CHATBOT TELEGRAM BERBASIS KECERDASAN BUATAN MENGGUNAKAN *LARGE LANGUAGE MODEL QWEN 2.5* DAN *RETRIEVAL AUGMENTED GENERATION* STUDI KASUS PPID UNIVERSITAS LAMPUNG

(Skripsi)

Oleh

CHELLY SABRINA

NPM. 2115061042

FAKULTAS TEKNIK
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

CHATBOT TELEGRAM BERBASIS KECERDASAN BUATAN MENGGUNAKAN *LARGE LANGUAGE MODEL QWEN 2.5* DAN *RETRIEVAL AUGMENTED GENERATION* STUDI KASUS PPID UNIVERSITAS LAMPUNG

Oleh

CHELLY SABRINA

Skripsi

Sebagai Salah Satu Syarat Untuk Mencapai Gelar
SARJANA TEKNIK

Pada

Program Studi Teknik Informatika
Fakultas Teknik Universitas Lampung

FAKULTAS TEKNIK
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

ABSTRAK

CHATBOT TELEGRAM BERBASIS KECERDASAN BUATAN MENGGUNAKAN *LARGE LANGUAGE MODEL QWEN 2.5* DAN *RETRIEVAL AUGMENTED GENERATION* STUDI KASUS PPID UNIVERSITAS LAMPUNG

Oleh

CHELLY SABRINA

Keterbukaan akses informasi publik diatur dalam Undang-Undang No 14 Tahun 2008 tentang Keterbukaan Informasi Publik sebagai elemen utama untuk mewujudkan transparansi dan tata kelola pemerintahan yang baik. Di Universitas Lampung, Pejabat Pengelola Informasi dan Dokumentasi (PPID) menyediakan berbagai data publik melalui website resminya. Namun, pengguna sering mengalami kesulitan dalam menemukan informasi spesifik akibat tersebarnya dokumen serta keterbatasan fitur pencarian. Untuk mengatasi permasalahan tersebut, dikembangkan *chatbot* telegram berbasis kecerdasan buatan menggunakan *Large Language Model* (LLM) dan *Retrieval Augmented Generation* (RAG). *Dataset chatbot* diperoleh dari informasi yang tersedia pada *website* resmi PPID Universitas Lampung yang kemudian diproses untuk mendukung pencarian berbasis semantik. Sistem ini dikembangkan menggunakan model *Qwen*2.5 VL 72B *Instruct* dan pencarian vektor semantik menggunakan vektor *database* FAISS. Evaluasi dilakukan menggunakan matriks RAGAS menunjukkan tingkat akurasi yang baik dengan skor di atas ambang batas minimum 80%. Selanjutnya dilakukan juga evaluasi perbadingan dengan LLM murni tanpa *retrieval*, di mana *chatbot* yang dibuat menunjukkan tingkat akurasi yang lebih tinggi. Selain itu, dilakukan juga pengujian dari sisi pengalaman pengguna menggunakan metode *Chatbot Usability Questionnaire* (CUQ) menghasilkan skor rata-rata sebesar 90,62 menunjukkan bahwa *chatbot* memberikan pengalaman penggunaan yang baik.

Kata kunci: *Chatbot,* Telegram*, Large Language Model, Retrieval Augmented Generation, FAISS, Qwen.*

ABSTRACT

TELEGRAM CHATBOT BASED ON ARTIFICIAL INTELLIGENCE USING LARGE LANGUAGE MODEL QWEN 2.5 AND RETRIEVAL AUGMENTED GENERATION A CASE STUDY OF PPID UNIVERSITAS LAMPUNG

By

CHELLY SABRINA

Public access to information is regulated under Law No. 14 of 2008 concerning Keterbukaan Informasi Publik (UU KIP) as a key element in promoting transparency and good governance. At the University of Lampung, Pejabat Pengelola Informasi dan Dokumentasi (PPID) provides various public data through its official website. However, users often face difficulties in finding specific information due to scattered documents and limited search features. To address this issue, a Telegram chatbot powered by artificial intelligence was developed using a Large Language Model (LLM) and Retrieval Augmented Generation (RAG). The chatbot dataset was obtained from the official PPID Universitas Lampung website and was further processed to support semantic-based search. The system was developed using the Qwen2.5 VL 72B Instruct model, with semantic vector search performed using the FAISS vector database. Evaluation using the RAGAS metrics showed good accuracy levels, with scores exceeding the minimum threshold of 80%. A comparative evaluation with a pure LLM (without retrieval) was also conducted, in which the developed chatbot demonstrated higher accuracy. In addition, user experience testing was carried out using the Chatbot Usability Questionnaire (CUQ), yielding an average score of 90.62, indicating that the chatbot provides a good user experience.

Keywords: *Chatbot, Telegram, Large Language Model, Retrieval Augmented Generation, FAISS, Qwen*

Judul Skripsi	:	***CHATBOT* TELEGRAM BERBASIS KECERDASAN BUATAN MENGGUNAKAN *LARGE LANGUAGE MODEL QWEN 2.5* DAN *RETRIEVAL AUGMENTED GENERATION* STUDI KASUS PPID UNIVERSITAS LAMPUNG**
Nama Mahasiswa	:	Chelly Sabrina
Nomor Pokok Mahasiswa	:	**2115061042**
Jurusan	:	**Teknik Informatika**
Fakultas	:	**Teknik**

MENYETUJUI

1. Komisi Pembimbing

Pembimbing Utama	Pembimbing Pendamping
Dr. Ir. M. Komarudin, S.T., M.T.	**Yessi Mulyani, S.T., M.T.**
NIP. 196812071997031006	**NIP. 197312262000122001**

2. Mengetahui

Ketua Jurusan Teknik Elektro	Ketua Program Studi Teknik Informatika
Herlinawati, S.T., M.T.	**Yessi Mulyani, S.T., M.T.**
NIP. 197103141999032001	**NIP. 197312262000122001**

MENGESAHKAN

1. Tim Penguji

   Ketua : **Dr. Ir. M. Komarudin, S.T., M.T.**

   Sekretaris : **Yessi Mulyani, S.T., M.T.**

   Penguji : **Puput Budi Wintoro, S.Kom., M.T.I.**

2. Dekan Fakultas Teknik

Dr. Eng. Ir. Helmy Fitriawan, S.T., M.Sc.
NIP. 197509282001121002

Tanggal Lulus Ujian Skripsi: **31 Juli 2025**

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini, menyatakan bahwa skripsi saya dengan judul "*Chatbot* Telegram Berbasis Kecerdasan Buatan Menggunakan *Large Language Model Qwen 2.5* dan *Retrieval Augmented Generation* Studi Kasus PPID Universitas Lampung" dibuat oleh saya sendiri. Semua hasil yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila di kemudian hari terbukti bahwa skripsi ini merupakan salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan hukurn atau akademik yang berlaku.

Bandar Lampung, 05 Agustus 2025

Pembuat Pernyataan,

Chelly Sabrina

2115061042

RIWAYAT HIDUP

Penulis lahir di Lubuk Linggau pada tanggal 18 Januari 2003 dari pasangan Bapak Serka (Purn.) Dwi Susanto dan Ibu Emil Dayeni, S.Pd. Penulis telah menyelesaikan pendidikan formal di SDN 13 Argamakmur pada tahun 2015, SMPN 1 Bangunrejo pada tahun 2018, dan SMAN 1 Bangunrejo pada tahun 2021. Pada tahun 2021, penulis terdaftar sebagai mahasiswa Program Studi Teknik Informatika Universitas Lampung melalui jalur SBMTN. Selama menjadi mahasiswa, penulis telah aktif dalam berbagai kegiatan pengembangan diri dan meraih penghargaan. Kegiatan dan penghargaan tersebut di antaranya adalah:

1. Menjuarai lomba kepenulisan (esai, artikel, dan sajak) tingkat nasional sebanyak 9 kali.
2. Penerima Beasiswa Karya Salemba Empat (KSE) pada tahun 2024 - 2025 oleh Yayasan Karya Salemba Empat.
3. Menjadi pembicara *Student Representative* Bangkit Academy perwakilan Provinsi Lampung pada tahun 2023.
4. Mengikuti program Magang Bersertifikat Kampus Merdeka *Batch* 6 di PT United Tractor sebagai *IT and Data Science* pada 16 Februari 2024 sampai 30 Juni 2024.
5. Mengikuti program Studi Independen Kampus Merdeka di Bangkit Academy Batch 5 tahun 2023 mengambil kelas *Machine Learning*.
6. Menjadi anggota Himpunan Mahasiswa Jurusan Teknik Elektro Universitas Lampung Departemen Sosial dan Kewirausahaan Divisi Sosial periode 2022/2023 dan 2023/2024.
7. Menjadi anggota Keluarga Muda Bina Rohani Mahasiswa Islam Universitas Lampung (KM Birohmah) Departemen Hubungan Masyarakat pada tahun 2021.

MOTTO

“Maka sesungguhnya bersama kesulitan ada kemudahan.”

(Q.S. Al-Insyirah: 5)

“Allah tidak membebani seseorang melainkan sesuai dengan kesanggupannya.”

(Q.S. Al-Baqarah: 286)

“Jangan ragu untuk ketuk pintu lain, jika satu pintu tertutup untukmu.”

(Maudy Ayunda)

“Nan mangseoriji anko, Idaero ga, It's my fate”
(Enhypen: Fate)

PERSEMBAHAN

Bismillahirrohmanirrohim, segala puji bagi Allah SWT. Tuhan Yang Maha Esa, karena atas nikmat dan karunia-Nya, Saya dapat menyelesaikan skripsi ini.

Kupersembahkan Skripsi ini Kepada:

"Diriku yang telah berjuang sejauh ini untuk menyelesaikan skripsi ini. Terima kasih karena telah berusaha sekuat tenaga, tidak menyerah, dan tetap percaya bahwa semua ini bisa dilalui. Mengorbankan pikiran, tenanga, dan waktu untuk untuk melewati segala rintangan demi mencapai tujuan ini. Semoga seluruh impian dan cita-cita mulia dikemudian hari dapat segera tercapai."

"Ibuku yang senantiasa memberikan dukungan tanpa henti, kasih sayang yang tulus, dan doa-doa terbaik disetiap sujudnya. Terima kasih telah menjadi sumber kekuatan, pelipurlara, dan cahaya dalam setiap gelap yang kulalui. Terima kasih karena telah bekerja keras setiap hari, mengangkat dagangan yang berat di bawah terik matahari, demi aku bisa melanjutkan kuliah. Tanpa restumu, mungkin langkah ini tidak akan sampai sejauh ini. Skripsi ini kupersembahkan dengan penuh cinta dan hormat untukmu, ibuku"

SANWACANA

Puji syukur kehadirat Allah SWT yang telah memberikan rahmat dan hidayat-Nya sehingga penulis dapat menyelesaikan skripsi ini. Salawat serta salam penulis sanjungkan kepada Nabi dan Rasul Muhammad SAW yang penulis harapkan syafaatnya di hari akhir kelak.

Skripsi yang berjudul "*Chatbot* Telegram Berbasis Kecerdasan Buatan Menggunakan *Large Language Model Qwen 2.5* dan *Retrieval Augmented Generation* Studi Kasus PPID Universitas Lampung" ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana teknik pada Program Studi Teknik Informatika, Universitas Lampung. Dalam proses penelitian ini, sangat banyak orang-orang yang terlibat dalam pelaksanaannya. Oleh karena itu, penulis mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Kedua orang tua dan keluarga yang selalu menyertai penulis dengan doa, kasih sayang, dan dukungan moril sepanjang perjalanan hidup penulis.
2. Bapak Dr. Eng. Helmy Fitriawan, S.T., M.Sc., selaku Dekan Fakultas Teknik Universitas Lampung.
3. Ibu Herlinawati, S.T., M.T., selaku Ketua Jurusan Teknik Elektro Universitas Lampung.
4. Ibu Yessi Mulyani, S.T., M.T. selaku Ketua Program Studi Teknik Informatika Universitas Lampung sekaligus pembimbing pendamping yang telah membantu proses kelancaran pengerjaan penelitian dan telah memberikan bimbingan, dukungan, serta masukan secara detail terhadap penyelesaian skripsi ini.
5. Bapak Dr. Ir. Muhamad Komarudin, S.T., M.T selaku pembimbing utama yang selalu bersedia meluangkan waktunya untuk memberikan bimbingan, arahan

dalam proses pengerjaan penelitian, dan dukungan kepada penulis dalam menyelesaikan penelitian ini.

6. Bapak Puput Budi Wintoro, S.Kom., M.T.I. selaku penguji yang telah memberikan banyak pelajaran dan masukan kepada penulis terkait dengan pelaksanaan penelitian ini.
7. Teman-teman *Livinginbrg* yang telah banyak membantu penulis dan menjadi tempat bertukar pikiran sejak semester satu hingga saat ini.
8. Teman-teman *Support System* OSIS SMAN 1 Bangunrejo yang telah banyak membantu penulis sejak masa SMA baik melalui dukungan semangat, masukan yang membangun, serta ajakan untuk terus bertawakal dan memperbaiki diri.
9. Teman-teman United Tractors *Intern* MBKM *Batch* 6 serta teman-teman Bangkit Academy MBKM *Batch* 5 yang telah memberikan banyak inspirasi kepada penulis.
10. Teman-teman PSTI 21 yang tidak bisa penulis sebutkan satu per satu, terima kasih atas segala bantuan, kebersamaan, dan semangat yang telah kalian berikan selama masa perkuliahan. Dukungan dan canda tawa kalian menjadi bagian berharga dalam perjalanan ini.

Penulis menyadari bahwa penelitian ini masih banyak kekurangannya, baik dari segi penulisan maupun sistem yang dikembangkan. Oleh karena itu penulis memohon maaf dan menerima saran terhadap apa yang telah penulis tuangkan dalam skripsi ini.

Bandar Lampung, 05 Agustus 2025
Penulis,

Chelly Sabrina
2115061042

DAFTAR ISI

Halaman

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1 Latar Belakang

Keterbukaan informasi publik merupakan elemen penting dalam mewujudkan tata kelola yang baik, transparan, dan akuntabel diberbagai sektor. Hal ini diatur dalam Undang-Undang Nomor 14 Tahun 2008 tentang Keterbukaan Informasi Publik (UU KIP) yang mewajibkan badan publik untuk menyediakan informasi secara terbuka dan dapat diakses oleh masyarakat [1]. Prinsip ini tidak hanya berlaku pada lembaga pemerintahan, tetapi juga pada organisasi non-pemerintah, institusi pendidikan, dan dunia usaha. Dengan keterbukaan informasi, masyarakat memiliki akses yang lebih luas terhadap data dan kebijakan yang memungkinkan partisipasi aktif dalam pengambilan keputusan serta pengawasan terhadap pelaksanaan program. Hal ini menjadi landasan guna menciptakan lingkungan yang lebih adil, demokratis, dan berorientasi pada kepentingan publik.

Sebagai institusi pendidikan tinggi, Universitas Lampung (Unila) memiliki tanggung jawab untuk melaksanakan keterbukaan informasi sesuai dengan amanat UU KIP. Dalam rangka memenuhi kewajiban tersebut, Unila membentuk Pejabat Pengelola Informasi dan Dokumentasi (PPID) sebagai pusat pengelolaan dan layanan informasi publik. PPID Unila bertugas menyediakan informasi yang relevan, akurat, dan mudah diakses baik yang berkaitan dengan kebijakan maupun program di universitas [2]. Kehadiran PPID ini tidak hanya mendukung tata kelola universitas yang transparan, tetapi juga mampu meningkatkan kualitas layanan pendidikan serta mendorong partisipasi aktif publik.

Dalam mendukung keterbukaan informasi, PPID Unila menggunakan *website* sebagai sarana utama untuk menyampaikan informasi kepada publik. Melalui situs ini, khalayak dapat mengakses berbagai dokumen seperti laporan kinerja, kebijakan institusi, serta prosedur layanan. Namun, berdasarkan hasil observasi awal yang dilakukan peneliti ditemukan bahwa tidak semua informasi tersedia dalam format yang langsung ditampilkan atau mudah diakses oleh pengguna. Informasi yang ada dalam website PPID Unila tersebar dalam banyak dokumen dan struktur navigasi situs memerlukan pemahaman khusus agar pengguna dapat menemukan data yang relevan dengan kebutuhannya.

Temuan ini menunjukkan adanya kebutuhan akan inovasi dalam penyajian informasi yang lebih adaptif, interaktif, dan mudah dipahami oleh pengguna. Salah satu solusi potensial yang dapat diterapkan adalah penggunaan *chatbot* berbasis kecerdasan buatan atau *Artificial Intelligence* (AI). Saat ini, teknologi kecerdasan buatan tengah mengalami perkembangan pesat dan menjadi salah satu tren utama di berbagai bidang. Hadirnya AI telah membawa banyak manfaat, terutama dalam mempercepat dan mempermudah akses terhadap informasi [3]. Dalam penelitian ini, *chatbot* dipilih sebagai solusi karena kemampuannya memberikan layanan pencarian informasi secara cepat, mudah, dan interaktif. Berbeda dengan metode pencarian manual di *website* yang mengharuskan pengguna membaca banyak dokumen atau halaman, *chatbot* memungkinkan pengguna memperoleh jawaban spesifik hanya dengan mengetikkan pertanyaan [4].

Hadirnya *chatbot* juga dapat meningkatkan kemudahan akses informasi dan kepuasan pengguna melalui pengalaman tanya jawab yang intuitif dan interaktif. Hal ini diperkuat oleh penelitian Xu et al. (2023) yang membandingkan *chatbot* seperti ChatGPT dengan *Google Search* untuk tugas pencarian informasi. Hasilnya menunjukkan bahwa penggunaan *chatbot* mengurangi waktu penyelesaian tugas pencarian sebesar 65,20% dengan tingkat akurasi jawaban yang setara. Artinya, *chatbot* memungkinkan pengguna menyelesaikan pencarian informasi dalam sepertiga waktu dibandingkan dengan mesin pencari konvensional, sehingga dapat meningkatkan pengalaman dan efisiensi dalam proses pencarian informasi [5].

Temuan serupa juga disampaikan oleh Arz Von Straussenburg (2023) yang menunjukkan bahwa pencarian informasi lebih mudah dilakukan dengan menggunakan *chatbot* berbasis AI dibandingkan dengan penelusuran melalui situs web secara manual. Hal ini terlihat pada pertanyaan-pertanyaan yang dapat dijawab secara langsung oleh *chatbot*, tanpa perlu menelusuri banyak halaman [6]. Dengan berbagai manfaat tersebut, penerapan *chatbot* berbasis kecerdasan buatan menjadi solusi strategis untuk meningkatkan aksesibilitas, kecepatan layanan, dan pengalaman pengguna dalam keterbukaan sistem informasi publik PPID Universitas Lampung.

Penelitian ini memilih mengembangkan *chatbot* berbasis *Large Language Model* (LLM) yang terintegrasi dengan arsitektur *Retrieval Augmented Generation* (RAG). *Chatbot* berbasis LLM dipilih karena memiliki kemampuan pemahaman bahasa alami yang lebih baik dibanding *chatbot* berbasis *Pattern-Based Chatbot* (aturan sederhana) atau *Retrivalbased Chatbot* (pengambilan teks). Keputusan ini didasarkan pada studi perbandingan yang menunjukkan bahwa penggunaan LLM murni seperti ChatGPT untuk menjawab pertanyaan khusus dari *website* PPID Unila masih menghasilkan sejumlah kesalahan (*hallucination*) karena model tersebut menghasilkan jawaban berdasarkan pola bahasa tanpa akses ke sumber informasi aktual. Oleh karena itu, penelitian ini berfokus pada pengembangan *chatbot* berbasis arsitektur RAG yang menyediakan data tambahan khusus dari *website* PPID Unila, sehingga dapat mengurangi terjadinya halusinasi yang sering kali menghasilkan informasi yang tidak akurat atau relevan. Halusinasi ini terjadi karena LLM hanya bergantung pada pola-pola dalam data pelatihan untuk menghasilkan teks, tanpa kemampuan untuk mengakses informasi eksternal.

Dengan ini, diharapkan *chatbot* dapat berfungsi secara optimal dalam mendukung kemudahan penyampaian informasi PPID Unila, memberikan informasi dengan lebih cepat, serta memastikan kualitas dan akurasi data yang disajikan. Hal ini akan menciptakan sistem informasi yang lebih transparan, memudahkan pengguna dalam mengakses data yang dibutuhkan, serta meningkatkan keterlibatan publik dalam memperoleh informasi.

1.2 Rumusan Masalah

Adapun rumusan masalah yang diangkat dari penelitian ini adalah:

1. Bagaimana membuat sistem kecerdasan buatan dalam pengembangan *chatbot* LLM dan RAG yang dapat menjadi solusi untuk meningkatkan kemudahan akses informasi publik di lingkungan PPID Universitas Lampung?
2. Bagaimana performa *chatbot* yang dikembangkan mampu memberikan jawaban seputar informasi publik yang tersedia di PPID Universitas Lampung?
3. Bagaimana penerapan arsitektur RAG pada *chatbot* dapat mengurangi potensi halusinasi jawaban dibandingkan dengan penggunaan LLM secara murni (tanpa *retrieval*)?

1.3 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Mengembangkan sistem kecerdasan buatan melalui *chatbot* LLM dan RAG untuk mendukung kemudahan akses informasi publik di lingkungan PPID Universitas Lampung.
2. Menganalisis performa jawaban yang diberikan oleh *chatbot* terhadap pertanyaan-pertanyaan seputar informasi publik yang tersedia pada website PPID Universitas Lampung.
3. Mengevaluasi efektivitas arsitektur RAG dalam mengurangi potensi halusinasi jawaban dibandingkan dengan penggunaan LLM secara murni (tanpa *retrieval)*.

1.4 Manfaat Penelitian

Adapun manfaat dari penelitian ini diantaranya:

1. Secara akademis, penelitian ini diharapkan dapat memberikan pengalaman dan pemahaman bagi peneliti dalam menerapkan sistem kecerdasan buatan, khususnya pengembangan *chatbot* berbasis LLM dan RAG.

2. Secara praktis, penelitian ini diharapkan dapat memberikan manfaat bagi PPID Universitas Lampung dalam menyediakan solusi inovatif untuk meningkatkan kemudahan penyampaian informasi publik melalui pengembangan *chatbot*.
3. Secara teknis, penelitian ini diharapkan dapat memberikan gambaran mengenai penerapan arsitektur RAG untuk meningkatkan akurasi jawaban *chatbot* dengan menggabungkan kemampuan pemrosesan bahasa alami dari LLM dan pengambilan informasi relevan dari sumber data aktual.

1.5 Batasan Masalah

Adapun batasan masalah pada penelitian ini di antaranya:

1. Data yang diigunakan terbatas pada data yang tersedia di *website* PPID Universitas Lampung.
2. Akurasi jawaban *chatbot* bergantung pada kualitas dan relevansi data yang dikelola RAG serta kemampuan model *Qwen* 2.5 VL 72B *Instruct* dalam memproses dan menghasilkan jawaban yang sesuai dengan konteks.
3. *Chatbot* akan diimplementasikan dengan platform telegram, sehingga terbatas oleh fitur yang disediakan oleh *Telegram Bot*.

1.6 Sistematika Penulisan

Dalam laporan penelitian ini, sistematika penulisan yang digunakan adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini membahas secara umum mengenai latar belakang penelitian, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan penelitian, serta sistematika penulisan dalam penelitian pengembangan *Chatbot* Telegram Berbasis Kecerdasan Buatan Menggunakan *Large Language Model Qwen 2.5* dan *Retrieval Augmented Generation* Studi Kasus PPID Universitas Lampung.

BAB II TINJAUAN PUSTAKA

Bab ini memuat teori-teori dasar yang digunakan sebagai sumber untuk mendukung penelitian, seperti *Artificial Intelligence, Machine Learning, Artificial Neural Network, Deep Learning, Natural Language Processing, Large Language Model, Transformer, Retrieval Augmented Generation, Chatbot, Tokenisasi, Embedding, Indexing, tools* yang digunakan, dan penelitian terkait.

BAB III METODE PENELITIAN

Bab ini memuat waktu dan tempat penelitian, alat dan bahan, tahapan penelitian, tahapan pengembangan *chatbot*, serta diagram alir dari tahapan penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini membahas mengenai hasil yang diperoleh dalam penelitian pengembangan *Chatbot* Telegram Berbasis Kecerdasan Buatan Menggunakan *Large Language Model Qwen 2.5* dan *Retrieval Augmented Generation* Studi Kasus PPID Universitas Lampung.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dan saran dari hasil penelitian yang telah dilakukan sebagai bahan penelitian lebih lanjut.

BAB II
TINJAUAN PUSTAKA

2.1 Landasan Teori

2.1.1 *Artificial Intelligence*

Artificial Intelligence (AI) atau kecerdasan buatan adalah bidang ilmu komputer yang memungkinkan komputer untuk meniru kemampuan kecerdasan manusia. Sistem berbasis AI akan menghasilkan keluaran (*output*) berupa solusi dari suatu permasalahan berdasarkan sekumpulan pengetahuan (*knowlagde*) yang telah diperlajari sebelumnya. Dalam penerapannya, AI bertujuan untuk mempermudah pekerjaan manusia menjadi lebih efisien dan efektif melalui pemprosesan data untuk mengambil keputusan [7].

AI melibatkan penggunaan algoritma dan model matematika yang dirancang untuk memungkinkan komputer belajar dari data, mengenali pola, dan membuat keputusan secara cerdas [8]. Proses ini dimulai dengan pengumpulan dan pengolahan data yang menjadi dasar pembelajaran sistem. Algoritma AI digunakan untuk melatih model agar mampu memahami hubungan antar data. Dengan mengenali pola yang ada dalam data, sistem AI dapat memprediksi hasil atau mengambil tindakan tertentu berdasarkan analisis yang telah dilakukan. Kemampuan ini membuat AI menjadi alat yang sangat *powerful* untuk berbagai aplikasi, mulai dari pengenalan suara dan wajah hingga analisis data besar dan otomatisasi proses pengambilan keputusan. Dengan terus berkembangnya teknologi dan ketersediaan data, AI semakin efektif dalam memberikan solusi inovatif yang dapat diterapkan di berbagai bidang kehidupan.

2.1.2 *Machine Learning*

Machine Learning (ML) adalah cabang dari kecerdasan buatan yang berfokus pada pengembangan algoritma yang memungkinkan sistem komputer belajar dari data dan pengalaman tanpa memerlukan instruksi eksplisit dengan menggunakan data besar dan beragam, ML memungkinkan sistem untuk mengidentifikasi pola dan hubungan yang mungkin tidak terlihat secara langsung. Proses ini melibatkan pelatihan model yang memproses data untuk mengekstrak informasi dan membuat keputusan berdasarkan pola teridentifikasi. Teknologi ini telah digunakan di berbagai bidang, seperti bidang kesehatan untuk mendeteksi penyakit, bidang bisnis untuk menganalisis perilaku konsumen, dan bidang teknologi untuk pengembangan pengenalan wajah serta pengolahan bahasa alami [9].

2.1.3 *Artificial Neural Network*

Artificial Neural Network (ANN) adalah algoritma *machine learning* yang arsitekturnya terinspirasi dari cara kerja otak manusia dalam memproses informasi. ANN terdiri dari unit-unit komputasi yang disebut *neuron* dan tersusun dalam bentuk jaringan yang saling terhubung antar lapisan. Masing-masing *neuron* dalam jaringan ini terhubung satu sama lain melalui bobot (*weights*) yang akan menentukan kekuatan sinyal antar *neuron*. Setiap *neuron* menerima *input*, mengolahnya melalui fungsi aktivasi, lalu meneruskan *output* ke *neuron* di lapisan berikutnya. Jaringan ini biasanya dibagi menjadi tiga lapisan utama, yaitu lapisan *input*, satu atau lebih lapisan tersembunyi (*hidden layers*), dan lapisan *output*. Melalui arsitektur ini, ANN mampu mengenali pola dan hubungan dalam data [10].

Proses ANN diawali dengan tahap *forward propagation* yaitu proses di mana data yang masuk melalui lapisan *input* diteruskan ke lapisan-lapisan jaringan secara berurutan. Setiap lapisan akan melakukan perhitungan berdasarkan nilai bobot dan bias yang dimilikinya, kemudian hasilnya diteruskan melalui fungsi aktivasi untuk menghasilkan keluaran (*output*) sementara. Proses ini berlangsung dari lapisan *input* hingga lapisan *output*, dan hasil akhirnya berupa prediksi yang dihasilkan oleh

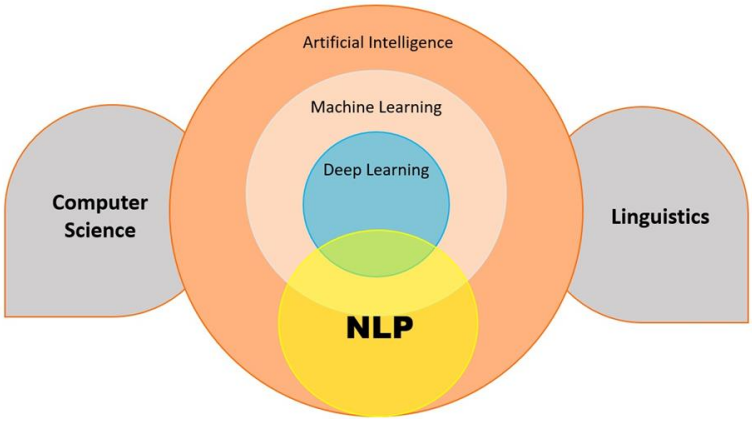
model. Setelah model menghasilkan prediksi, dilakukan evaluasi dengan cara membandingkan hasil prediksi tersebut dengan nilai sebenarnya (*ground truth*) menggunakan fungsi kesalahan (*loss function*). Selisih antara prediksi dan nilai sebenarnya ini disebut *error*. Untuk memperbaiki kesalahan tersebut, digunakan algoritma *back propagation*, yaitu proses menyebarkan kembali *error* dari lapisan *output* ke lapisan-lapisan sebelumnya. Melalui pendekatan ini, *back propagation* menghitung seberapa besar pengaruh setiap bobot terhadap kesalahan tersebut. Informasi ini kemudian digunakan untuk memperbarui nilai bobot dengan tujuan agar model dapat menghasilkan prediksi yang lebih akurat diiterasi berikutnya [11].

2.1.4 *Deep Learning*

Deep learning adalah pengembangan dari ANN yang menggunakan jaringan dengan banyak lapisan tersembunyi (*hidden layers*). *Deep Learning* ini juga dirancang untuk mengajarkan komputer cara memproses data yang terinspirasi dari cara kerja otak manusia. Metode ini menggunakan *neuron network* yang terdiri dari banyak lapisan untuk mengenali, mengklasifikasikan, dan menggambarkan objek dalam data [12]. *Deep Learning* bekerja dengan memproses data melalui berbagai lapisan yang memungkinkan model belajar melalui data yang sangat kompleks. Proses dalam *deep learning* dimulai dengan lapisan *input* yang menerima data mentah, seperti gambar, teks, atau suara. Data ini kemudian diproses melalui beberapa lapisan *hidden layers*. Lapisan-lapisan ini bekerja secara berurutan, dengan setiap lapisan memperdalam pemahaman model mengenai data. Hingga akhirnya, lapisan *output* berhasil menghasilkan prediksi atau keputusan berdasarkan analisis yang dilakukan oleh lapisan-lapisan sebelumnya [13].

2.1.5 *Natural Language Processing*

Natural Language Processing (NLP) adalah cabang dari kecerdasan buatan yang yang mampu mempelajari komunikasi antara manusia dengan komputer melalui bahasa alami [14]. Bahasa alami sendiri adalah bahasa yang secara umum digunakan oleh manusia dalam berkomunikasi satu sama lain. Agar komputer dapat memahami dan memproses bahasa alami, perlu dilakukan proses yang bertujuan untuk menerjemahkan maksud pengguna ke dalam format yang bisa dimengerti oleh komputer, sehingga sistem dapat merespons dengan tepat. NLP digunakan untuk melakukan tugas-tugas yang berkaitan dengan teks, seperti *machine translation, digital assistants, search engines, customer service*, dan *chatbot* [15].



Gambar 1. Posisi NLP dalam Bidang AI [16]

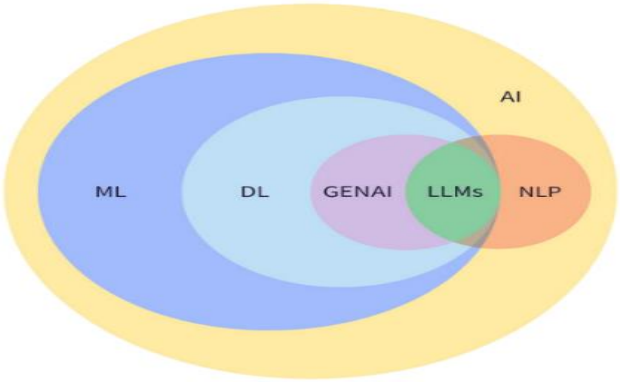
Dalam NLP, pemrosesan data sering dilakukan menggunakan arsitektur model *Deep Learning* yang dirancang untuk memahami pola, hubungan, dan makna mendalam suatu teks [17]. Model ini memanfaatkan berbagai teknik untuk menganalisis data berupa kata atau kalimat, seperti pengenalan pola sekuensial dan konteks antar kata. Salah satu model awal yang digunakan adalah *Recurrent Neural Networks* (RNN) yang mampu menangkap hubungan antar kata dalam urutan teks, meskipun memiliki keterbatasan dalam menangani hubungan jangka panjang.

2.1.6 *Large Language Model*

Large Language Model (LLM) adalah model yang dilatih menggunakan *dataset* teks dalam jumlah besar yang digunakan dalam berbagai penerapan model NLP. Model ini dilatih dengan menggunakan jutaan hingga miliar parameter teks untuk menangani tugas-tugas yang memerlukan pemrosesan bahasa alami. Melalui pemanfaatan data yang sangat besar, LLM mampu mempelajari pola bahasa, struktur kalimat, dan konteks bahasa manusia [18].

Cara kerja LLM dalam menghasilkan teks dimulai dengan memprediksi satu token (kata atau subkata) pada satu waktu, berdasarkan token-token yang telah dihasilkan sebelumnya. Proses ini diawali dengan *pre-training* pada data teks yang sangat besar dan beragam, di mana model belajar memahami pola bahasa, tata bahasa, dan informasi faktual dengan memprediksi kata berikutnya dalam suatu kalimat. Ketika diberikan *prompt* atau masukan teks, model akan memecah teks tersebut menjadi token-token individual dan mengubahnya menjadi representasi numerik yang disebut dengan *embedding*.

Large Language Model bekerja dengan cara menghasilkan teks baru berdasarkan prediksi kata yang paling mungkin muncul berikutnya. Prediksi ini hadir didasari oleh pola yang telah dipelajari dari data teks selama pelatihan. Namun, kekurangan utama LLM adalah terbatasnya pengetahuan yang dimiliki karena informasi yang diperoleh hanya dari data pelatihan dan tidak dapat diperbarui secara *real-time*. Akibatnya, LLM dapat menghasilkan halusinasi, yaitu informasi yang salah atau tidak faktual. Halusinasi terjadi karena model mencoba mengisi kekosongan pengetahuan dengan asumsi berdasarkan pola yang dipelajari tanpa pemahaman tentang benar atau salah informasi tersebut.



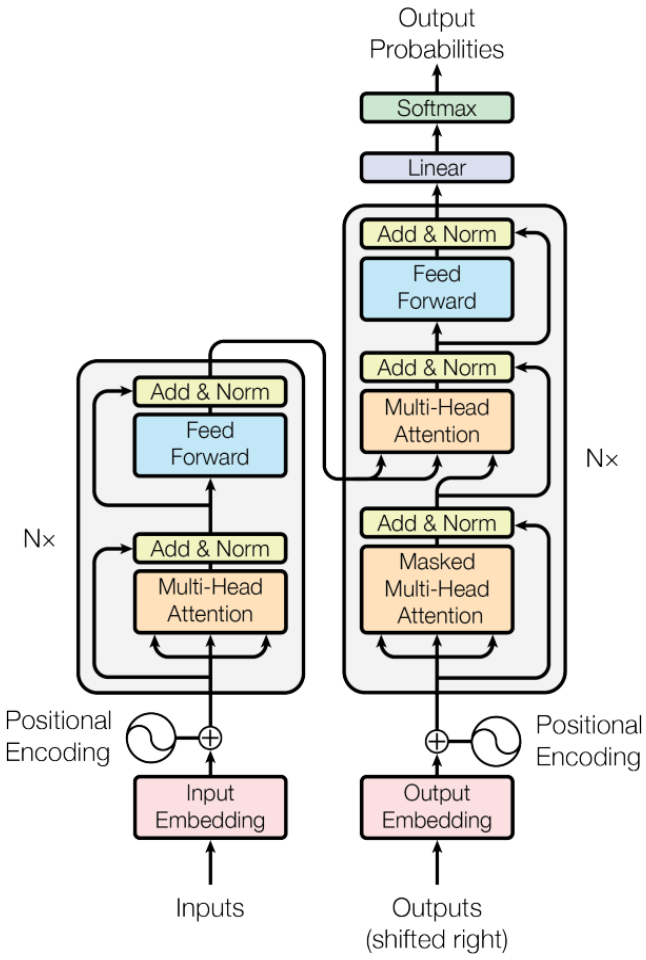
Gambar 2. Posisi LLM dalam Bidang AI [19]

2.1.7 *Transformer*

Transformer adalah sebuah arsitektur model yang menghilangkan penggunaan mekanisme *recurrence* seperti pada RNN dan sepenuhnya mengandalkan mekanisme *attention* untuk menangkap ketergantungan global antara *input* dan *output* [20]. Mekanisme ini memungkinkan *transformer* memproses data secara paralel yang menjadikannya lebih efisien dan mampu mencapai standar baru dalam berbagai tugas pemrosesan bahasa alami (NLP). Dalam konteks LLM, *transformer* memungkinkan pemrosesan token secara bersamaan, sehingga mempercepat waktu pelatihan dan inferensi dibandingkan dengan model berbasis *sequence* tradisional. Komponen kunci dalam arsitektur *transformer* adalah *self-attention* yang memungkinkan model untuk memahami hubungan antar kata atau token, bahkan jika kata-kata tersebut tidak berada dalam posisi yang berdekatan.

Arsitektur *transformer* terdiri dari dua komponen utama, yaitu *encoder* dan *decoder*. *Encoder* bertugas memproses *input* data menjadi representasi vektor yang kaya informasi. *Encoder* terdiri dari lapisan *self-attention* dan *feed-forward* yang memungkinkan setiap kata dalam *input* memperhatikan kata-kata lain dan memperkaya representasi informasi secara independen pada setiap token. Sementara itu, *decoder* menggunakan informasi dari *encoder* untuk memprediksi dan menghasilkan *output*. *Decoder* juga memiliki lapisan *self-attention* yang memungkinkan prediksi token berikutnya dengan tetap mempertahankan

konsistensi konteks sebelumnya. Kombinasi ini membuat *transformer* menjadi dasar bagi perkembangan LLM terkemuka seperti GPT (*Generative Pre-trained Transformer*) dan BERT (*Bidirectional Encoder Representations from Transformers*).



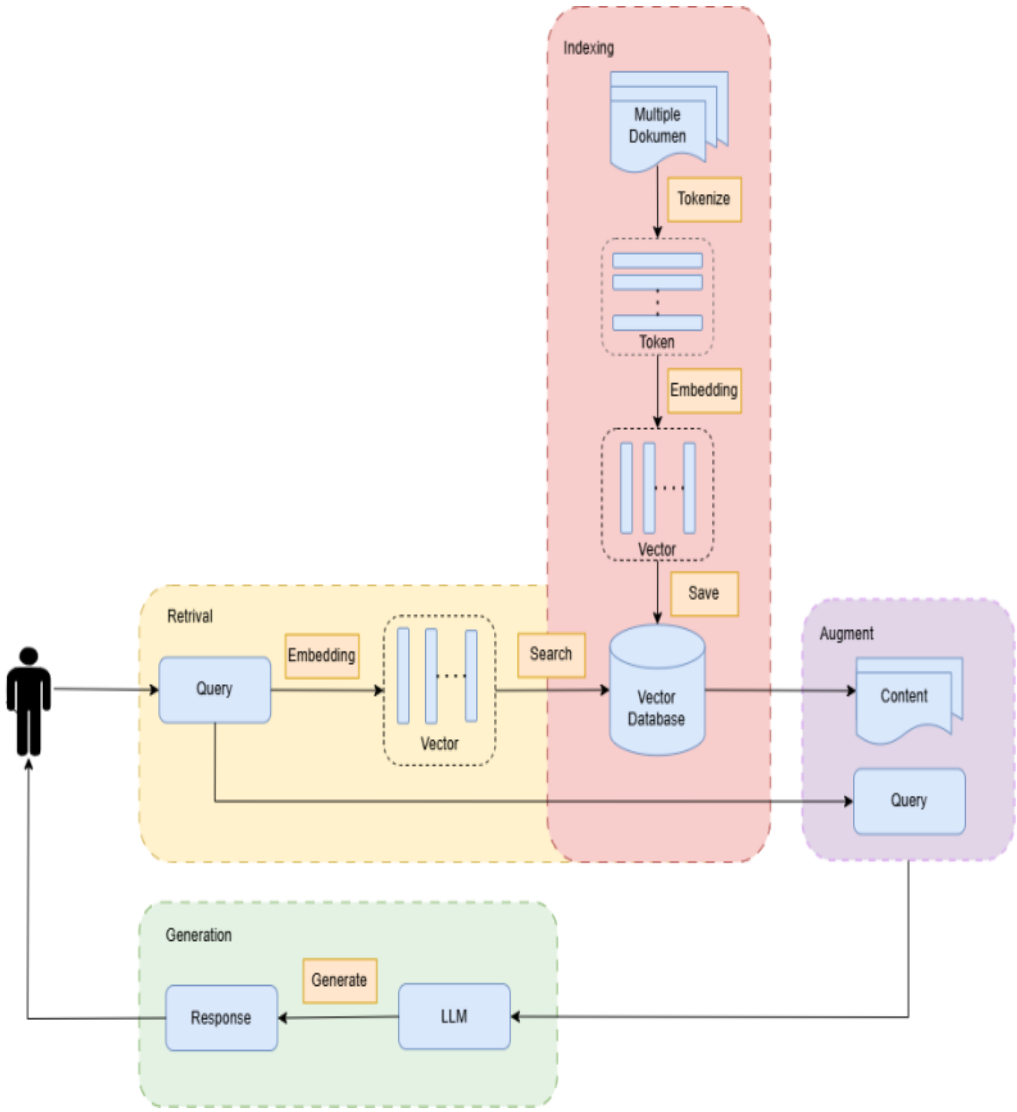
Gambar 3. Arsitektur *Transformer* [20]

2.1.8 *Retrieval Augmented Generation*

Retrieval Augmented Generation (RAG) adalah teknik NLP yang menggabungkan kemampuan LLM dengan mekanisme pengambilan informasi (*Information Retriever*) dari sumber eksternal [21]. Teknik ini memungkinkan LLM memperluas kemampuannya ke data tertentu atau basis pengetahuan internal organisasi tanpa perlu melatih ulang model sehingga mengatasi keterbatasan

pengetahuan LLM yang hanya bergantung pada data *pre-training*. Selain itu, RAG juga efektif dalam mengurangi halusinasi yaitu kecenderungan model menghasilkan informasi yang tidak faktual atau tidak relevan, sehingga mampu meningkatkan akurasi dan relevansi *output* LLM [22].

Dengan adanya RAG, model dapat mengambil informasi dari sumber eksternal dengan cara mengambil potongan dokumen yang relevan dari basis pengetahuan eksternal melalui kalkulasi kesamaan semantik. RAG terdiri dari empat tahap utama yaitu *indexing*, *retrieval*, *augment* dan *generation*. Proses pertama *indexing* yaitu proses ketika sumber pengetahuan eksternal dipecah menjadi potongan-potongan informasi yang lebih kecil yang disebut token. Setiap token diubah menjadi representasi numerik (vektor) menggunakan teknik *embedding* yang kemudian disimpan dalam *vector database* untuk memudahkan pencarian informasi. Tahap selanjutnya adalah *retrieval* yang terjadi ketika LLM menerima *Query* atau pertanyaan dari pengguna, sistem RAG melalui teknik *embedding* akan mengubah *Query* tersebut menjadi bentuk vektor juga dan membandingkannya dengan vektor-vektor yang ada di dalam indeks untuk mengambil potongan-potongan informasi yang paling relevan secara semantik. Kemudian tahap *augment* terjadi ketika informasi relevan ditemukan, data yang diambil digabungkan dengan *input* pengguna untuk memberikan kontekstualisasi. Tahap terakhir adalah *generation*, pada tahap ini menghasilkan jawaban dengan memanfaatkan LLM untuk menghasilkan teks sesuai dengan data yang telah diambil [23].



Gambar 4. Cara Kerja *Retrieval Augmented Generation*

Dalam konteks *chatbot* untuk *website* PPID Unila, penerapan RAG dapat meningkatkan kualitas pelayanan informasi publik. Dengan memanfaatkan mekanisme RAG, *chatbot* dapat memberikan jawaban yang cepat dan relevan terhadap permintaan informasi berdasarkan dokumen resmi dan data publik yang tersedia di *website* PPID Unila. Selain itu, RAG memungkinkan *chatbot* untuk menjawab pertanyaan kompleks dengan merujuk langsung pada sumber informasi yang spesifik, sehingga memastikan transparansi, kecepatan, dan keakuratan dalam menyampaikan informasi kepada masyarakat.

2.1.9 *Chatbot*

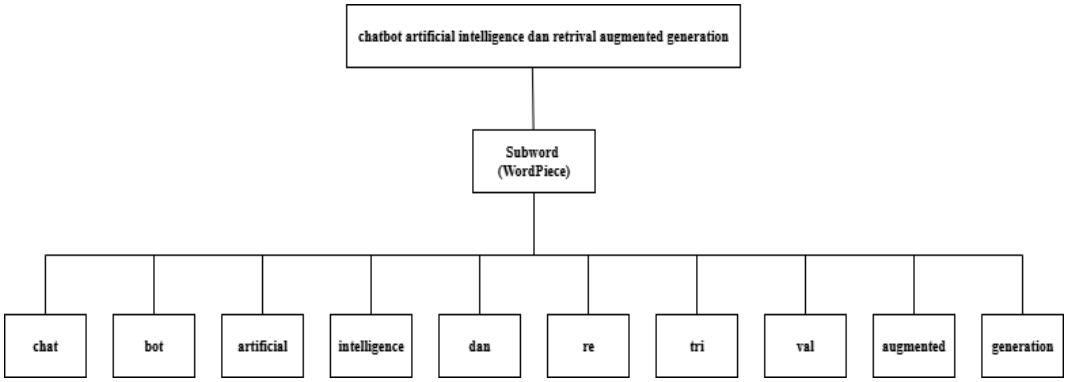
Chatbot adalah program komputer yang memberikan layanan interaksi dengan pengguna melalui percakapan teks. *Chatbot* menggantikan peran manusia dalam melayani pembicaraan melalui aplikasi pesan dengan menjawab setiap kalimat yang dikirimkan pengguna dan berinteraksi layaknya manusia berkat adanya kecerdasan buatan [24]. *Chatbot* termasuk dalam program komputer independen yang dapat diintegrasikan ke berbagai platform *open source*. *Chatbot* juga merupakan salah satu bentuk pengaplikasian dari sistem cerdas melalui NLP.

Chatbot dapat dikategorikan menjadi tiga jenis utama, yaitu *pattern-based chatbot*, *retrivalbased chatbot*, dan *generative-based chatbot*. *Pattern-based chatbot* bekerja berdasarkan aturan atau pola yang telah ditentukan sebelumnya, sehingga responsnya terbatas pada skenario yang diatur. *Retrivalbased chatbot* menggunakan teknik pemrosesan nlp untuk memahami maksud pengguna dan mencocokkannya dengan jawaban yang paling relevan dari basis data. Sementara itu, *generative-based chatbot* menggunakan model *pre-trained* untuk menghasilkan respons secara dinamis sehingga lebih fleksibel dalam menjawab pertanyaan kompleks meskipun berisiko menghasilkan informasi yang kurang akurat [25].

2.1.10 *Tokenisasi*

Tokenisasi adalah proses dasar dalam pemrosesan bahasa alami (*Natural Language Processing*/NLP) yang bertujuan untuk memecah teks mentah menjadi satuan-satuan lebih kecil yang disebut token. Token dapat berupa kata, frasa pendek, potongan kata (*subword*), atau bahkan karakter, tergantung pada pendekatan dan algoritma yang digunakan. Tujuan utama dari *tokenisasi* adalah untuk mengubah representasi teks menjadi bentuk yang dapat diproses lebih lanjut oleh sistem komputer, khususnya dalam model pembelajaran mesin atau model bahasa berbasis *transformer*.

Dalam konteks model bahasa modern seperti BERT atau *Sentence Transformer*, *tokenisasi* berfungsi sebagai langkah awal sebelum teks dikonversi ke dalam representasi numerik. Model-model ini tidak menerima teks dalam bentuk asli, melainkan membutuhkan *input* berupa urutan angka yang merepresentasikan token sesuai dengan *vocabulary* atau kamus token yang telah ditentukan sebelumnya. *Tokenisasi* juga berperan penting dalam menangani kata-kata baru atau tidak umum. Oleh karena itu, banyak model menggunakan metode *tokenisasi* berbasis *subword* seperti *WordPiece*, *Byte Pair Encoding* (BPE), atau *Unigram* yang memungkinkan pemecahan kata menjadi bagian-bagian lebih kecil agar tetap bisa dikenali oleh model meskipun tidak ditemukan secara utuh dalam kamus (*vocabulary)* yang digunakan. *Tokenisasi* ini menjadi samgat penting karena model tidak dapat menerima *input* dalam bentuk teks mentah, melainkan dalam bentuk urutan token numerik sesuai dengan *vocabulary* yang digunakan.



Gambar 5. Ilustrasi *Tokenisasi* Berbasis *Subword* [26]

Gambar 5 menunjukkan ilustrasi dari *tokenisasi* berbasis Subword dengan menggunakan model *WordPiece*, di mana kalimat "*chatbot* artificial intelligence dan *retrieval* augmented generation" dipecah menjadi potongan-potongan kata (*subword*). Dalam proses ini, kata "*chatbot*" dipecah menjadi "chat" dan "bot", sedangkan "*retrieval*" dipecah menjadi "re", "tri", dan "val", karena bentuk utuhnya tidak tersedia dalam *vocabulary* model. Sebaliknya, kata-kata seperti "artificial", "intelligence", "dan", "augmented", dan "generation" tetap utuh karena sudah tersedia dalam *vocabulary*. Teknik ini memungkinkan model tetap dapat

memproses kata-kata baru atau jarang muncul dengan mengandalkan bagian-bagian kata yang umum dan telah dikenal sebelumnya.

2.1.11 *Embedding*

Embedding adalah proses yang digunakan untuk mengubah teks menjadi representasi numerik dalam bentuk angka. Proses ini penting karena algoritma ML tidak dapat memproses data dalam bentuk *string* atau teks biasa, sehingga diperlukan angka sebagai *input* [27]. Dalam konteks *Natural Language Processing* (NLP), *embedding* merepresentasikan kata-kata dalam bentuk vektor berdimensi tinggi yang dirancang untuk membantu model AI memahami, memproses, dan menganalisis teks. Setiap kata direpresentasikan sebagai vektor dalam ruang berdimensi, di mana kata-kata dengan makna atau konteks serupa memiliki posisi yang berdekatan [28]. Representasi ini tidak hanya mengubah kata menjadi angka, tetapi juga mencerminkan hubungan semantik antar kata sehingga mampu menghitung kemiripan kata berdasarkan nilai-nilai vektornya.

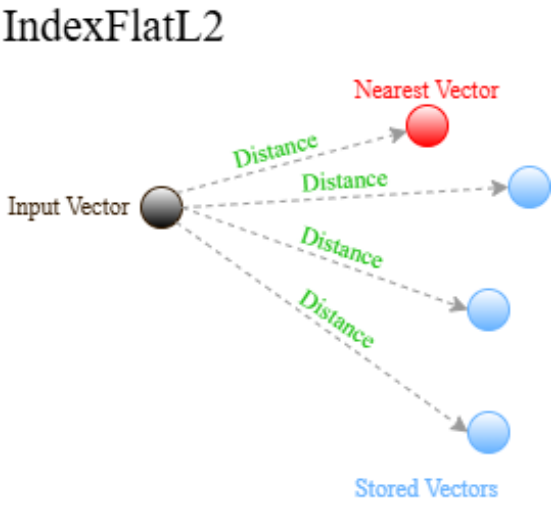
2.1.12 *Indexing*

Indexing adalah proses menyimpan dan mengatur data agar dapat dicari kembali dengan cepat dan efisien. Dalam konteks sistem berbasis teks atau bahasa alami, *indexing* dilakukan setelah data teks diubah terlebih dahulu menjadi bentuk vektor melalui proses *embedding*. Vektor-vektor ini mewakili makna dari kalimat atau kata dalam bentuk angka-angka, sehingga komputer dapat melakukan pencarian berdasarkan kemiripan makna, bukan hanya kecocokan kata secara langsung [29].

Melalui *indexing*, seluruh vektor yang telah dihasilkan akan disusun ke dalam suatu struktur data tertentu yang memungkinkan sistem untuk menemukan informasi yang paling relevan dalam waktu singkat. Misalnya, ketika pengguna mengajukan pertanyaan, sistem akan mencocokkannya dengan vektor-vektor yang sudah ada dan mencari yang paling mirip secara semantik. Salah satu tools yang banyak

digunakan untuk *indexing* adalah FAISS (*Facebook AI Similarity Search*). FAISS dirancang untuk menangani jutaan vektor dengan sangat efisien, dan sering digunakan dalam sistem berbasis *retrieval*, seperti *chatbot*, sistem pencarian dokumen, atau aplikasi tanya-jawab.

Dalam sistem FAISS, vektor-vektor yang mewakili pertanyaan disimpan di dalam sebuah struktur yang disebut indeks. Indeks ini merupakan komponen inti yang bertugas mengatur penyimpanan dan pencarian vektor berdasarkan tingkat kemiripannya. Pada penelitian ini, jenis indeks yang digunakan adalah *IndexFlatL2*, yaitu salah satu tipe indeks dalam FAISS yang menggunakan matriks jarak *Euclidean* (L2 *distance*) untuk mengukur kemiripan antar vektor. Semakin kecil jarak antara dua vektor, semakin besar tingkat kemiripan maknanya [30].



Gambar 6. Ilustrasi *Indexing*

Gambar 6 terlihat ilustrasi prinsip kerja *IndexFlatL2*, dimana titik hitam mewakili *input vector* (vektor dari pertanyaan pengguna), sedangkan titik-titik biru merupakan *stored vectors* (vektor-vektor dari *dataset*). Garis putus-putus menunjukkan jarak yang dihitung dari vektor *input* ke masing-masing vektor yang tersimpan. Vektor yang paling dekat ditandai dengan warna merah sebagai *nearest vector* yang menunjukkan vektor dengan makna atau konteks paling relevan

terhadap pertanyaan pengguna. Vektor-vektor terdekat inilah yang kemudian digunakan sebagai konteks untuk menghasilkan jawaban melalui LLM.

Dalam implementasinya, seluruh pertanyaan dalam dataset terlebih dahulu diubah menjadi vektor numerik menggunakan *Sentence Transformer*, lalu dimasukkan ke dalam indeks FAISS. Ketika pengguna mengajukan pertanyaan baru, sistem akan mengubahnya menjadi vektor dan mencocokkannya dengan semua vektor dalam indeks. *IndexFlatL2* kemudian menghitung jarak antara vektor *input* pengguna dengan seluruh vektor yang tersimpan, dan mengembalikan beberapa vektor terdekat yang memiliki kemiripan tertinggi. Proses inilah yang menjadi dasar utama dalam pengambilan konteks pada sistem *chatbot* RAG.

2.2 Tools Yang Digunakan

2.2.1 *Python*

Python adalah salah satu bahasa pemrograman yang populer dan banyak digunakan oleh perusahaan besar maupun *developer* untuk mengembangkan berbagai jenis aplikasi berbasis desktop, web, dan *mobile* [31]. Bahasa ini memiliki sejumlah keunggulan, seperti sifatnya yang *open source*, kemampuan untuk berjalan di berbagai platform, dan sintaksis yang sederhana sehingga relatif mudah dipahami oleh pemula maupun profesional. Selain itu, *python* juga dilengkapi dengan beragam *library* yang sangat membantu dalam mempercepat pengembangan aplikasi. *Library* ini berisi kumpulan kode yang siap digunakan untuk menghemat waktu dan usaha dalam menulis kode dari awal, menjadikan *python* sebagai pilihan utama dalam berbagai proyek teknologi modern.

2.2.2 *Visual Studio Code*

Visual Studio Code (VS Code) adalah perangkat lunak kode editor yang dikembangkan oleh Microsoft dan pertama kali dirilis pada tahun 2015. *VS Code* dirancang untuk mendukung pengembangan perangkat lunak lintas platform, sehingga dapat dijalankan di berbagai sistem operasi seperti Windows, macOS, dan

Linux [32]. *VS Code* memiliki antarmuka yang ramah pengguna dan mendukung berbagai bahasa pemrograman seperti *Java, JavaScript, Python*, C, C++, dan masih banyak lagi. Selain itu, *VS Code* dilengkapi dengan fitur-fitur unggulan seperti *IntelliSense* yang dapat memberikan saran penulisan kode, *debugging* terintegrasi, terminal bawaan, dan dukungan Git. *VS Code* juga menyediakan ribuan ekstensi gratis yang memungkinkan pengguna menyesuaikan fitur dan tampilan sesuai dengan kebutuhan proyek mereka, menjadikan *VS Code* salah satu kode editor favorit di kalangan pengembang di seluruh dunia.

2.2.3 FAISS

FAISS (*Facebook AI Similarity Search*) adalah mesin *database* vektor *open source* yang dirancang khusus untuk menyimpan, mencari, dan mengelola data yang disimpan dalam bentuk vektor, seperti hasil *embedding* dari model kecerdasan buatan (AI)**.** FAISS memungkinkan penyimpanan vektor dalam jumlah besar yang sering digunakan dalam aplikasi kecerdasan buatan untuk analisis teks. FAISS menyediakan layanan siap pakai yang mudah diintegrasikan ke dalam aplikasi dengan pustaka yang dapat mengelola vektor dengan cepat dan efisien [33].

2.2.4 *OpenRouter*

OpenRouter merupakan sebuah platform yang menyediakan layanan akses terhadap berbagai model *Large Language Model* (LLM) dari berbagai penyedia melalui satu antarmuka pemrograman aplikasi (API) yang terintegrasi. Dengan menggunakan *OpenRouter*, pengguna tidak perlu melakukan registrasi atau integrasi secara langsung ke masing-masing penyedia model, karena seluruh akses dapat dilakukan melalui satu API *key* yang disediakan oleh *OpenRouter*. Platform ini bertujuan untuk memudahkan penggunaan beragam LLM, baik untuk kebutuhan percobaan, pengembangan, maupun produksi, dengan proses yang lebih sederhana dan efisien. Selain itu, *OpenRouter* juga menawarkan fleksibilitas harga serta ketersediaan model-model tertentu yang dapat diakses secara gratis [34].

2.2.5 *Pre-Trained* Model *Qwen*

Dalam kecerdasan buatan (AI), model merujuk pada sistem matematis atau algoritma yang dilatih untuk memahami pola, struktur, atau hubungan dalam data. Model ini digunakan untuk memprediksi hasil atau menghasilkan *output* berdasarkan data *input* yang diberikan. Sementara itu *pre-trained* model yang sudah dilatih sebelumnya menggunakan *dataset* besar, sehingga model ini sudah memiliki pemahaman yang umum mengenai suatu pola. *Pre-trained* model tidak langsung siap untuk digunakan pada tugas spesifik tetapi sudah cukup pintar untuk dapat diadaptasi ke berbagai aplikasi [35].

Dalam bidang AI, terdapat banyak *pre-trained* model yang populer dan digunakan secara luas untuk berbagai aplikasi, salah satunya adalah model *Qwen* (Tongyi Qianwen) yang dikembangkan oleh Alibaba *Cloud*. *Qwen* merupakan model bahasa generatif berbasis arsitektur *transformer* yang diluncurkan pada awal tahun 2025. Model ini dirancang untuk mendukung berbagai tugas pemrosesan bahasa alami, seperti penerjemahan dan pengembangan *chatbot*. Sejak diluncurkan, *Qwen* terus mengalami peningkatan, baik dari segi ukuran model maupun kemampuannya dalam memahami konteks dan menghasilkan respons. Dalam pengembangan *chatbot*, *Qwen* menawarkan keunggulan dalam pemrosesan bahasa yang efisien dan kemampuan memahami percakapan multi-bahasa [36]. Hal ini menjadikan *Qwen* sebagai salah satu alternatif model yang menjanjikan, terutama bagi pengembang yang mencari model *open source* dengan performa tinggi untuk menciptakan pengalaman interaksi yang lebih alami, responsif, dan adaptif.

2.2.6 RAGAS *Scores*

RAGAS (*Retrival Augmented Generation Assessment*) adalah sebuah *framework* yang digunakan untuk mengevaluasi sistem RAG tanpa memerlukan referensi manual. RAGAS memberikan analisis mendalam tentang kualitas jawaban yang dihasilkan dengan mempertimbangkan relevansi konteks dan keakuratan informasi [37]. Berbeda dengan *framework* perhitungan lainnya, RAGAS tidak hanya fokus

pada kesamaan teks, tetapi juga pada sejauh mana jawaban mendukung pertanyaan dengan data yang diambil dari sumber yang relevan. Keunggulan utama RAGAS adalah kemampuannya untuk memberikan evaluasi otomatis yang komprehensif terhadap sistem RAG yang sering kali menghasilkan jawaban bervariasi karena terintegrasi dengan model generatif. Pengujian ini bertujuan untuk mengetahui sejauh mana sistem mampu memberikan jawaban yang benar, relevan, dan berbasis pada data yang tersedia. Adapun matriks-matriks yang digunakan dalam RAGAS adalah sebagai berikut:

a) *Faithfulness*

Matriks ini digunakan untuk mengukur tingkat konsistensi faktual antara jawaban yang diberikan oleh *chatbot* dengan konteks yang diambil dari hasil *retrieval*. Sebuah jawaban dianggap faithful jika seluruh klaim atau informasi yang disampaikan dapat didukung atau dibuktikan oleh konteks yang diberikan, tanpa menyimpang, menambah, atau mengubah fakta.

b) *Answer Relevancy*

Matriks ini digunakan untuk mengevaluasi seberapa tepat dan relevan jawaban yang dihasilkan oleh *chatbot* terhadap pertanyaan yang diajukan oleh pengguna. Matriks ini memastikan bahwa jawaban yang diberikan bukan hanya secara teknis benar, tetapi juga langsung menjawab inti dari pertanyaan tanpa mengandung informasi yang tidak dibutuhkan.

c) *Context Recall*

Context Recall mengukur sejauh mana informasi penting dalam konteks yang diambil (retrieved *Context*) benar-benar digunakan untuk membentuk jawaban. Dalam evaluasi ini, semakin banyak bagian penting dari konteks yang tercermin dalam jawaban, semakin tinggi skor *recall*-nya [38].

Dalam pengujian ini skor 0.8 (atau 80%) dijadikan sebagai ambang batas minimal **(***threshold***)** untuk setiap matriks [38]. Jika skor di bawah nilai tersebut maka dilakukan proses perbaikan *dataset*. Perbaikan ini meliputi penyempurnaan format jawaban, penghapusan informasi yang membingungkan, atau menyesuaikan gaya bahasa agar lebih mudah dipahami oleh model. Proses evaluasi ini dilakukan secara

iteratif, yaitu melalui perbaikan berulang terhadap *dataset* hingga seluruh matriks RAGAS mencapai atau melebihi skor ambang batas (*threshold*)**.** Pendekatan ini bertujuan untuk memastikan bahwa *chatbot* mampu memberikan jawaban yang relevan**,** akurat**,** serta berdasarkan data resmi yang bersumber dari *website* PPID Universitas Lampung. Berikut adalah hasil perhitungan untuk masing-masing matriks.

2.2.7 *Chatbot Usability Quesionnaire*

Chatbot Usability Questionnaire (CUQ) adalah instrumen evaluasi yang dirancang khusus untuk menilai tingkat kegunaan (usability) *chatbot* dari perspektif pengguna. CUQ dikembangkan berdasarkan prinsip user experience (UX) dengan fokus pada aspek utama interaksi pengguna dengan *chatbot*, meliputi kepribadian (*personality*), pengalaman awal penggunaan (*onboarding*), navigasi, pemahaman, kualitas respons, penanganan kesalahan (*error handling*), dan kecerdasan *chatbot* [39]. CUQ terdiri dari 16 pernyataan yang mencakup aspek positif maupun negatif dari pengalaman pengguna. Setiap pernyataan dinilai menggunakan skala *likert* 1 hingga 5. Skala *likert* yang digunakan dalam kuesioner ini memiliki rentang nilai dari 1 hingga 5, di mana skor 1 menunjukkan “sangat tidak setuju”, skor 2 “tidak setuju”, skor 3 “netral”, skor 4 “setuju”, dan skor 5 berarti “sangat setuju”. Total nilai maksimum dari seluruh pernyataan adalah 160 yang kemudian dinormalisasi menjadi skala 100 untuk memudahkan interpretasi hasil [40].

CUQ ini dirancang untuk menilai berbagai aspek pengalaman pengguna, meliputi kemudahan penggunaan, kejelasan jawaban, serta manfaat *chatbot* dalam memberikan informasi yang relevan dan dibutuhkan oleh pengguna. Kuesioner ini mencakup dua jenis pertanyaan, yaitu positif dan negatif. Pertanyaan positif menilai aspek yang mendukung pengalaman pengguna yang baik, seperti kenyamanan dan kepuasan, sementara pertanyaan negatif fokus pada potensi masalah atau hambatan yang mungkin ditemui pengguna, seperti kesulitan dalam memahami jawaban atau ketidaksesuaian informasi yang diberikan. Dengan demikian, CUQ mampu

memberikan gambaran yang komprehensif tentang sejauh mana *chatbot* memenuhi ekspektasi pengguna, dan hasil dari kedua jenis pertanyaan ini memungkinkan penilaian yang lebih mendalam tentang area yang perlu diperbaiki.

2.2.8 Telegram

Telegram adalah layanan pesan instan berbasis *cloud* yang memungkinkan pengguna untuk bertukar pesan, berbagi media dan file, serta melakukan panggilan suara atau video secara privat maupun dalam grup. Aplikasi ini didirikan pada tahun 2013 dan sejak itu berkembang menjadi salah satu platform komunikasi yang populer di seluruh dunia. Telegram tersedia di berbagai platform termasuk Android, iOS, Windows, macOS, Linux, dan browser web, menjadikannya sangat fleksibel untuk digunakan di berbagai perangkat.

Selain sebagai platform komunikasi, telegram juga menyediakan fitur yang memungkinkan pengembangan *chatbot*, yang dapat digunakan untuk berbagai tujuan [41]. *Chatbot* di telegram adalah program otomatis yang berinteraksi dengan pengguna melalui antarmuka obrolan. Bot ini dapat dirancang untuk menjalankan berbagai fungsi, seperti memberikan informasi, mengelola tugas, memproses pembayaran, atau bahkan mengintegrasikan layanan pihak ketiga. Kemampuan ini menjadikan telegram tidak hanya sebagai aplikasi pesan, tetapi juga sebagai alat serbaguna untuk mendukung bisnis, pendidikan, hingga pengembangan teknologi berbasis AI. Dengan API yang terbuka, telegram mempermudah pengembang untuk menciptakan bot yang dapat diakses langsung oleh banyak orang [42].

2.3 Penelitian Terkait

Dalam penelitian ini terdapat beberapa studi literatur yang digunakan sebacai acuan dalam penulisan penelitian dengan tujuan menambah pengetahuan sekaligus mengembangkan *insight* baru dari penelitian yang telah dilakukan sebelumnya.

Penelitian yang dilakukan oleh Pujiono, et al. pada tahun 2024 yang berjudul *Implementing Retrival Augmented Generation and Vector databases for Chatbots in Public Services Agencies Context* membahas mengenai penerapan teknologi RAG dalam pengembangan *chatbot* untuk layanan *Financial Advisor* Badan Layanan Umum (BLU). Penelitian ini membandingkan beberapa model *pre-trained* yaitu Davinci-002, Babbage-002, GPT-3.5 Turbo, dan GPT-4 dengan fokus pada peningkatan efisiensi *chatbot* dalam merespons permintaan informasi pengguna secara akurat dan relevan. RAG digunakan untuk menggabungkan kemampuan pencarian informasi dari berbagai sumber eksternal dengan kemampuan pembuatan teks dari model bahasa besar (LLM) yang memungkinkan *chatbot* memberikan jawaban berdasarkan data yang disimpan dalam basis data vektor. Hasil dari penelitian ini menunjukkan bahwa model GPT-4 memiliki nilai rata-rata *cosine similarity* tertinggi yang menandakan kinerja terbaik dalam menghasilkan jawaban yang relevan dan akurat dibandingkan dengan model lainnya. [43] .

Selanjutnya penelitian yang dilakukan oleh Irfan pada tahun 2024 dengan judul *Penerapan Retrival Augmented Generation Menggunakan LangChain dalam Pengembangan Sistem Tanya Jawab Hadis Berbasis Web* berfokus pada pengembangan sistem tanya jawab hadis berbasis web menggunakan RAG dengan *framework LangChain* dan model *pre-trained* GPT-4. Penelitian ini berhasil meningkatkan kualitas sistem tanya jawab hadis dengan mengintegrasikan pencarian berdasarkan sumber dari 9 kitab hadis. Hasil pengujian menunjukkan performa yang baik, dengan indeks kepuasan pengguna mencapai 89,4%, menandakan "sangat setuju" terhadap jawaban yang dihasilkan [44].

Kemudian penelitian yang dilakukan oleh Qingqing Zhou et al. tahun 2024 dengan judul *Gastrobot: A Chinese Gastrointestinal Disease Chatbot Based on The Retrival Augmented Generation* membahas pengembangan GastroBot, *chatbot* yang dirancang untuk menyediakan informasi klinis terkait penyakit *gastrointestinal* menggunakan metode RAG. Penelitian ini memanfaatkan model *pre-prained* GPT-3.5 Turbo yang ditambahkan dengan data 25 pedoman klinis dan 40 artikel literatur terkini dari Tiongkok untuk meningkatkan relevansi dan akurasi

jawaban. GastroBot dinilai unggul dalam evaluasi performa mencapai relevansi jawaban 92,28% yang diukur dengan RAGAS *Scores*. Penelitian ini menyoroti potensi RAG untuk meningkatkan aplikasi klinis LLM, khususnya di bidang *gastroenterology* yang memberikan solusi inovatif untuk mendukung diagnosis serta edukasi medis [45].

Selanjutnya terdapat penelitian yang dilakukan oleh Melanie M. Cooper dan Michael W. Klymkowsky pada tahun 2024 dengan judul *Let Us Not Squander the Affordances of LLMs for the Sake of Expedience: Using Retrieval Augmented Generative AI Chatbots to Support and Evaluate Student Reasoning* membahas mengenai pengembangan *chatbot* RAG sebagai inovasi dalam mendukung pembelajaran siswa sekaligus mengatasi keterbatasan model bahasa besar (LLM). Penelitian ini menggunakan model GPT-4 yang diintegrasikan dengan informasi dari berbagai sumber pembelajaran kimia untuk mendukung pengajaran berbasis *three-dimensional learning* (3DL) yang dapat membantu guru merancang tugas, mendukung penalaran siswa, dan mengevaluasi pembelajaran. [46].

Terakhir, penelitian yang dilakukan oleh Jing Miao et al. pada tahun 2024 dengan judul *Integrating Retrival Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications* menjelaskan mengenai integrasi sistem RAG dengan LLM yang menunjukkan potensi besar dalam mendukung aplikasi praktis di bidang nefrologi. Penelitian ini menggunakan model *pre-prained* GPT-4 yang menambahkan informasi panduan medis khusus yang selaras dengan pedoman klinis terbaru, seperti KDIGO (*Kidney Disease: Improving Global Outcomes*) tahun 2023 untuk penyakit ginjal kronis. Selain itu, model berbasis RAG ini mampu meningkatkan efisiensi keputusan klinis, mendukung pembelajaran berbasis kasus, dan membantu pengembangan pendidikan kedokteran. Studi ini memberikan contoh penggunaan RAG, termasuk penyempurnaan diagnosis dan pengelolaan pasien yang menunjukkan manfaatnya dalam menciptakan praktik medis yang lebih canggih dan akurat [47].

Beberapa penelitian relevan yang dijabarkan sebelumnya membantu proses pelaksanaan penelitian ini, khususnya dalam menentukan posisi penelitian. Studi terdahulu mengenai implementasi teknologi *chatbot* memberikan gambaran mengenai pendekatan yang paling efektif dalam pengembangan sistem serupa. Implementasi RAG yang digabungkan dengan LLM menunjukkan potensi yang besar dalam meningkatkan kinerja *chatbot* baik dalam konteks layanan publik, pendidikan, kesehatan maupun bidang lainnya. Dengan mengintegrasikan kemampuan *retrieval* (pencarian informasi) dari sumber eksternal dan *generation* (pembuatan teks), RAG memungkinkan *chatbot* untuk memberikan jawaban yang lebih akurat dan relevan sesuai dengan basis data. Teknologi ini mengurangi risiko "halusinasi" atau kesalahan informasi yang sering muncul pada LLM murni, serta mendukung proses pengambilan keputusan yang lebih tepat. Dengan demikian, penerapan RAG dalam LLM berpotensi menjadi solusi inovatif untuk berbagai tantangan dalam pengembangan sistem berbasis kecerdasan buatan yang efektif dan akurat.

BAB III
METODE PENELITIAN

3.1 Waktu dan Tempat

Waktu dan tempat pelaksanaan penelitian ini dilakukan pada:

1. Waktu penelitian : Desember 2024 sampai dengan Juli 2025
2. Tempat penelitian : Universitas Lampung

Tabel 1. Waktu Pelaksanaan Penelitian

No	Aktivitas	2024	2025					
		Des	Jan	Feb	Mar	Apr	Mei	Jun	Jul
1	Persiapan Penelitian								
2	Studi Literatur								
3	Pengumpulan Data								
4	Persiapan Data								
5	*Embedding*								
6	Integrasi *Pre-trained* Model								
7	Integrasi Telegram Bot								
8	Evaluasi								
9	Penulisan Laporan								

3.2 Alat dan Bahan Penelitian

3.2.1 Alat Penelitian

Penelitian ini menggunakan perangkat keras (*hardware*) dan perangkat lunak (*software*) dengan spesifikasi berikut.

Tabel 2. Alat Penelitian

No	Perangkat	Spesifikasi	Deskripsi
1	Laptop	Intel Core i5, RAM 8 GB, dan SSD 512 GB	Perangkat keras yang digunakan untuk mengembangkan *chatbot*.
2	*Python*	Versi 3.12.7	*Python* yang digukanan sebagai bahasa pemrograman dalam membangun model.
3	*Visual Studio Code*	-	Perangkat lunak kode editor untuk membuat kode program
4	FAISS	-	Mesin penyimpanan dan pencarian *database* vektor.
5	*OpenRouter*	-	Platform yang menyediakan layanan akses terhadap berbagai LLM
6	Telegram	Versi 5.8.0	Media pengimplementasian *chatbot.*
7	*Draw.io*	-	*Tools* untuk membuat diagram.
8	*Microsoft Word*	Versi 2021	Pendokumentasikan penelitian.
9	*Google Drive*	-	Tempat penyimpanan *dataset*.

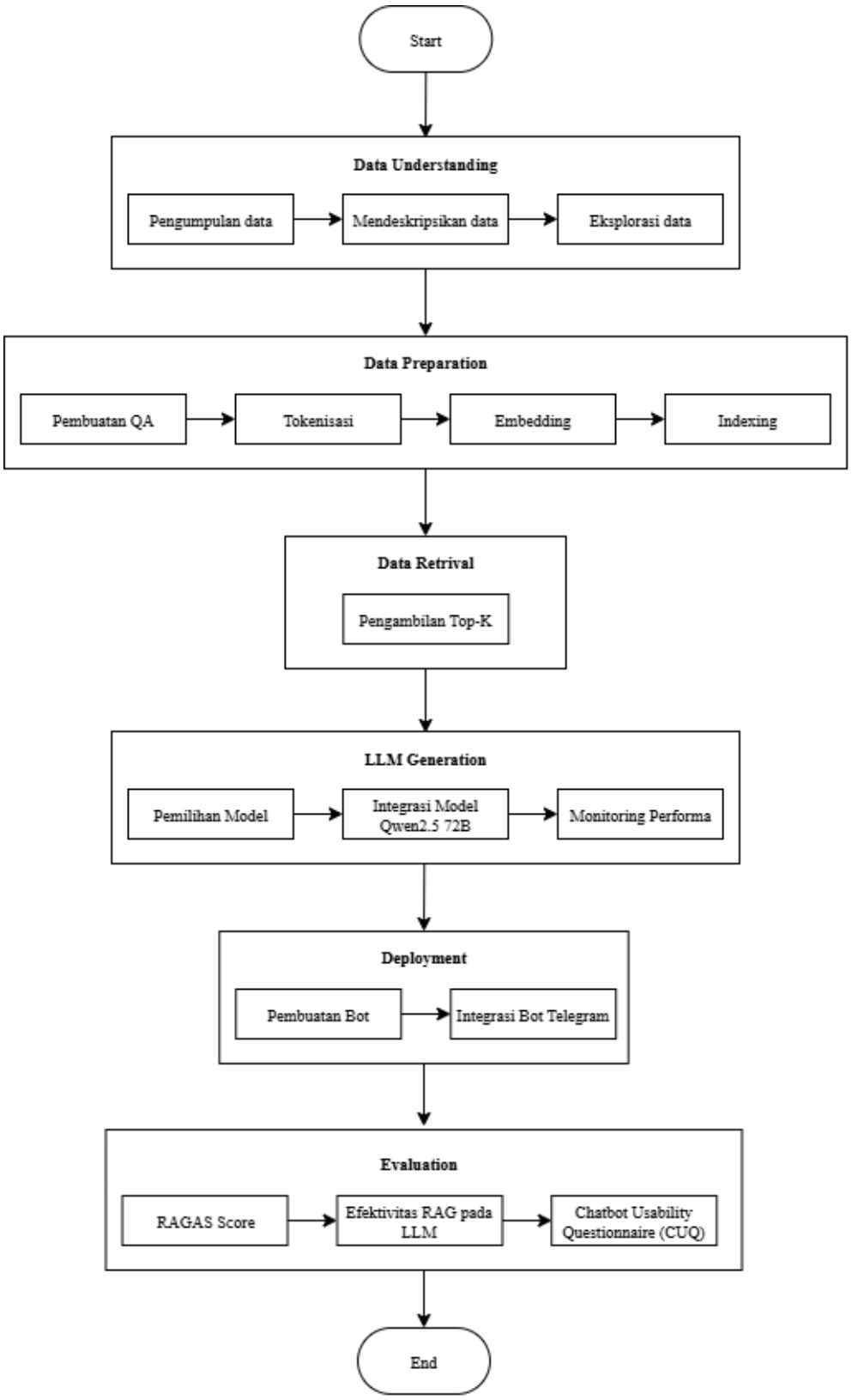
3.2.2 Bahan Penelitian

Pada penelitian ini, digunakan beberapa bahan penelitian untuk mendukung informasi dan proses pengembangan *chatbot* pada *website* PPID Unila, yaitu:

1. Beberapa penelitian terdahulu yang relevan.
2. Jurnal nasional dan internasional yang digunakan sebagai acuan penelitian.
3. Dokumentasi *website* PPID Unila.
4. API *pre-trained* model *Qwen 2.5 VL 72B Instruct*.
5. *Telegram bot.*

3.3 Tahapan Penelitian

Penelitian dimulai dengan studi literatur yang melibatkan pengumpulan dan pembelajaran ilmu pengetahuan dari berbagai sumber seperti jurnal, buku, artikel dan *website* yang memberikan dasar pengetahuan. Selanjutnya menyiapkan alat dan bahan yang akan digunakan untuk medukung penelitian ini. Kemudian dilakukan proses pengembangan *chatbot* melalui enam langkah utama, yaitu *data understanding***,** *data preparation***,** *data retrieval***,** LLM *generation***,** *deployment***,** dan *Evaluation***.** Setelah seluruh proses selesai, dilakukan analisis hasil untuk menarik kesimpulan dan memberikan saran yang dapat digunakan dalam penelitian selanjutnya. Seluruh tahapan penelitian dapat dilihat melalui diagram alir pada Gambar 7.



Gambar 7. Tahapan Penelitian

3.3.1 *Data Understanding*

Pada tahap ini melibatkan identifikasi sumber data, pengumpulan data, serta eksplorasi awal terhadap konten yang akan digunakan. Data dikumpulkan dari situs resmi PPID Universitas Lampung yang berisi berbagai dokumen dan informasi publik. Eksplorasi dilakukan untuk memahami karakteristik data yang mencakup jenis informasi yang tersedia, struktur data, dan kelengkapan atau kejelasan konten.

3.3.2 *Data Preparation*

Pada tahap ini dilakukan pembuatan dataset QA (*question–answer*) dengan menyusun pertanyaan yang relevan berdasarkan isi dokumen dan membuat jawaban yang sesuai. Dataset kemudian dibersihkan dari data yang tidak relevan atau redundan. Selanjutnya dilakukan proses teknis seperti *tokenisasi*, pembuatan *embedding*, *indexing* dengan FAISS, dan pengaturan mekanisme pencarian vektor berbasis kemiripan semantik. Semua proses ini bertujuan untuk menyiapkan data dalam format yang optimal untuk dimodelkan oleh sistem.

3.3.3 *Data Retrival*

Data *retrieval* merupakan proses pengambilan informasi relevan berdasarkan pertanyaan yang diajukan pengguna. Dalam sistem *chatbot* berbasis RAG ini, setiap pertanyaan diubah menjadi representasi vektor menggunakan model *embedding* yang sama dengan yang digunakan pada tahap *indexing*. Vektor pertanyaan tersebut kemudian dibandingkan dengan vektor dokumen yang telah disimpan dalam FAISS untuk menemukan potongan konteks yang paling relevan. Hasil pencarian ini berupa *top-k* konteks yang selanjutnya digunakan untuk membentuk *prompt* yang akan dikirimkan ke model LLM. Tahapan ini berperan penting dalam memastikan bahwa model memperoleh informasi yang akurat dan relevan sebelum menghasilkan jawaban.

3.3.4 LLM *Generation*

Pada tahap ini dilakukan proses pembangunan dan menerapkan LLM untuk menjawab pertanyaan pengguna berdasarkan hasil pencarian konteks. Konteks kemudian digabungkan dengan pertanyaan untuk menghasilkan jawaban menggunakan pendekatan RAG. Tahap ini juga mencakup pemantauan performa model untuk memastikan bahwa sistem bekerja secara optimal dan sesuai dengan harapan.

3.3.5 *Deployment*

Pada tahap *deployment*, model yang telah dibangun diimplementasikan ke dalam platform Telegram agar dapat digunakan secara langsung oleh pengguna. Proses *deployment* mencakup pembuatan bot, integrasi telegram dengan sistem RAG, serta visualisasi antarmuka *chatbot* PPID Universitas Lampung.

3.3.6 *Evaluation*

Pada tahap ini dilakukan proses evaluasi untuk mengukur efektivitas *chatbot* yang telah dikembangkan. Evaluasi dilakukan melalui tiga metode, yaitu pengujian menggunakan matriks RAGAS untuk menilai sejauh mana sistem mampu memberikan jawaban yang akurat, relevan, dan kontekstual berdasarkan dokumen PPID. Selanjutnya dilakukan perbandingan kualitas jawaban yang dihasilkan oleh model LLM murni (tanpa *retrieval*) dengan sistem *chatbot* berbasis RAG yang dikembangkan. Kemudian dilakukan juga penilaian dari sisi pengalaman pengguna menggunakan metode *Chatbot Usability Questionnaire* (CUQ), guna mengukur tingkat kepuasan, kenyamanan, dan kemudahan penggunaan *chatbot*. Evaluasi ini bertujuan untuk memastikan bahwa sistem tidak hanya unggul secara teknis, tetapi juga memenuhi aspek fungsionalitas dan kegunaan dalam konteks pelayanan informasi publik.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pengembangan *chatbot* PPID Universitas Lampung menggunakan *Indexing Augmented Generation* (RAG) dan *Large Language Model* (LLM), dapat disimpulkan sebagai berikut:

1. *Chatbot* PPID Universitas Lampung berhasil dikembangkan dan diintegrasikan ke Telegram menggunakan pendekatan RAG dan LLM *Qwen* 2.5 VL 72B dengan hasil evaluasi melalui metode CUQ yang menunjukkan bahwa sistem ini efektif dalam mempermudah akses informasi.
2. Performa *chatbot* menunjukkan hasil yang sangat baik dengan skor masing-masing matriks RAGAS yang berhasil melampaui angka 0,8 sebagai ambang batas minimum, yaitu *faithfulness* sebesar 0,9629, *answer relevancy* sebesar 0,8760, dan *Context recall* sebesar 0,8681. Selain itu, evaluasi CUQ oleh responden menghasilkan skor rata-rata 90,62 yang menandakan sistem berfungsi baik dan memberikan pengalaman pengguna yang positif.
3. Pendekatan RAG terbukti lebih efektif dalam mengurangi potensi halusinasi jawaban dibandingkan dengan penggunaan model LLM *Qwen* 2.5 VL 72B tanpa penambahan konteks. Hasil pengujian terhadap 20 pertanyaan acak menunjukkan bahwa *chatbot* yang dikembangkan mampu menjawaban semua pertanyaan dengan tepat dan sesuai dengan informasi yang tersedia. Sebaliknya, model LLM murni tanpa penambahan konteks hanya berhasil menjawab 4 pertanyaan dengan benar, sementara sisanya cenderung menghasilkan jawaban yang umum, tidak spesifik, dan menyimpang.

5.2 Saran

Berdasarkan hasil penelitian dan pengembangan *chatbot* PPID Universitas Lampung menggunakan *Indexing Augmented Generation* (RAG) dan *Large Language Model* (LLM), terdapat beberapa saran yang dapat dijadikan acuan untuk pengembangan lebih lanjut:

1. *Chatbot* dapat diintegrasikan lebih lanjut dengan berbagai sistem informasi di lingkungan Universitas Lampung, seperti sistem akademik, kepegawaian, dan pelayanan publik lainnya. Hal ini bertujuan agar *chatbot* dapat memberikan informasi yang lebih luas dan menyeluruh kepada pengguna.
2. Untuk menjaga akurasi dan relevansi jawaban yang diberikan *chatbot*, perlu dilakukan pembaruan rutin terhadap *dataset*, terutama saat terjadi perubahan kebijakan.
3. Untuk meningkatkan ketersediaan layanan, *chatbot* sebaiknya di-*deploy* ke platform *hosting* berbasis *cloud* agar dapat diakses kapan saja melalui Telegram tanpa bergantung pada perangkat pengembang yang harus selalu aktif.

DAFTAR PUSTAKA

[1] *Undang-Undang Dasar Republik Indonesia 14 Tahun 2008*. Kementrian Komunikasi dan Informatika RI, 2010. [Daring]. Tersedia pada: https://eppid.kominfo.go.id/storage/uploads/1_9_2-Undang_Undang_Nomor_14_Tahun_2008.pdf

[2] "Pejabat Pengelola Informasi dan Dokumentasi Universitas Lampung." [Daring]. Tersedia pada: https://ppid.unila.ac.id/profil/

[3] E. Yulianto, T. Murdianto, dan Al-Amin, "Peran Artificial Intelligence (AI) dalam Manajemen Arsip dan Dokumen," vol. 1, no. 6, hal. 484–499, 2024.

[4] T. Tourism *et al.*, "Pengembangan Virtual Assistant (Chatbot) Berbasis NLP (Natural Language Processing) Untuk Portal Informasi Terpadu Pariwisata Tasikmalaya," vol. 17, no. 1, hal. 136–148, 2025.

[5] R. Xu, Y. (Katherine) Feng, dan H. Chen, "ChatGPT vs. Google: A Comparative Study of Search Performance and User Experience," *SSRN Electron. J.*, no. July, 2023, doi: 10.2139/ssrn.4498671.

[6] A. F. Arz von Straussenburg dan A. Wolters, "Towards Hybrid Architectures: Integrating Large Language Models in Informative Chatbots," in *Wirtschaftsinformatik 2023*, 2023. [Daring]. Tersedia pada: https://aisel.aisnet.org/wi2023/9

[7] Muksalmina, *Rahasia Menguasai Artificial Intelligence: Kenali, Pelajari dan Taklukan Siapkan Dirimu untuk Era Artificial Intelligence*. Banda Aceh: Bandar Publishing, 2024.

[8] S. R. Sari *et al.*, "Peran Teknologi dalam Pengambilan Keputusan Bisnis: Integrasi Artificial Intelligence dalam Teori Pengambilan Keputusan," *Neraca Manajemen, Ekon.*, vol. 10, no. 1, 2024.

[9] S. Junaidi *et al.*, *Buku Ajar Machine Learning*. Jambi: PT. Sonpedia

Publishing Indonesia, 2024.

[10] R. Dastres dan M. Soori, "Artificial Neural Network," *Int. J. Imaging Robot.*, vol. 21, no. 2, hal. 13–25, 2021.

[11] O. A. Montesinos López, A. Montesinos López, dan J. Crossa, *Fundamentals of Artificial Neural Networks and Deep Learning*. 2022. doi: 10.1007/978-3-030-89010-0_10.

[12] H. A. Pratiwi, M. Cahyanti, dan M. Lamsani, "Implementasi Deep Learning Flower Scanner Menggunakan Metode Convolutional Neural Network," *Sebatik*, vol. 25, no. 1, hal. 124–130, 2021, doi: 10.46984/sebatik.v25i1.1297.

[13] Y. Hafifah, K. Muchtar, A. Ahmadiar, dan S. Esabella, "Perbandingan Kinerja Deep Learning Dalam Pendeteksian Kerusakan Biji Kopi," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 6, hal. 1928, 2022, doi: 10.30865/jurikom.v9i6.5151.

[14] H. Pratiwi, *Buku ajar kecerdasan buatan : disertai praktik baik pemanfaatannya*. Kalimantan Tengah: PT. Asadel Liamsindo Teknologi, 2024.

[15] E. S. Eriana dan D. A. Zein, *Artificial Intelligence (AI)*. Purbalingga: CV. Eureka Media Aksara, 2023.

[16] T. Chauhan, "NLP 101 - NLP Architecture Explained," E2E Cloud. [Daring]. Tersedia pada: https://www.e2enetworks.com/blog/nlp-101-nlp-architecture-explained

[17] M. M. Lopez dan J. Kalita, "Deep Learning applied to NLP," *Univ. Color. Color. Springs*, 2017.

[18] H. A. C. Utomo, Y. M. Saputra, dan A. Prasetiadi, "Implementasi Sistem Konfigurasi Router Berbasis Natural Language Processing dengan Pendekatan Low Rank Adaptation Finetuning dan 8-Bit Quantization," *J. Internet Softw. Eng.*, vol. 4, no. 2, hal. 1–7, 2023, doi: 10.22146/jise.v4i2.9093.

[19] "What Is a Large Language Model?," dataiku. [Daring]. Tersedia pada: https://www.dataiku.com/stories/detail/what-is-a-large-language-model/

[20] K. Mohiuddin *et al.*, "Attention Is All You Need," in *International Conference on Information and Knowledge Management, Proceedings*, 2017, hal. 4752–4758. doi: 10.1145/3583780.3615497.

[21] S. Zhao, Y. Yang, Z. Wang, Z. He, L. K. Qiu, dan L. Qiu, "*Indexing* Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely," *Microsoft Res. Asia*, 2024.

[22] K. Bourne, *Unlocking Data with Generative AI and RAG*. Barmingham: Packt Publishing Ltd, 2024.

[23] Y. Gao *et al.*, "*Indexing*-Augmented Generation for Large Language Models: A Survey," hal. 1–21, 2023, [Daring]. Tersedia pada: http://arxiv.org/abs/2312.10997

[24] E. Larasati Amalia dan D. Wahyu Wibowo, "Rancang Bangun Chatbot Untuk Meningkatkan Performa Bisnis," *J. Ilm. Teknol. Inf. Asia*, vol. 13, no. 2, hal. 137–142, 2019.

[25] H. T. Hien, C. Pham-Nguyen, N. H. N. Le, L. T. K. N. Ho, dan D. T. Le, "Intelligent Assistants in Higher-Education Environments: The FITEBot, a Chatbot for Administrative and Learning Support," 2018, *SoICT*.

[26] C. Si *et al.*, "Sub-Character Tokenization for Chinese Pretrained Language Models," *Trans. Assoc. Comput. Linguist.*, vol. 11, hal. 469–487, 2023, doi: 10.1162/tacl_a_00560.

[27] F. A. Nugraha, N. H. Harani, dan R. Habibi, *Analisis Sentimen Terhadap Pembatasan Sosial Menggunakan Deep Learning*. Kreatif, 2020.

[28] S. Ghannay, B. Favre, Y. Estève, dan N. Camelin, "Word *embedding*s evaluation and combination," *Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016*, hal. 300–305, 2016.

[29] J. R. Smith, M. Naphade, dan A. Natsev, "Multimedia semantic *indexing* using model vectors," *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 2, hal. 445–448, 2003, doi: 10.1109/ICME.2003.1221649.

[30] L. D. Krisnawati, A. W. Mahastama, S. C. Haw, K. W. Ng, dan P. Naveen, "Indonesian-English Textual Similarity Detection Using Universal Sentence

Encoder (USE) and Facebook AI Similarity Search (FAISS)," *CommIT J.*, vol. 18, no. 2, hal. 183–195, 2024, doi: 10.21512/commit.v18i2.11274.

[31] Muhammad Romzi dan B. Kurniawan, "Pembelajaran Pemrograman Python Dengan Pendekatan Logika Algoritma," *JTIM J. Tek. Inform. Mahakarya*, vol. 03, no. 2, hal. 37–44, 2020.

[32] B. Kurniawan dan M. Romzi, "Pembuatan dan Pelatihan Administrator Website pada Dinas Kesehatan Kabupaten Ogan Komering Ulu," *J. Pengabdi. Masy.*, vol. 2, no. 3, hal. 253–258, 2022, doi: 10.31004/abdira.v2i3.202.

[33] A. A. Nur Hakim, A. C. Murti, dan R. Nindyasari, "Implementasi Artificial Intelligence Dalam Sistem Pencarian Orang Hilang Dengan Face Recognition Studi Kasus Polres Kudus," *SKANIKA Sist. Komput. dan Tek. Inform.*, vol. 8, no. 1, hal. 168–180, 2025, doi: 10.36080/skanika.v8i1.3334.

[34] "OpenRouter," *Prompt*Layer. [Daring]. Tersedia pada: https://openrouter.ai/docs/quickstart

[35] A. Lee, "What Is a Pretrained AI Model?," NVIDIA. [Daring]. Tersedia pada: https://blogs.nvidia.com/blog/what-is-a-pretrained-ai-model/

[36] "*Qwen* LLMs," Alibaba Cloud. Diakses: 25 April 2025. [Daring]. Tersedia pada: https://www.alibabacloud.com/help/en/model-studio/what-is-*Qwen*-llm

[37] S. Es, J. James, L. Espinosa-Anke, dan S. Schockaert, "RAGAS: Automated Evaluation of *Indexing* Augmented Generation," *EACL 2024 - 18th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc. Syst. Demonstr.*, hal. 150–158, 2024.

[38] Shahul, "Evaluating RAG pipelines with Ragas and Openlayer," Openlayer. [Daring]. Tersedia pada: https://www.openlayer.com/blog/post/evaluating-rag-pipelines-with-ragas-and-openlayer?utm_source=chatgpt.com

[39] S. Holmes, A. Moorhead, R. Bond, H. Zheng, V. Coates, dan M. McTear, "Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?," *ECCE 2019 - Proc. 31st Eur. Conf. Cogn. Ergon. ' 'Design Cogn.*, hal. 207–214, 2019, doi:

10.1145/3335082.3335094.

[40] A. Alabbas dan K. Alomar, "A Weighted Composite Metric for Evaluating User Experience in Educational Chatbots: Balancing Usability, Engagement, and Effectiveness," *Futur. Internet*, vol. 17, no. 2, hal. 1–35, 2025, doi: 10.3390/fi17020064.

[41] I. P. G. A. Sudiatmika, "E-Learning Berbasis Telegram Bot," *KERNEL J. Ris. Inov. Bid. Inform. dan Pendidik. Inform.*, vol. 1, no. 2, hal. 49–60, 2021, doi: 10.31284/j.kernel.2020.v1i2.1469.

[42] N. Fernando, Humaira, dan E. Asri, "Monitoring Jaringan dan Notifikasi dengan Telegram pada Dinas Komunikasi dan Informatika Kota Padang," *JITSI J. Ilm. Teknol. Sist. Inf.*, vol. 1, no. 4, hal. 121–126, 2020, doi: 10.30630/jitsi.1.4.17.

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

[44] A. T. U. B. Lubis, N. S. Harahap, S. Agustian, M. Irsyad, dan I. Afrianty, "*Question Answer*ing System pada Chatbot Telegram Menggunakan Large Language Models (LLM) dan Langchain (Studi Kasus UU Kesehatan)," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, hal. 955–964, 2024, doi: 10.57152/malcom.v4i3.1378.

[45] Q. Zhou *et al.*, "GastroBot: a Chinese gastrointestinal disease chatbot based on the *indexing*-augmented generation," *Front. Med.*, vol. 11, no. May, 2024, doi: 10.3389/fmed.2024.1392555.

[46] M. M. Cooper dan M. W. Klymkowsky, "Let Us Not Squander the Affordances of LLMs for the Sake of Expedience: Using *Indexing* Augmented Generative AI Chatbots to Support and Evaluate Student Reasoning," *J. Chem. Educ.*, 2024, doi: 10.1021/acs.jchemed.4c00765.

[47] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. Garcia Valencia, dan W. Cheungpasitporn, "Integrating *Indexing*-Augmented Generation with Large

Language Models in Nephrology: Advancing Practical Applications," *Med.*, vol. 60, no. 3, hal. 1–15, 2024, doi: 10.3390/medicina60030445.