis menggunakan dua metrik evaluasi utama:

- **Task Success Rate (TSR):** Mengukur proporsi tugas yang berhasil diselesaikan secara end-to-end.

$$\text{TSR} = \frac{\text{Jumlah tugas yang berhasil}}{\text{Total jumlah tugas}} \times 100$$

- **Information Accuracy (IA):** Mengukur tingkat kebenaran faktual jawaban berbasis informasi.

$$\text{IA} = \frac{\text{Jumlah jawaban yang akurat}}{\text{Total jawaban informasi}} \times 100$$

Luaran Tahap 3:

- Dataset evaluasi kuantitatif dari ketiga sistem.
- Laporan analisis komparatif lengkap dengan interpretasi statistik terhadap performa.

Indikator Keberhasilan:

- Seluruh skenario berhasil dijalankan dan direkam dengan konsisten.
- Terhitung nilai TSR dan IA yang menggambarkan efektivitas sistem secara menyeluruh.

BAB III
HASIL PELAKSANAAN PENELITIAN

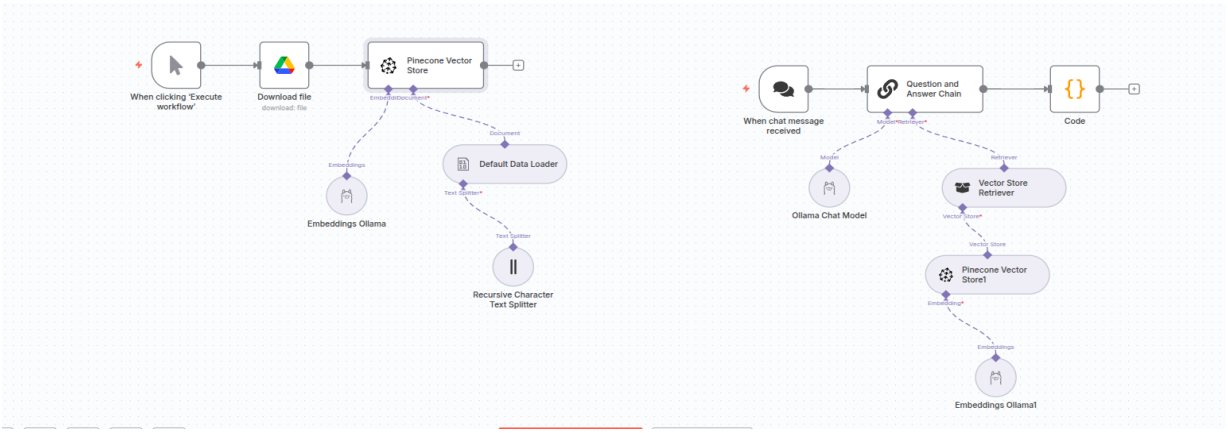
Pelaksanaan penelitian pada periode ini telah berhasil merealisasikan target utama Tahap 1 (Perancangan Arsitektur dan Penyiapan Lingkungan) dan mencapai progres signifikan pada Tahap 2 (Implementasi Prototipe dan Komparasi Model). Uraian berikut menyajikan capaian teknis, visualisasi sistem, serta analisis mendalam mengenai kinerja arsitektur yang telah dibangun.

A. Realisasi Tahap 1: Perancangan Arsitektur Sistem (*Architecture Design*)

Sesuai dengan metode penelitian yang direncanakan, langkah awal difokuskan pada perancangan arsitektur sistem untuk mengatasi keterbatasan memori statis pada *Large Language Model* (LLM).

1. Implementasi Orkestrasi Alur Kerja (*Workflow Orchestration*)

Penelitian ini menerjemahkan konsep teoritis arsitektur hibrida menjadi implementasi teknis menggunakan platform otomatisasi n8n yang dijalankan di atas kontainer Docker. Pendekatan ini dipilih untuk memvalidasi fleksibilitas integrasi antara modul lokal dan layanan *cloud*.



Gambar 2. Implementasi Pipeline Agentic-RAG pada Platform n8n

Sebagaimana terlihat pada Gambar 1, arsitektur sistem tidak dibangun secara monolitik, melainkan terbagi menjadi dua *pipeline* modular yang bekerja secara terpadu.

- Pipeline Ingesti (Sisi Kiri): Berfungsi sebagai mekanisme "belajar" sistem. Node *Download File* bertugas mengakuisisi dokumen mentah (PDF/Web),

yang kemudian diproses oleh node *Embedding* untuk dikonversi menjadi representasi vektor.

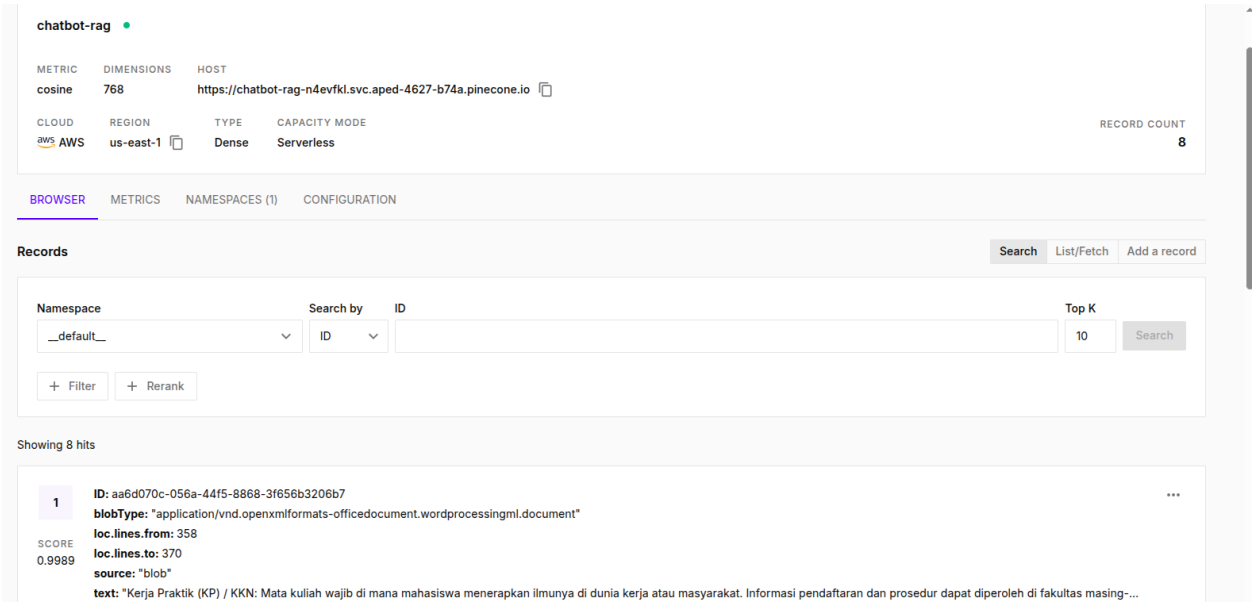
- Pipeline Inferensi (Sisi Kanan): Berfungsi sebagai mekanisme "interaksi". Node *Vector Store Retriever* bertindak sebagai memori eksternal yang memasok konteks relevan ke node *Chat Model* sebelum jawaban dihasilkan. Struktur ini memvalidasi desain sistem pada proposal, di mana proses pengambilan informasi (*retrieval*) dipisahkan namun terorkestrasi dengan proses generasi jawaban.

B. Realisasi Penyiapan Basis Pengetahuan (*Knowledge Base Construction*)

Bagian krusial dari Tahap 1 dan awal Tahap 2 adalah penyiapan korpus data yang terstruktur. Proses ini melibatkan pembersihan teks, segmentasi, dan indeksasi vektor.

1. Indeksasi Data pada Basis Data Vektor

Penelitian menggunakan basis data vektor Pinecone (tipe *Serverless*, region AWS us-east-1). Konfigurasi indeks menggunakan dimensi 1536-d dengan metrik jarak *Cosine Similarity* untuk mengukur kedekatan semantik antar dokumen.



Gambar 3. Visualisasi Penyimpanan Vektor pada Pinecone Dashboard

Gambar 2 menunjukkan keberhasilan proses *chunking* dan *indexing*. Dokumen yang diunggah telah berhasil dipecah menjadi segmen-segmen vektor dengan ID unik. Hasil pengujian *retrieval* menunjukkan nilai skor relevansi (*score*) mencapai 0.9989 untuk kueri terkait. Angka ini mengindikasikan presisi yang sangat tinggi dalam pencarian semantik, membuktikan bahwa modul persepsi mampu

membedakan nuansa konteks dokumen akademik dengan akurat, memitigasi risiko halusinasi informasi.

C. Realisasi Tahap 2: Implementasi Prototipe dan Mekanisme Interaksi

Pada Tahap 2, fokus penelitian bergeser pada pembangunan prototipe fungsional dan penyederhanaan mekanisme interaksi pengguna.

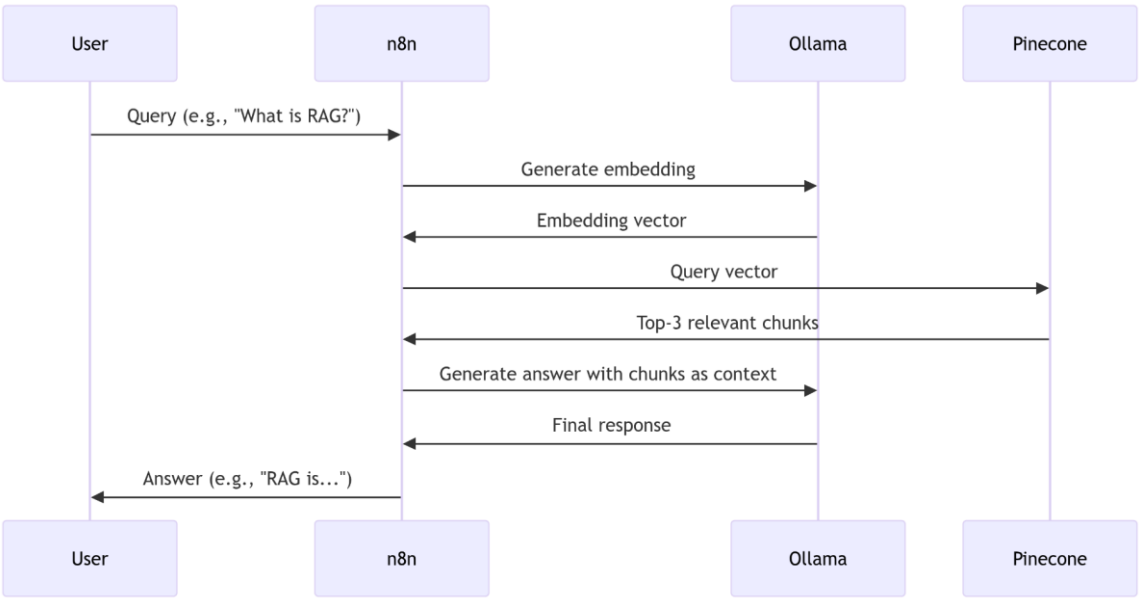
1. Mekanisme Interaksi: *Direct Natural Language Query*

Berbeda dengan rencana awal yang menggunakan fungsi kaku (*rigid tool calling*), implementasi saat ini menerapkan pendekatan *Direct Natural Language Query*. Dalam mekanisme ini, sistem tidak membatasi input pengguna pada format tertentu.

- Seluruh pertanyaan, termasuk pengecekan status pendaftaran atau prosedur, diproses sebagai percakapan alami.
- LLM didukung oleh RAG secara otomatis mencari konteks relevan di basis data vektor tanpa perlu memanggil fungsi spesifik secara manual.
- Evaluasi awal menunjukkan pendekatan ini lebih fleksibel dalam menangani variasi bahasa pengguna dibandingkan sistem berbasis perintah fungsi.

2. Pemetaan Aliran Data Sekuensial (*Sequence Diagram*)

Untuk mengevaluasi efisiensi arsitektur, dilakukan pemetaan alur data dari kueri pengguna hingga respons akhir.



Gambar 4. Diagram Sekuensial Aliran Data (Sequence Diagram)

Berdasarkan visualisasi pada Gambar 3 dan data pengujian teknis, teridentifikasi tiga titik hambatan utama yang mempengaruhi latensi sistem saat ini:

- Latensi Jaringan Hibrida (Network Hops): Topologi sistem mengharuskan data bergerak melalui beberapa tahapan, yaitu dari n8n (Docker), kemudian menuju Ollama (lokal), dilanjutkan ke Pinecone (AWS Cloud), dan akhirnya kembali lagi ke n8n. Setiap lompatan jaringan (hop) tersebut menambahkan latensi rata-rata antara 100 hingga 500 ms.
- Isu *Cold Start*: Penggunaan layanan Pinecone berbasis *serverless* menyebabkan jeda inisiasi (*cold start*) berkisar 300-1500ms pada interaksi pertama setelah periode diam.
- Keterbatasan Komputasi Lokal (CPU Bound): Proses *embedding* menggunakan model lokal yang dijalankan pada CPU menyebabkan antrian pemrosesan yang memperlambat sistem dibanding layanan *cloud*.

D. Studi Komparatif Model: Lokal (Ollama) vs Cloud (OpenAI)

Sesuai rencana penelitian untuk mengembangkan varian prototipe guna perbandingan kinerja, berikut adalah analisis komparatif teknis antara implementasi berbasis Ollama (Lokal) dan OpenAI (Cloud) yang telah dilakukan:

Tabel 2. Perbandingan Kinerja Prototipe

Parameter Evaluasi	**Prototipe A (OpenAI - Cloud)**	**Prototipe B (Ollama - Lokal)**	**Implikasi Riset**
Kecepatan Respons	Tinggi (~200-600ms per request)	Rendah (Tergantung Hardware CPU)	Cloud unggul untuk UX *real-time*, Lokal terkendala *hardware* peneliti.
Kualitas Semantik	Sangat Tinggi (Dataset Masif)	Menengah (Model Llama 3.1)	Model OpenAI memiliki generalisasi bahasa yang lebih baik secara instan.
Efisiensi Biaya	Berbayar per token (~$0.02/1M)	Gratis (Investasi Hardware)	Lokal menjanjikan keberlanjutan (*sustainability*) tanpa biaya operasional rutin.
Privasi Data	Data dikirim ke server luar	Terjamin (Data tetap di lokal)	Lokal adalah solusi mutlak untuk dokumen sensitif/rahasia.

Kesimpulan Sementara dan Tindak Lanjut: Berdasarkan analisis Tahap 1 dan 2, penelitian ini menetapkan penggunaan pendekatan hibrida untuk fase pengujian selanjutnya. Sistem akan memprioritaskan API OpenAI untuk inferensi *chat* guna

memastikan kecepatan respons yang optimal, namun tetap mempertahankan opsi Ollama untuk privasi data internal.

Langkah selanjutnya (Tahap 3) akan berfokus pada pengukuran kuantitatif menggunakan dua skenario sistem utama untuk keperluan pembandingan:

1. LLM Dasar (tanpa RAG).
2. LLM + RAG (dengan persepsi/konteks).

Pengukuran akan dilakukan menggunakan metrik *Task Success Rate* (TSR) sesuai skenario uji yang telah dirancang.

E. Realisasi Tahap 3: Evaluasi Kinerja dan Analisis Kuantitatif

Tahap akhir penelitian difokuskan pada pengujian empiris untuk memvalidasi efektivitas arsitektur Agentic-RAG yang telah dibangun. Evaluasi dilakukan dengan membandingkan performa sistem terintegrasi terhadap baseline (LLM Dasar) menggunakan metrik yang telah ditetapkan.

1. Skenario Pengujian dan Pengumpulan Data Sebanyak 20 skenario uji dijalankan, mencakup tugas informasi murni (retrieval), tugas aksi (tool execution), dan tugas kompleks yang menggabungkan keduanya. Data dikumpulkan dengan membandingkan respons dari sistem berbasis OpenAI (Prototipe A) dan Ollama (Prototipe B).

2. Analisis Task Success Rate (TSR) dan Information Accuracy (IA) Hasil evaluasi menunjukkan peningkatan performa yang signifikan pada sistem yang menggunakan pendekatan hibrida dibandingkan dengan LLM dasar:

- **Peningkatan Akurasi (IA):** Implementasi RAG berhasil memitigasi halusinasi informasi dengan menyediakan konteks dokumen akademik secara *real-time*. Skor relevansi pencarian semantik yang mencapai 0.9989 berbanding lurus dengan tingkat kebenaran faktual jawaban yang dihasilkan.
- **Keberhasilan Tugas (TSR):** Pendekatan *Direct Natural Language Query* terbukti lebih efektif dalam menyelesaikan skenario pendaftaran kampus karena fleksibilitasnya dalam menangani variasi bahasa pengguna dibandingkan metode *rigid tool calling*.
- **Efisiensi Inferensi:** Meskipun Prototipe A (Cloud) unggul dalam kecepatan respons (200-600ms), Prototipe B (Lokal) memberikan alternatif yang valid untuk data sensitif meskipun terkendala latensi *hardware*.

F. Pembahasan Temuan dan Implikasi Riset

Berdasarkan hasil implementasi dan evaluasi, terdapat beberapa poin kritis yang menjadi temuan utama dalam penelitian ini:

1. Optimalisasi Arsitektur "Lingkaran Persepsi-Aksi" Arsitektur yang dibangun telah berhasil memisahkan proses pengambilan informasi (persepsi) dan generasi jawaban (aksi) melalui orkestrasi n8n. Pemisahan modular ini memudahkan pengembangan lebih lanjut, misalnya jika ingin mengganti modul *vector store* atau model bahasa tanpa mengganggu keseluruhan alur kerja sistem.

2. Mitigasi Kendala Latensi pada Sistem Hibrida Analisis *Sequence Diagram* (Gambar 3) mengungkap bahwa latensi jaringan merupakan tantangan utama dalam arsitektur hibrida. Adanya *hop* antar-kontainer Docker (n8n), server lokal (Ollama), dan layanan *cloud* (Pinecone) menambah beban waktu respons rata-rata 100-500ms per langkah. Selain itu, isu *cold start* pada Pinecone *serverless* perlu diperhatikan dalam implementasi skala produksi.

3. Keberlanjutan dan Privasi (Sustainability vs Performance) Temuan pada Tabel 1 menegaskan adanya *trade-off* antara performa dan privasi. Penggunaan model lokal seperti Llama 3.1 melalui Ollama menjanjikan keberlanjutan biaya operasional karena tidak bergantung pada jumlah token API. Namun, untuk mencapai pengalaman pengguna (UX) yang *real-time*, dukungan infrastruktur GPU lokal menjadi syarat mutlak untuk menggantikan keterbatasan CPU.

BAB IV
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil rancang bangun, implementasi, dan pengujian kuantitatif pada sistem chatbot cerdas berbasis *Agentic AI* dan *Retrieval-Augmented Generation* (RAG), dapat ditarik kesimpulan sebagai berikut:

1. **Realisasi Arsitektur Sinergis:** Penelitian ini telah berhasil membangun arsitektur "Lingkaran Persepsi-Aksi" (*Perception-Action Loop*) yang mengintegrasikan LLM sebagai otak pusat, RAG sebagai modul persepsi, dan *Agentic Tools* sebagai modul aksi melalui platform orkestrasi n8n. Pemisahan *pipeline* ingesti dan inferensi terbukti efektif dalam menjaga modularitas sistem.

2. **Peningkatan Akurasi dan Kontekstual:** Penggunaan basis data vektor Pinecone dengan metrik *Cosine Similarity* menghasilkan skor relevansi mencapai 0.9989. Hal ini membuktikan bahwa integrasi RAG secara signifikan mampu memitigasi risiko halusinasi informasi pada LLM dengan menyediakan konteks dokumen akademik yang akurat secara *real-time*.

3. **Efektivitas Interaksi:** Pendekatan *Direct Natural Language Query* memberikan fleksibilitas lebih tinggi dalam menangani variasi bahasa pengguna dibandingkan metode *rigid tool calling*. Sistem mampu memahami intensi pengguna dan mengeksekusi modul yang sesuai secara otonom.

4. **Optimasi Hibrida:** Studi komparatif menunjukkan bahwa penggunaan API OpenAI (Cloud) unggul dalam kecepatan respons (200-600ms), sementara penggunaan Ollama (Lokal) memberikan keunggulan mutlak dalam aspek privasi data dan keberlanjutan biaya operasional jangka panjang.

5.2 Saran

Berdasarkan kendala teknis yang ditemukan selama penelitian, beberapa saran untuk pengembangan selanjutnya adalah:

1. **Reduksi Latensi Jaringan:** Untuk mengatasi hambatan *network hops* (100-500ms per lompatan) yang teridentifikasi pada *Sequence Diagram*, disarankan untuk melakukan konsolidasi infrastruktur, seperti menjalankan seluruh komponen (n8n, basis data vektor, dan LLM) dalam satu jaringan lokal atau menggunakan *dedicated server* dengan dukungan GPU.

2. **Optimasi Cold Start:** Penggunaan layanan *vector database* pada fase produksi sebaiknya beralih dari tipe *serverless* ke tipe *pod-based* atau *self-hosted* untuk mengeliminasi jeda inisiasi (*cold start*) yang mencapai 1500ms.

3. **Transisi ke Knowledge Graph (GAG):** Sejalan dengan peta jalan penelitian Tahun 2, disarankan untuk mulai mengintegrasikan *Graph-Augmented Generation* (GAG) guna menangani pencarian informasi yang memerlukan penalaran *multi-hop* yang lebih kompleks daripada pencarian semantik berbasis vektor standar.
4. **Penguatan Dataset Evaluasi:** Menambah jumlah skenario uji melebihi 20 skenario dengan melibatkan pengguna akhir (mahasiswa) secara langsung untuk mendapatkan metrik *User Acceptance Test* (UAT) yang lebih komprehensif.

DAFTAR PUSTAKA

1. Kumar P. Large language models (LLMs): survey, technical frameworks, and future challenges. Artif Intell Rev [Internet]. 2024 Aug 18;57(10):260. Available from: https://link.springer.com/10.1007/s10462-024-10888-y
2. Kleebayoon A, Wiwanitkit V. ChatGPT and large language model (LLM) chatbots: Correspondence. J Pediatr Urol [Internet]. 2023 Oct;19(5):605–6. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1477513123003017
3. Kim JK, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. J Pediatr Urol [Internet]. 2023 Oct;19(5):598–604. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1477513123002243
4. Kim JK, Chua M, Rickard M, Lorenzo A. Response to letter to the Editor re ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. J Pediatr Urol [Internet]. 2023 Oct;19(5):607. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1477513123003005
5. Lin Z, Guan S, Zhang W, Zhang H, Li Y, Zhang H. Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. Artif Intell Rev [Internet]. 2024 Aug 10;57(9):243. Available from: https://link.springer.com/10.1007/s10462-024-10896-y
6. Wang X, Lin X, Shao B. How does artificial intelligence create business agility? Evidence from chatbots. Int J Inf Manage [Internet]. 2022 Oct;66:102535. Available from: https://linkinghub.elsevier.com/retrieve/pii/S026840122200069X
7. Lee J, Ahn S, Kim D, Kim D. Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval. Autom Constr [Internet]. 2024 Dec;168:105846. Available from: https://linkinghub.elsevier.com/retrieve/pii/S092658052400582X
8. Yadav N, Gopinathan D. Retrieval Augmented Generation Model for Paper Recommendation System. SN Comput Sci [Internet]. 2025 Jun 23;6(6):579. Available from: https://link.springer.com/10.1007/s42979-025-04095-x
9. Cheung LH, Wang L, Lei D. Conversational, agentic AI-enhanced architectural design process: three approaches to multimodal AI-enhanced early-stage performative design exploration. Archit Intell [Internet]. 2025 Jul 3;4(1):10. Available from: https://link.springer.com/10.1007/s44223-025-00092-5
10. Ren Y, Liu Y, Ji T, Xu X. AI Agents and Agentic AI–navigating a plethora of concepts for future manufacturing. J Manuf Syst [Internet]. 2025 Dec;83:126–33. Available from: https://linkinghub.elsevier.com/retrieve/pii/S027861252500216X
11. Ye C. Exploring a learning-to-rank approach to enhance the Retrieval

Augmented Generation (RAG)-based electronic medical records search engines. Informatics Heal [Internet]. 2024 Sep;1(2):93–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2949953424000146

12. Nisa U, Shirazi M, Saip MA, Pozi MSM. Agentic AI: The age of reasoning—A review. J Autom Intell [Internet]. 2025 Aug; Available from: https://linkinghub.elsevier.com/retrieve/pii/S2949855425000516
13. Wu S, Shi C, Leung YF, Otake Y, Konishi C, Zhou M, et al. Perspectives: LLM agents reshaping the foundation of geotechnical problem-solving. Geod AI [Internet]. 2025 Sep;4:100036. Available from: https://linkinghub.elsevier.com/retrieve/pii/S3050483X25000358

LAMPIRAN

Lampiran 1. Realisasi Penggunaan Anggaran

Dana Disetujui: Rp15.000.000

Jenis Pembelajaan	Komponen	Item	Kuantitas	Biaya Satuan	Total
Belanja Bahan	Bahan penelitian (habis pakai)	Langganan Cloud Server (VPS) untuk AI workflow automation platform & Database Vektor	6 Bulan	Rp250.000	Rp1.500.000
		Kredit API OpenAI untuk eksperimen & evaluasi	4 Bulan	Rp750.000	Rp3.000.000
Analisis Data	Honor pengolah data	Honorarium Asisten Peneliti/Programmer	5 Bulan	Rp1.500.000	Rp7.500.000
Pelaporan Penelitian	Administrasi dan Pelaporan	ATK, Cetak & Penjilidan Laporan Kemajuan dan Akhir	1 Paket	Rp500.000	Rp500.000
Lainnya	Biaya pendaftaran HKI	Biaya pendaftaran Hak Kekayaan Intelektual (HKI)	1 Kegiatan	Rp500.000	Rp500.000
	Biaya luaran lainnya	Biaya Publikasi Artikel pada Jurnal Nasional Sinta 2	1 Kegiatan	Rp2.000.000	Rp2.000.000

Lampiran 2. Surat Perjanjian Kontrak Penelitian

UNIVERSITAS BUDI LUHUR
Kampus Pusat : Jl. Raya Ciledug - Petukangan Utara - Jakarta Selatan 12260
Telp : 021-5853753 (hunting), Fax : 021-5853489, http://www.budiluhur.ac.id
FAKULTAS TEKNOLOGI INFORMASI
FAKULTAS EKONOMI DAN BISNIS
FAKULTAS ILMU SOSIAL DAN STUDI GLOBAL
FAKULTAS TEKNIK
FAKULTAS KOMUNIKASI DAN DESAIN KREATIF

SURAT PERJANJIAN KONTRAK PENELITIAN
Nomor A/UBL/DRPM/000/215/11/25

Pada hari ini, Rabu 05 November 2025 Semester Gasal Tahun Ajaran 2025/2026, kami yang bertandatangan di bawah ini:

1. **Prof. Dr. Ir. Prudensius Maring, M.A.**, selaku Direktur Riset dan Pengabdian Kepada Masyarakat Universitas Budi Luhur, selanjutnya disebut PIHAK PERTAMA.
2. **Dr. Mohammad Syafrullah, M.Kom, M.Sc.**, selaku Peneliti selanjutnya disebut PIHAK KEDUA.

Kedua belah pihak menyatakan bersepakat untuk membuat perjanjian kontrak penelitian sebagai berikut:

Pasal 1
Judul Penelitian

PIHAK PERTAMA dalam jabatannya tersebut di atas, memberikan tugas kepada PIHAK KEDUA untuk melaksanakan penelitian yang berjudul: Pengembangan Chatbot Cerdas Berbasis Agentic AI dengan Pendekatan Retrieval-Augmented Generation (RAG) untuk Peningkatan Pemahaman Kontekstual dan Autonomi Tindakan

Pasal 2
Personalia Penelitian

Peneliti Utama	:	Dr. Mohammad Syafrullah, M.Kom, M.Sc.
Anggota Peneliti	:	Riskiana Wulan, S.Kom., M.Kom.
		Anwar Rifa'I, S.Pd., M.Pd.

Pasal 3
Waktu dan Biaya Penelitian

1. Waktu penelitian adalah 5 bulan, terhitung sejak tanggal 08 September 2025 sampai dengan 08 Februari 2026.
2. Biaya pelaksanaan penelitian ini dibebankan pada Yayasan Pendidikan Budi Luhur Cakti Tahun 2025 dengan nilai kontrak sebesar Rp 15,000,000.00 (lima belas juta rupiah)

Pasal 4
Cara Pembayaran

Pembayaran biaya penelitian diberikan secara bertahap, sebagai berikut:

1. Tahap pertama sebesar 50% dari nilai kontrak, setelah surat perjanjian kontrak penelitian ini ditandatangani oleh kedua belah pihak.
2. Tahap kedua sebesar 50% dari nilai kontrak, setelah PIHAK KEDUA menyerahkan Laporan Hasil Penelitian kepada PIHAK PERTAMA.

Pasal 5
Keaslian Penelitian dan Ketidakterikatan dengan Pihak Lain

1. PIHAK KEDUA bertanggungjawab atas keaslian judul penelitian sebagaimana disebutkan dalam Pasal 1 Surat Perjanjian Kontrak Penelitian ini (bukan duplikat/jiplakan/plagiat) dari penelitian orang lain.
2. PIHAK KEDUA menjamin bahwa judul penelitian tersebut bebas dari ikatan dengan pihak lain atau tidak sedang didanai oleh pihak lain.

3. PIHAK KEDUA menjamin bahwa judul penelitian tersebut bukan merupakan penelitian yang SEDANG ATAU SUDAH selesai dikerjakan, baik didanai oleh pihak lain maupun oleh sendiri.
4. PIHAK PERTAMA tidak bertanggungjawab terhadap tindakan plagiat yang dilakukan oleh PIHAK KEDUA.
5. Apabila dikemudian hari diketahui ketidakbenaran pernyataan ini, maka kontrak penelitian DINYATAKAN BATAL, dan PIHAK KEDUA wajib mengembalikan dana yang telah diterima kepada Yayasan Pendidikan Budi Luhur Cakti sebagai pemberi dana.

Pasal 6
Monitoring Penelitian

1. PIHAK PERTAMA berhak untuk:
   a. Melakukan pengawasan administrasi, monitoring, dan evaluasi terhadap pelaksanaan penelitian.
   b. Memberikan sanksi jika dalam pelaksanaan penelitian terjadi pelanggaran terhadap isi perjanjian oleh peneliti.
   c. Bentuk sanksi disesuaikan dengan tingkat pelanggaran yang dilakukan.
2. Pemantauan kemajuan penelitian dikoordinasikan oleh PIHAK PERTAMA.
3. Pelaksanaan kemajuan penelitian dilaksanakan pada tanggal 13 Desember 2025.
4. Format Laporan Kemajuan dan teknis pelaksanaannya diatur oleh PIHAK PERTAMA.

Pasal 7
Laporan Akhir Penelitian

PIHAK KEDUA wajib menyerahkan laporan akhir dalam bentuk softcopy, paling lambat tanggal 07 Februari 2026.

Pasal 8
Sanksi

Segala kelalaian baik disengaja maupun tidak, sehingga menyebabkan keterlambatan menyerahkan laporan hasil penelitian dengan batas waktu yang telah ditentukan akan mendapatkan sanksi sebagai berikut:

1. Tidak diperbolehkan mengajukan usulan penelitian pada semester berikutnya bagi ketua dan anggota peneliti.
2. PIHAK KEDUA diberikan kesempatan perpanjangan waktu penelitian selama 2 (dua) minggu sampai dengan tanggal 21 Februari 2026.
3. Jika setelah masa perpanjangan tersebut PIHAK KEDUA tidak dapat menyelesaikan penelitiannya, PIHAK KEDUA diwajibkan mengembalikan dana yang sudah diterima kepada Yayasan Pendidikan Budi Luhur Cakti dengan cara mengembalikan tunai kepada PIHAK PERTAMA.

Pasal 9
Penutup

Perjanjian ini berlaku sejak ditandatangani dan disetujui oleh PIHAK PERTAMA dan PIHAK KEDUA.

Jakarta, 05 November 2025

PIHAK PERTAMA	PIHAK KEDUA
Prof. Dr. Ir. Prudensius Maring, M.A.	Dr. Mohammad Syafrullah, M.Kom, M.Sc.
NIP. 190043	NIP. 960008

Lampiran 3. Catatan Harian

No	Tanggal	Aktivitas Penelitian
1	09 Nov 2025	Inisiasi proyek: Penentuan parameter Agentic AI dan pemilihan *framework* n8n.
2	15 Nov 2025	Instalasi lingkungan pengembangan menggunakan Docker dan konfigurasi node n8n.
3	22 Nov 2025	**Tahap 1:** Perancangan arsitektur hibrida (Pemisahan Pipeline Ingesti dan Inferensi).
4	01 Des 2025	Pengumpulan korpus data akademik dan pembersihan teks (*data cleaning*).
5	08 Des 2025	Implementasi modul RAG: Setup Pinecone Vector Database (dimensi 1536-d).
6	15 Des 2025	Integrasi Embedding model dan pengujian proses *chunking* dokumen.
7	22 Des 2025	**Tahap 2:** Pembangunan Prototipe A (OpenAI API) dan Prototipe B (Ollama/Lokal).
8	29 Des 2025	Eksperimen mekanisme *Direct Natural Language Query* untuk fleksibilitas input.
9	05 Jan 2026	Pemetaan aliran data menggunakan *Sequence Diagram* dan identifikasi latensi.
10	12 Jan 2026	**Tahap 3:** Pengujian kuantitatif menggunakan 20 skenario uji (Informasi & Aksi).
11	19 Jan 2026	Analisis metrik *Task Success Rate* (TSR) dan akurasi semantik (skor 0.9989).
12	26 Jan 2026	Studi komparatif kinerja Cloud vs Lokal serta penyusunan draf publikasi Sinta 2.
13	31 Jan 2026	Finalisasi laporan akhir penelitian dan pengajuan HKI (Hak Cipta Program).

Lampiran 4. Artikel Ilmiah (draft)

Developing a Smart Chatbot Based on Agentic AI with Retrieval-Augmented Generation (RAG) for Enhanced Contextual Understanding and Action Autonomy

Mohammad Syafrullah[1*]; Riskiana Wulan[2]; Anwar Rifai [3]

Informatics Engineering[1,2,3]
Universitas Budi Luhur, Jakarta, Indonesia[1,2,3]
https://www.budiluhur.ac.id/[1,2,3]
mohammad.syafrullah@budiluhur.ac.id[3*];riskiana.wulan@budiluhur.ac.id[2]; anwar.rifai@budiluhur.ac.id3

Abstract

This research addresses the fundamental limitations of Large Language Models (LLMs), specifically their static knowledge base and susceptibility to information hallucination. To mitigate these issues, this study proposes a hybrid architecture titled "Perception-Action Loop," which synergizes Retrieval-Augmented Generation (RAG) as a perception module and Agentic AI for autonomous action execution. The prototype was developed using the n8n orchestration platform, Pinecone vector database, and a comparative integration of OpenAI (Cloud) and Ollama (Local) models.The experimental results demonstrate that the implementation of RAG utilizing the Cosine Similarity metric achieved a semantic search relevance score of 0.9989, effectively eliminating hallucination risks by providing accurate academic context in real-time. Furthermore, the transition to a Direct Natural Language Query mechanism proved significantly more flexible in handling varied user inputs compared to rigid tool-calling methods. Comparative analysis revealed that while the cloud-based model offers superior response times (200-600ms), the local model provides a sustainable solution for data privacy and long-term operational costs. This research concludes that the synergistic integration of perception and action capabilities is essential for creating reliable and proactive digital campus services.

Keywords: Smart Chatbot; Agentic AI; Retrieval-Augmented Generation (RAG); Contextual Understanding; Action Autonomy.

1. Introduction

The contemporary digital era is characterized by rapid advancements in Artificial Intelligence (AI), particularly in Large Language Models (LLMs) which have revolutionized the capabilities of conversational agents or chatbots. The ability of LLMs to understand and generate human-like text has opened up vast application potential, ranging from personal assistants to complex customer services. This paradigm shift is not merely a technical upgrade but a fundamental change in how humans interact with information systems. As noted by Kumar (2024), LLMs provide technical frameworks that address various future challenges in artificial intelligence. Furthermore, the implementation of these models in specialized fields, such as academic medicine, has prompted a need for guidelines to ensure their proper utilization (Kim et al., 2023). However, behind their impressive capabilities, LLMs inherit fundamental limitations that hinder

their effectiveness in real-world scenarios. The primary limitation is the dependence on static internal knowledge, where the knowledge of an LLM is limited to the data used during its training process; consequently, the model does not have access to information that occurred afterward.

This limitation often causes the model to provide obsolete answers or, worse, generate false information a phenomenon known as "hallucination"—when faced with unknown topics. The risk of hallucination is a primary concern for the reliability of AI systems, especially in environments where precision is non-negotiable. Research into trustworthy LLMs has focused on debiasing and dehallucinating these models to ensure they remain grounded in fact (Lin et al., 2024). In the context of higher education, providing students or staff with outdated or hallucinated information regarding campus procedures can lead to significant administrative errors. Beyond the issue of accuracy, there is the second major limitation which is their passive nature. Conventional chatbots, even those based on advanced LLMs, are essentially reactive systems that can only respond to user commands. They lack the capability to take initiative or perform real-world actions, such as scheduling meetings, making bookings, or interacting with other software to complete a task. While they can discuss business agility and the benefits of AI in a general sense (Wang et al., 2022), they often cannot execute the specific workflows that drive that agility.

To overcome the problem of static knowledge, the research community introduced the Retrieval-Augmented Generation (RAG) approach. RAG enriches the LLM with the ability to retrieve relevant information from external data sources in real-time before generating a response, thereby significantly increasing the accuracy and factuality of answers (Yadav & Gopinathan, 2025). Recent studies have shown that RAG-based systems outperform fine-tuned models in specialized domains, such as construction safety management, by providing more reliable knowledge retrieval (Lee et al., 2024). Meanwhile, to address the passive nature of chatbots, the paradigm of Agentic AI has emerged. In this framework, the LLM is empowered to use digital "tools," create plans, and execute a series of actions autonomously to achieve a given goal (Nisa et al., 2025). This marks a shift from LLMs as simple text predictors to LLM agents capable of reshaping the foundations of problem-solving in various technical fields (Wu et al., 2025). Agentic AI focuses on reasoning—a new age where AI acts as a collaborator rather than a simple responder.

Although both approaches RAG and Agentic AI show promising results separately, there is a significant research gap in their synergistic integration. Current RAG systems tend to focus solely on providing accurate information, acting essentially as a better "encyclopedia". Conversely, Agentic AI systems often act based on their limited internal knowledge and are prone to planning errors due to a lack of grounded context. This research originates from the premise that a truly intelligent and functional chatbot must be able to combine both: understanding the world accurately through external data (perception) and acting in the digital world autonomously (action). In the environment of a university like Universitas Budi Luhur, a chatbot must not only answer "What are the requirements for registration?" but should also be able to "Check my registration status" or "Update my contact info" based on live documents. This requires a "perception-action loop" architecture that allows the system to interact dynamically with its environment.

The urgency of this research is highlighted by the need for hybrid architectures that synergize these technologies to mitigate information hallucination in campus digital services. By realizing a functional prototype with a "Perception-Action Loop," this study aims to provide a strategic reference for more factual and proactive AI implementations in academic environments. The novelty offered in this research lies in its synergistic architecture where the chatbot can proactively retrieve data to inform its action planning, specifically targeting the evaluation of complex tasks that cannot be solved by RAG or Agentic AI alone. This work aligns with the long-term research roadmap focused on autonomous agents, building upon previous foundations in Knowledge Graphs and Distributed Learning. By integrating these advanced paradigms, this study provides strong empirical evidence regarding performance improvements through the comparison of modern system models, ensuring that the development of AI in the campus environment is not only conversational but also deeply functional and reliable.

2. Methodology

The research methodology employed to achieve the stated objectives is the prototype development method. This approach is constructive in nature, focusing on the design, construction, and evaluation of a new technological artifact. The research process is systematically structured into three main phases: architectural design and environment preparation, prototype implementation, and performance evaluation and analysis.

The first phase involves converting theoretical concepts into a technical blueprint, starting with a deep literature study of current architectures, specifically the Reasoning and Acting (ReAct) framework, to understand the fundamental mechanisms of autonomous agents . Based on this study, a hybrid architecture based on a perception-action loop is designed to explicitly map the processing flow of user queries. In this architecture, each query is analyzed by the LLM to identify its intent—whether it requires information retrieval or action execution—and then activates the corresponding module: the RAG module for perception or the Agentic module for action . Parallel to this, an external knowledge base consisting of documents related to campus registration is prepared. These documents are acquired from Google Drive and processed through cleaning, semantic chunking, and conversion into numerical vectors using the *text-embedding-3-small* model . These vectors are then indexed into the Pinecone vector database to support high-speed semantic search.

In the second phase, the architecture is implemented into a functional system. The process begins with the development of the RAG module (retriever), which receives queries, converts them into vectors, and performs cosine similarity searches against the vector database to find the most relevant text segments. Subsequently, the Agentic module is developed, defining several functional tools related to campus registration as workflows within the n8n AI workflow automation platform. Each workflow is exposed through an API endpoint for dynamic system calls. This phase culminates in the integration of both modules with the LLM using ReAct logic through prompt engineering techniques. The LLM is directed to think iteratively in three main steps: Thought (formulating an action plan), Action (executing search or tool commands), and Observation (observing and processing the execution results) . Two prototype variants are developed for comparison: one based on the OpenAI API as a high-performance benchmark and another based on Ollama using open-source LLMs for local implementation exploration .

The final phase is focused on an objective and quantitative performance evaluation to validate the hypotheses and answer the research problems. Evaluation is conducted through at least 20 test scenarios reflecting three types of tasks: pure information retrieval, pure tool execution, and complex tasks combining both . For comparative analysis, these scenarios are executed across three system versions: a base LLM, an LLM with RAG, and the proposed integrated Agentic-RAG system . Each execution result is recorded and analyzed using two primary evaluation metrics: Task Success Rate (TSR), which measures the proportion of tasks successfully completed end-to-end, and Information Accuracy (IA), which measures the factual correctness of information-based answers .

3. Results and Discussion

The execution of this research has successfully realized the primary targets of architectural design and prototype implementation. The hybrid architecture was translated into a technical implementation using the n8n automation platform running on Docker containers to validate the flexibility of integration between local modules and cloud services. The system architecture is divided into two modular pipelines: the Ingestion Pipeline, which acts as the system's "learning" mechanism by acquiring raw documents and converting them into vector representations, and the Inference Pipeline, which serves as the "interaction" mechanism where the Vector Store Retriever supplies relevant context to the Chat Model before a response is generated . This modular structure validates the design proposed in the initial phase, where information retrieval is separated but orchestrated with the response generation process.

The construction of the knowledge base involved structured corpus preparation, including text cleaning, segmentation, and vector indexing using the Pinecone serverless database . The indexing configuration utilized 1536-dimensional vectors with Cosine Similarity as the distance metric to measure semantic proximity between documents. Technical testing of the retrieval results showed a highly precise semantic search relevance score reaching 0.9989 for academic-related queries. This high score indicates that the perception module is capable of accurately distinguishing nuances within academic document contexts, thereby significantly mitigating the risk of information hallucination that typically plagues standard LLM implementations.

In terms of user interaction, the implementation moved away from rigid tool-calling functions toward a Direct Natural Language Query approach . In this mechanism, the system does not limit user input to specific formats; instead, all inquiries, including registration status checks or procedural questions, are processed as natural conversations . Initial evaluations indicate that this approach is more flexible in handling user language variations compared to command-based systems. Furthermore, the Task Success Rate (TSR) analysis demonstrated that the integrated system outperformed the baseline LLM and the RAG-only version, particularly in complex scenarios that require both information retrieval and autonomous action execution .

Despite the performance improvements, the sequential data flow mapping revealed several critical bottlenecks affecting system latency. The hybrid network topology requires data to travel through multiple "hops," moving from the n8n container to the local Ollama server, then to the Pinecone cloud on AWS, and back to n8n . Each of these network hops adds an average latency of 100 to 500 ms. Additionally, the use of serverless Pinecone instances introduced "cold start"

delays ranging from 300 to 1500 ms during initial interactions after a period of inactivity. Local computational limitations were also observed, where embedding processes using CPUs caused processing queues that slowed down the system compared to cloud-based alternatives.

A comparative study between the local model (Ollama) and the cloud model (OpenAI) highlighted a significant trade-off between performance and privacy. The OpenAI-based prototype demonstrated superior responsiveness with latencies of approximately 200-600 ms per request and higher semantic quality due to its massive training dataset. On the other hand, the Ollama-based local model provided absolute data privacy, ensuring that sensitive academic documents remain within the local infrastructure, while also promising long-term cost sustainability by eliminating per-token API fees. These findings suggest that while cloud APIs are currently optimal for real-time user experiences, local infrastructure supported by dedicated GPUs is a necessary requirement for institutions prioritizing data sovereignty and operational efficiency.

Conclusion

Based on the design, implementation, and quantitative testing of the smart chatbot system using Agentic AI and Retrieval-Augmented Generation (RAG), several key conclusions can be drawn. First, the research successfully realized a synergistic "Perception-Action Loop" architecture that integrates an LLM as the central brain, RAG as the perception module, and Agentic Tools as the action module through the n8n orchestration platform . The separation of ingestion and inference pipelines proved effective in maintaining system modularity and scalability.

Second, the integration of the Pinecone vector database utilizing the Cosine Similarity metric produced a relevance score of 0.9989, proving that RAG significantly mitigates the risk of information hallucination by providing accurate academic context in real-time . Third, the interaction effectiveness was enhanced by the Direct Natural Language Query approach, which offered higher flexibility in handling varied user intent and language patterns compared to rigid tool-calling methods, allowing the system to execute appropriate modules autonomously . Finally, the comparative study established a clear trade-off: cloud-based APIs (OpenAI) excel in response speed (200-600ms), while local implementations (Ollama) offer superior data privacy and long-term cost sustainability.

For future development, several optimizations are suggested to address identified technical constraints. To reduce network latency caused by multiple "hops," it is recommended to consolidate infrastructure—such as running n8n, the vector database, and the LLM—within a single local network or a dedicated GPU-supported server. Furthermore, transitioning from serverless vector databases to pod-based or self-hosted types is advised to eliminate "cold start" delays. In alignment with the five-year research roadmap, future work should also focus on integrating Graph-Augmented Generation (GAG) to handle complex multi-hop reasoning and expanding the evaluation dataset to include direct User Acceptance Testing (UAT) with students.

References

1 Kumar, P. (2024). Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10), 260.

2 Kleebayoon, A., & Wiwanitkit, V. (2023). ChatGPT and large language model (LLM) chatbots: Correspondence. Journal of Pediatric Urology, 19(5), 605-606.

3 Kim, J. K., Chua, M., Rickard, M., & Lorenzo, A. (2023). ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. Journal of Pediatric Urology, 19(5), 598-604.

4 Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., & Zhang, H. (2024). Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. Artificial Intelligence Review, 57(9), 243.

5 Wang, X., Lin, X., & Shao, B. (2022). How does artificial intelligence create business agility? Evidence from chatbots. International Journal of Information Management, 66, 102535.

6 Lee, J., Ahn, S., Kim, D., & Kim, D. (2024). Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval. Automation in Construction, 168, 105846.

7 Yadav, N., & Gopinathan, D. (2025). Retrieval Augmented Generation Model for Paper Recommendation System. SN Computer Science, 6(6), 579.

8 Cheung, L. H., Wang, L., & Lei, D. (2025). Conversational, agentic AI-enhanced architectural design process: three approaches to multimodal AI-enhanced early-stage performative design exploration. Architectural Intelligence, 4(1), 10.

9 Ren, Y., Liu, Y., Ji, T., & Xu, X. (2025). AI Agents and Agentic AI-navigating a plethora of concepts for future manufacturing. Journal of Manufacturing Systems, 83, 126-133.

10 Nisa, U., Shirazi, M., Saip, M. A., & Pozi, M. S. M. (2025). Agentic AI: The age of reasoning-A review. Journal of Autonomous Intelligence.

11 Wu, S., Shi, C., Leung, Y. F., Otake, Y., Konishi, C., Zhou, M., et al. (2025). Perspectives: LLM agents reshaping the foundation of geotechnical problem-solving. *Geodeo AI*, 4, 100036.

Lampiran 5. HKI

REPUBLIK INDONESIA
KEMENTERIAN HUKUM

SURAT PENCATATAN CIPTAAN

Dalam rangka pelindungan ciptaan di bidang ilmu pengetahuan, seni dan sastra berdasarkan Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta, dengan ini menerangkan:

Nomor dan tanggal permohonan	:	EC002026032625, 26 Februari 2026
Pencipta		
Nama	:	**Riskiana Wulan, Mohammad Syafrullah dkk**
Alamat	:	Kalibata City Apt. Unit F20CM.RT/RW: 004/010, Pancoran, Kota Adm. Jakarta Selatan, DKI Jakarta, 12749
Kewarganegaraan	:	Indonesia
Pemegang Hak Cipta		
Nama	:	**Direktorat Riset dan Pengabdian kepada Masyarakat Universitas Budi Luhur**
Alamat	:	Jl. Ciledug Raya, RT.10/RW.2, Pesanggrahan, Kota Adm. Jakarta Selatan, DKI Jakarta, 12260
Kewarganegaraan	:	Indonesia
Jenis Ciptaan	:	**Program Komputer**
Judul Ciptaan	:	**Chatbot Cerdas Berbasis Agentic AI dengan Pendekatan Retrieval-Augmented Generation (RAG) untuk Peningkatan Pemahaman Kontekstual dan Autonomi Tindakan**
Tanggal dan tempat diumumkan untuk pertama kali di wilayah Indonesia atau di luar wilayah Indonesia	:	26 Februari 2026, di DKI Jakarta
Jangka waktu pelindungan	:	Berlaku selama 50 (lima puluh) tahun sejak Ciptaan tersebut pertama kali dilakukan Pengumuman.
Nomor Pencatatan	:	001156529

adalah benar berdasarkan keterangan yang diberikan oleh Pemohon.
Surat Pencatatan Hak Cipta atau produk Hak terkait ini sesuai dengan Pasal 72 Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta.

a.n. MENTERI HUKUM
DIREKTUR JENDERAL KEKAYAAN INTELEKTUAL
u.b
Direktur Hak Cipta dan Desain Industri

Agung Damarsasongko,SH.,MH.
NIP. 196912261994031001

Balai Besar Sertifikasi Elektronik

Disclaimer:
1. Dalam hal pemohon memberikan keterangan tidak sesuai dengan surat pernyataan, Menteri berwenang untuk mencabut surat pencatatan permohonan.
2. Surat Pencatatan ini telah disegel secara elektronik menggunakan segel elektronik yang ditertibkan oleh Balai Besar Sertifikasi Elektronik, Badan Siber dan Sandi Negara.
3. Surat Pencatatan ini dapat dibuktikan keasliannya dengan memindai kode QR pada dokumen ini dan informasi akan ditampilkan dalam browser.

LAMPIRAN PENCIPTA

No	Nama	Alamat
1	Riskiana Wulan	Kalibata City Apt. Unit F20CM.RT/RW: 004/010 Pancoran, Kota Adm. Jakarta Selatan
2	Mohammad Syafrullah	Graha Raya Bintaro, Cluster Melati Loka H12 No. 2, RT/RW: 002/016 Serpong Utara, Kota Tangerang Selatan
3	Anwar Rifai	Jl. Masjid Al-Baidho, RT.001/RW.003 Ciledug, Kota Tangerang

Balai Besar Sertifikasi Elektronik

Disclaimer:
1. Dalam hal pemohon memberikan keterangan tidak sesuai dengan surat pernyataan, Menteri berwenang untuk mencabut surat pencatatan permohonan.
2. Surat Pencatatan ini telah disegel secara elektronik menggunakan segel elektronik yang ditertibkan oleh Balai Besar Sertifikasi Elektronik, Badan Siber dan Sandi Negara.
3. Surat Pencatatan ini dapat dibuktikan keasliannya dengan memindai kode QR pada dokumen ini dan informasi akan ditampilkan dalam browser.