

# Implementasi *Retrieval Augmented Generation* dalam Sistem *Chatbot* Dermatologi Berbasis Website

<http://dx.doi.org/10.28932/jutisi.v11i3.12258>

Riwayat Artikel

Received: 23 Juni 2025 | Final Revision: 10 November 2025 | Accepted: 22 November 2025



Creative Commons License 4.0 (CC BY – NC)

Ivana Lucia Kharisma<sup>✉ #1</sup>, Muhammad Syarif Hidayat<sup>#2</sup>, Somantri<sup>#3</sup>, Kamdan<sup>#4</sup>

<sup>#</sup> Program Studi Teknik Informatika, Universitas Nusa Putra

Jl. Raya Cibolang Cisaat - Sukabumi No.21, Cibolang Kaler, Kec. Cisaat, Kab. Sukabumi, Jawa Barat 43152, Indonesia

<sup>1</sup>ivana.lucia@nusaputra.ac.id

<sup>2</sup>muhammad.syarif\_ti21@nusaputra.ac.id

<sup>3</sup>somantri@nusaputra.ac.id

<sup>4</sup>kamdan@nusaputra.ac.id

<sup>✉</sup>Corresponding author: ivana.lucia@nusaputra.ac.id

**Abstrak** — Indonesia yang beriklim tropis, memiliki sanitasi yang kurang memadai, serta terbatasnya akses terhadap layanan medis, khususnya di daerah terpencil, menjadi faktor utama penyebab tingginya kasus penyakit kulit. Akses langsung ke dokter spesialis kulit masih menjadi kendala bagi sebagian masyarakat. Penelitian ini bertujuan untuk mengembangkan *Chatbot* konsultasi penyakit kulit dengan pendekatan *Retrieval Augmented Generation* (RAG), menggunakan *framework* LangChain, model LLaMA, dan *database* vektor Qdrant. *Dataset* penelitian mencakup 30 jenis penyakit kulit yang diambil dari *National Library of Medicine*. Tahapan *preprocessing* data meliputi normalisasi spasi, penghapusan karakter non-standar, serta penanganan data kosong agar konsisten sebelum proses vektorisasi. Hasil evaluasi menunjukkan skor *Faithfulness* sebesar 0,9429 dan *LLMContextRecall* sebesar 0,9600, yang mencerminkan kualitas jawaban yang relevan dan sesuai dengan dokumen sumber. Namun, skor *Precision* yang masih rendah (0,4720) menunjukkan bahwa aspek ketepatan informasi perlu ditingkatkan. *Chatbot* ini diintegrasikan dengan platform *Chainlit* untuk menyediakan antarmuka pengguna yang interaktif, termasuk fitur *login*, histori percakapan, dan pemberian umpan balik, yang mendukung pengembangan sistem berbasis masukan pengguna. Waktu pengambilan informasi terbilang cepat (0,08–0,29 detik), meskipun proses pembuatan jawaban masih cukup lama karena keterbatasan penggunaan CPU (255–283 detik). Pengembangan ke depan disarankan mencakup peningkatan akurasi jawaban, optimalisasi performa model, penambahan referensi data medis, serta fitur validasi medis otomatis guna meningkatkan keandalan sistem. Dengan demikian, *Chatbot* ini berpotensi menjadi solusi alternatif yang terjangkau dan efektif untuk memberikan informasi awal terkait penyakit kulit bagi masyarakat dengan akses terbatas ke layanan medis.

**Kata kunci** — *Chainlit*; *Chatbot*; *LangChain*; *LlaMA*; *Retrieval Augmented Generation*.

## Implementation of *Retrieval Augmented Generation* in a Web-Based Dermatological Chatbot System

**Abstract** — Indonesia's tropical climate, poor sanitation, and limited access to medical services especially in remote areas are key factors contributing to the high prevalence of skin diseases. Direct access to dermatologists remains difficult for many people. This study aims to develop a dermatological consultation Chatbot using a Retrieval Augmented Generation (RAG) approach, leveraging the LangChain framework, the LLaMA model, and the Qdrant vector database. The dataset includes 30 types of skin diseases sourced from the National Library of Medicine. The preprocessing stage involved whitespace normalization, removal of special characters, and handling of missing values to ensure data consistency before vectorization. Evaluation results showed high scores for Faithfulness (0,9429) and LLMContextRecall (0,9600), indicating that the responses were relevant and aligned with the source documents. However, a relatively low Precision score (0,4720) suggests a need for improved information accuracy. The Chatbot is integrated with the Chainlit platform, offering an interactive user interface that supports login, conversation history, and user feedback features to facilitate system development based on user input. The system demonstrated fast retrieval times (0,08–0,29 seconds), though answer generation remains slow due to CPU infrastructure limitations (255–283 seconds). Future improvements should focus on enhancing answer accuracy, optimizing the model's performance, enriching the medical reference dataset, and adding automated medical validation features to ensure the reliability of consultations. Therefore, this Chatbot system is expected to serve as a cost-effective and efficient alternative for providing initial information on skin conditions to individuals with limited access to healthcare services.

**Keywords** — Chainlit; Chatbot; LangChain; LLaMA; Retrieval Augmented Generation.

## I. PENDAHULUAN

Kulit, organ terbesar dan terluar tubuh manusia, berfungsi sebagai penerima sensasi seperti sentuhan, nyeri, dan suhu dan melindungi tubuh dari elemen luar. Kulit berfungsi sebagai pertahanan pertama tubuh dan sangat penting untuk menjaga homeostasis dan mencegah patogen masuk. Oleh karena itu, kulit harus dijaga dengan baik untuk menghindari masalah kulit yang dapat mempengaruhi kualitas hidup seseorang [1]. Penyakit kulit adalah kelainan pada jaringan kulit yang disebabkan oleh infeksi jamur, bakteri, parasit, atau virus. Penyakit ini dapat menyerang siapa saja, tanpa memandang usia, dan dapat menyerang sebagian atau seluruh tubuh. Jika tidak ditangani dengan tepat, penyakit kulit dapat memperburuk kondisi pasien dan bahkan menimbulkan risiko komplikasi lebih lanjut [2].

Lingkungan Indonesia dapat menyebabkan banyak masalah kulit, seperti kelembapan tinggi, paparan sinar matahari yang kuat, dan sanitasi lingkungan yang buruk, karena iklimnya yang tropis. Selain itu, keadaan ini akan diperburuk oleh keterbatasan akses terhadap layanan kesehatan dan rendahnya kesadaran masyarakat akan pentingnya menjaga kebersihan kulit. Sebagian besar penyakit kulit tidak mematikan, tetapi dapat menyebar dan menyebabkan ketidaknyamanan fisik dan mental. Biasanya diperlukan konsultasi langsung dengan dokter kulit untuk mendapatkan diagnosis yang akurat. Namun, beberapa individu tidak memiliki dana yang diperlukan untuk mendapatkan layanan ini, terutama di daerah terpencil [1].

Chatbot adalah program perangkat lunak yang dirancang untuk berinteraksi dengan orang menggunakan bahasa alami, terutama dengan meniru suara atau teks orang lain. Chatbot telah berkembang. Saat ini, Chatbot memiliki banyak fitur, yang menghasilkan istilah baru seperti *conversational interface*, *multimodal agent*, dan *chatterbots*. Chatbot telah mengubah interaksi antara manusia dan komputer [3]. Kemampuan Chatbot telah ditingkatkan berkat kemajuan *Generative AI*. Chatbot telah terbukti sangat bermanfaat bagi industri kesehatan karena dapat memenuhi kebutuhan individu dengan menyediakan layanan yang disesuaikan, yang dapat menghasilkan manajemen perawatan yang lebih efisien dan hasil kesehatan yang lebih baik. Chatbot ini dapat memberi tahu orang tentang efek samping, mengingatkan mereka untuk minum obat, dan bahkan membantu menjadwalkan janji temu dokter [4].

Metode pemrosesan bahasa alami dan pembelajaran mesin yang disebut *Retrieval Augmented Generation* (RAG) menggabungkan keunggulan model *generative* dan *retrieval* untuk meningkatkan kualitas teks yang dihasilkan. Ini adalah salah satu teknologi *Generative AI* yang sangat bermanfaat. Membuat agen percakapan, menjawab pertanyaan, dan merangkum dokumen adalah beberapa contoh aplikasinya. Berkat kemampuan sistem RAG untuk mengakses *database* medis eksternal secara *real-time*, *Large Language Models* (LLM) dapat memberikan *responses* berdasarkan penelitian terbaru, panduan klinis, dan informasi obat dalam bidang medis yang selalu berubah. Dua komponen utama RAG adalah *retrieval* dan *generasi*. Komponen generasi bertanggung jawab untuk mengumpulkan informasi yang relevan dari berbagai sumber pengetahuan, sedangkan komponen *retrieval* menghasilkan tanggapan yang mempertimbangkan konteks dokumen yang diambil, yang menghasilkan jawaban yang lebih akurat, informatif, dan relevan secara kontekstual [5].

Chainlit adalah *framework* multifungsi yang bertujuan untuk membuat proses pembuatan antarmuka web untuk agen percakapan lebih mudah. Chainlit, sebuah lembaga *open-source*, menggunakan kerangka kerja React untuk mengintegrasikan Chatbot dengan model bahasa besar (LLM) [6].

Beberapa penelitian sebelumnya telah mengeksplorasi penggunaan metode RAG dalam berbagai konteks. Misalnya, penelitian oleh Muhammad Irfan Syah et al. [7] menggunakan RAG, LangChain, dan Chainlit untuk membuat sistem tanya jawab hadis berbasis web, meskipun menggunakan model GPT-4-1106-preview dan *database* ChromaDB. Harsh Jain [8] mengembangkan Chatbot medis berbasis LLaMA 2 13B yang mampu menjawab pertanyaan secara efektif. Studi oleh Jing

Miao et al. [5] mengintegrasikan RAG dan *zero-shot prompt engineering* dalam bidang nefrologi dan berhasil mencapai akurasi hingga 99,25% dalam tugas perangkuman. Penelitian oleh Wenqi Shi et al. [9] memanfaatkan pendekatan *Chain-of-thought prompting* dalam pengambilan keputusan bersama untuk pasien skoliosis. Sementara itu, Ahmet Yusuf Alan et al. [10] menggunakan RAG untuk mengembangkan sistem tanya jawab Islam dan menunjukkan keunggulan dibandingkan model lain seperti ChatGPT. Penelitian lainnya oleh Surajit Sarma dan Nabankur Pathak [11] berfokus pada pengembangan *Chatbot* AI menggunakan pendekatan *deep learning* Bi-LSTM dan *attention mechanism*.

Penelitian ini memiliki beberapa perbedaan signifikan dari studi-studi sebelumnya. Sistem LLM yang digunakan adalah LLaMA 3.1, *database* vektor yang digunakan adalah Qdrant, dan antarmuka Chainlit dan *framework* LangChain digunakan. Tidak seperti metode manual atau *BERTScore*, evaluasi dilakukan menggunakan metode otomatis RAGAS. Selain itu, konsultasi penyakit kulit yang diangkat masih jarang diteliti dengan *Chatbot* berbasis RAG.

Penelitian ini memiliki nilai kebaruan (*novelty*) dalam hal pendekatan, arsitektur sistem, dan domain aplikasi karena menggabungkan berbagai elemen teknologi terbaru dan fokus pada topik yang relevan dan kontekstual di Indonesia. Penelitian ini berbeda dari penelitian sebelumnya karena topik penyakit kulit, integrasi teknologi LLaMA 3.1, LangChain, Qdrant, dan Chainlit, serta sistem evaluasi otomatis RAGAS dan antarmuka interaktif.

Dengan menggunakan pendekatan *Retrieval Augmented Generation* yang menggabungkan model LLaMA 3.1, *database* Qdrant, *framework* LangChain, dan antarmuka Chainlit, penelitian ini bertujuan untuk mengembangkan *Chatbot* konsultasi kulit berbasis web yang memberikan informasi kesehatan kulit yang akurat, relevan, dan mudah diakses oleh masyarakat luas.

## II. METODE

Penelitian ini menggunakan pendekatan kualitatif deskriptif studi kasus. Metode *Retrieval Augmented Generation* (RAG) digunakan dalam penelitian ini untuk mengembangkan dan menerapkan sistem *Chatbot* berbasis web yang dapat digunakan untuk memberikan konsultasi tentang penyakit kulit. Metode ini digunakan untuk mempelajari proses perancangan sistem, pemanfaatan teknologi LLM (*Large Language Model*), dan bagaimana *Chatbot* yang dibangun berfungsi.

### A. Pengumpulan Data

Untuk mengumpulkan data, metode kualitatif digunakan, dan metode *Retrieval Augmented Generation* (RAG) digunakan untuk mendukung pengembangan *Chatbot* konsultasi penyakit kulit berbasis web. Studi literatur dan wawancara dengan ahli medis adalah beberapa metode yang digunakan untuk mendapatkan informasi tentang subjek penelitian. Data yang digunakan mencakup informasi tentang penggunaan *framework* LangChain, penggunaan model bahasa besar seperti LLaMA, dan penggunaan *database* vektor Qdrant. Selain itu, sebagai sumber data utama, menggunakan kumpulan data yang mencakup 30 jenis penyakit kulit yang diambil dari situs web resmi *National Library of Medicine* [12].

### B. Analisis Kebutuhan Sistem

Untuk membuat sistem *Chatbot* konsultasi kulit berbasis *Retrieval Augmented Generation* (RAG) berjalan dengan baik, perencanaan kebutuhan sistem harus dilakukan dengan baik. Ini akan memastikan bahwa proses pengembangan dan pelaksanaan sistem berjalan dengan lancar. Sistem ini menggabungkan berbagai komponen perangkat keras dan perangkat lunak untuk memastikan kinerja yang *responsif* dan kemampuan untuk memberikan jawaban yang relevan terhadap pertanyaan pengguna.

Tabel 1 menggambarkan berbagai perangkat lunak yang digunakan untuk membuat sistem RAG *Chatbot* Konsultasi Kulit Berbasis Web dan fungsi masing-masing. Untuk memastikan sistem beroperasi secara optimal, setiap perangkat lunak memiliki peran penting, mulai dari tahap eksperimen dan pengembangan awal, proses penyimpanan dan pencarian data berbasis vektor, hingga penyediaan antarmuka pengguna yang interaktif.

TABEL 1  
KEBUTUHAN PERANGKAT LUNAK

| No | Nama                     | Perangkat Lunak | Fungsi/Kegunaan  |
|----|--------------------------|-----------------|--|
| 1  | Jupyter Notebook         |                 | Untuk logika sistem secara interaktif, eksperimen awal, prototyping, dan pengujian model   |
| 2  | Visual Studio Code       |                 | Lingkungan pengembangan utama yang mengintegrasikan RAG dan memungkinkan <i>Chatbot</i> untuk menulis, mengedit, dan mengelola kode proyek |
| 3  | Chainlit                 |                 | Antarmuka pengguna berbasis web untuk berinteraksi dengan <i>Chatbot</i> secara langsung   |
| 4  | Google Cloud             |                 | Infrastruktur <i>cloud</i> yang digunakan untuk menjalankan model dan menyimpan data secara skalabel                                       |
| 5  | LangChain                |                 | <i>Framework</i> yang menangani alur kerja RAG ( <i>embedding</i> , <i>retrieval</i> , LLM) secara modular dan terintegrasi                |
| 6  | bge-m3 (embedding model) |                 | Mengkonversi teks ke dalam bentuk vektor numerik untuk pencocokan semantik dalam proses <i>retrieval</i>                                   |
| 7  | Qdrant                   |                 | <i>Database</i> vektor yang menyimpan dan mencari dokumen berbasis representasi <i>embedding</i> untuk keperluan pencarian semantik        |

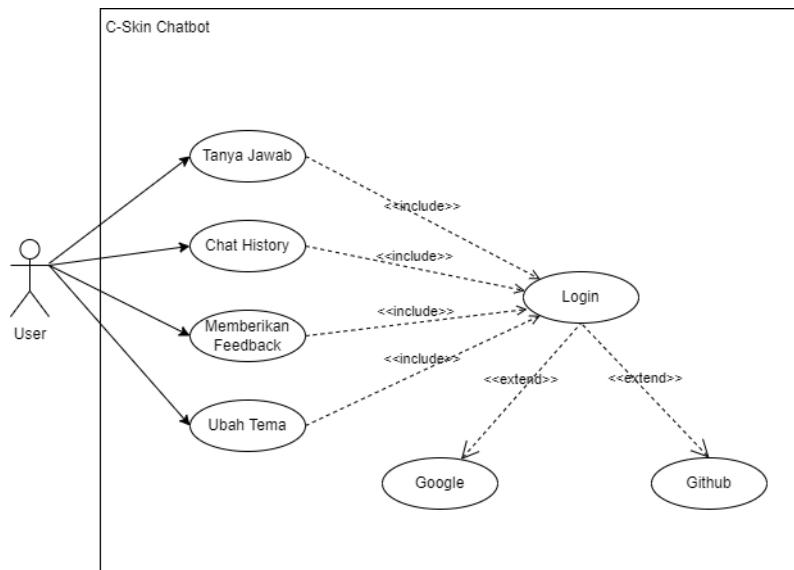
| No | Nama         | Perangkat Lunak | Fungsi/Kegunaan   |
|----|--------------|-----------------|---|
| 8  | LLaMA 3.1 8B |                 | Model Bahasa Besar (LLM) yang menghasilkan jawaban berdasarkan pertanyaan dan konteks hasil pencarian dokumen |

Tabel 2 menunjukkan bahwa *server* virtual berbasis *cloud* penelitian ini menggunakan mesin tipe *n2-standard-8* yang memiliki 8 vCPU dan RAM 32 GB. Ubuntu 22.04 LTS adalah sistem operasi yang dipilih karena kestabilannya dan biasanya digunakan untuk membuat aplikasi berbasis Linux. Media penyimpanannya adalah SSD berkapasitas 100 GB, yang dapat mendukung proses baca tulis data dengan cepat. Spesifikasi ini dipilih untuk memenuhi keterbatasan sumber daya komputasi dan memastikan proses pelatihan model, pengolahan data, dan pengujian sistem berjalan lancar.

TABEL 2  
KEBUTUHAN PERANGKAT KERAS

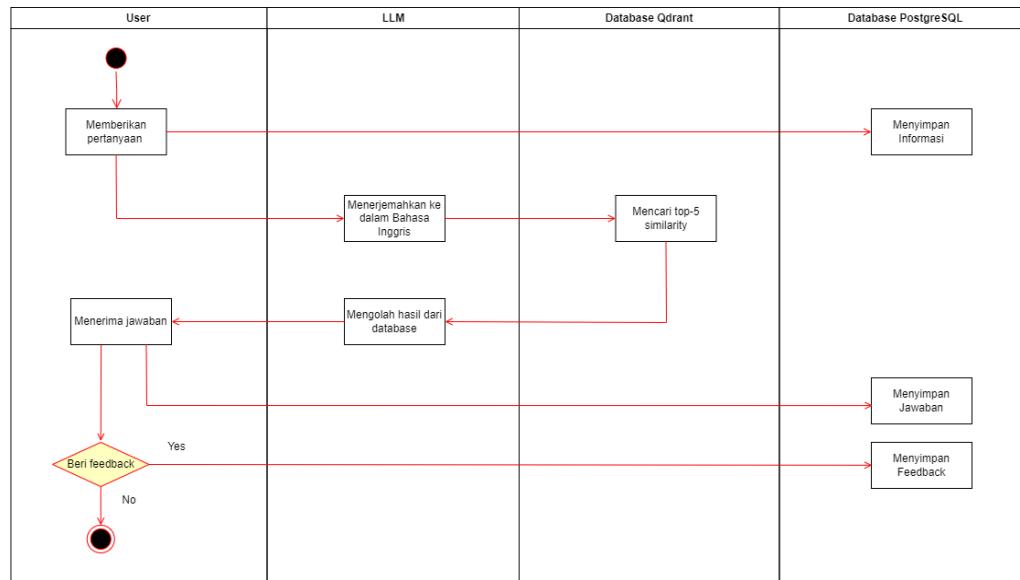
| No | Nama           | Spesifikasi          |
|----|----------------|----------------------|
| 1  | Jenis Mesin    | <i>n2-standard-8</i> |
| 2  | vCPU           | 8                    |
| 3  | RAM            | 32 GB                |
| 4  | Sistem Operasi | Ubuntu 22.04         |
| 5  | Penyimpanan    | SSD 100 GB           |

Penelitian ini menggunakan pendekatan pemodelan visual untuk menunjukkan fungsionalitas sistem. Ini memanfaatkan *Use Case Diagram*, yang menunjukkan interaksi antara sistem *Chatbot C-Skin* dan pengguna. Gambar 1 menunjukkan bahwa pengguna perlu melakukan proses autentikasi sebelum dapat mengakses fitur utama dari sistem, seperti mengajukan pertanyaan, melihat riwayat percakapan, memberikan umpan balik terhadap jawaban yang dihasilkan oleh *Chatbot*, serta melakukan pengaturan tema tampilan antarmuka. Proses login merupakan prasyarat untuk fitur-fitur utama lainnya, ditunjukkan melalui relasi *<<include>>*. Selain itu, sistem menyediakan dua metode autentikasi tambahan, yaitu melalui akun Google dan GitHub, yang direpresentasikan melalui relasi *<<extend>>*, memberikan fleksibilitas bagi pengguna dalam melakukan proses masuk ke dalam sistem.



Gambar 1. Use Case Diagram

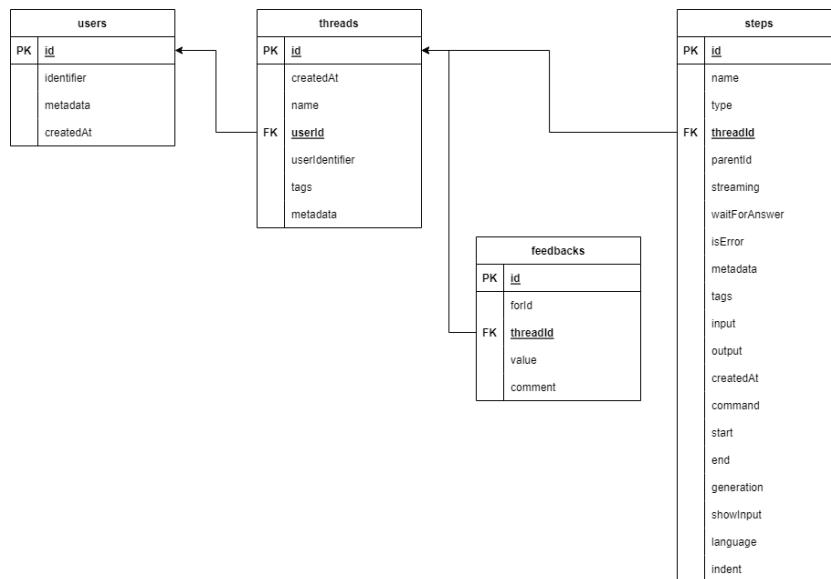
Gambar 2 menunjukkan interaksi antar komponen utama sistem. Diagram aktivitas menunjukkan alur aktivitas sistem yang menangani pertanyaan pengguna dari proses *input* hingga penyimpanan hasil.



Gambar 2. Activity Diagram

Ketika pengguna mengirimkan pertanyaan melalui antarmuka *Chatbot*, perjalanan dimulai. Sistem kemudian menerjemahkan pertanyaan ke dalam bahasa Inggris untuk memastikan bahasa dokumen dalam basis data sesuai. Setelah proses translasi selesai, model *embedding* digunakan untuk mengubah pertanyaan menjadi representasi vektor. Selanjutnya, metode *cosine similarity* digunakan untuk mencari informasi pada basis data vektor Qdrant. Model bahasa besar (LLM) kemudian memproses hasil pencarian dan dokumen yang relevan untuk menghasilkan jawaban yang tepat. Seluruh riwayat interaksi, termasuk umpan balik pengguna, disimpan dalam basis data PostgreSQL, dan tanggapan ditampilkan kepada pengguna. Proses ini menunjukkan bagaimana komponen pencarian generatif dan semantik diintegrasikan dalam sistem pencarian peningkatan generasi yang telah dibangun.

Sistem *Chatbot* menggunakan struktur basis data yang dirancang untuk membantu pengguna berinteraksi dengan *Chatbot* secara efisien dan terstruktur. Gambar 3 menunjukkan *Entity Relationship Diagram* (ERD), yang menggambarkan hubungan antar entitas utama dalam sistem.



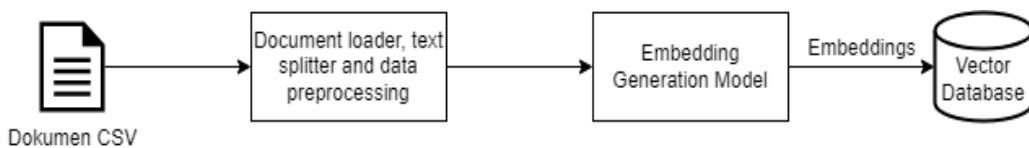
Gambar 3. Entity Relationship Diagram

Seperi yang ditunjukkan pada Gambar 3, tabel *users* memiliki relasi satu ke banyak (1:N) dengan tabel *threads*, yang menunjukkan bahwa satu pengguna dapat memiliki lebih dari satu *thread*. Selain itu, tabel *steps* memiliki relasi satu ke banyak (1:N) dengan tabel *threads*, karena satu *thread* dapat terdiri dari sejumlah *steps* atau proses yang dilakukan secara berurutan. Selain itu, tabel *feedbacks* memiliki hubungan langsung dengan tabel *threads* melalui atribut *threadId*. Fitur ini menunjukkan bahwa setiap *feedbacks* atau umpan balik diberikan terhadap *thread* tertentu. Setiap tabel memiliki fitur unik yang disesuaikan dengan fungsinya di sistem. Beberapa fitur, termasuk kolom metadata, disediakan dalam format JSON agar lebih fleksibel dan terorganisir. Ini adalah tujuan dari hubungan yang dibangun antara entitas ini untuk menunjukkan alur interaksi yang ada dalam sistem.

### C. Indexing

Tujuan dari tahap indeksing penelitian ini adalah untuk mengorganisasi dan menyimpan dokumen referensi ke dalam struktur indeks berbasis vektor, sehingga sistem dapat secara efektif mencari informasi yang relevan selama proses inferensi. Metode *Retrieval Augmented Generation* (RAG) digunakan untuk indeksing; proses ini menjadi bagian integral dari sistem untuk mendukung kemampuan pencarian berbasis semantik.

Pada tahap ini, fokus utama adalah optimalisasi struktur indeks dan kueri. Optimalisasi indeks mencakup peningkatan *granularitas* data, penambahan metadata, dan penggabungan dan penyelarasan strategi pencarian. Sementara itu, optimalisasi kueri bertujuan untuk menyederhanakan dan menyesuaikan pertanyaan pengguna agar lebih sesuai dengan proses pencarian dokumen. Teknik seperti penulisan ulang, transformasi, dan perluasan kueri juga digunakan [13]. Alur *indexing* data teks yang diambil dari file CSV yang berisi dokumen referensi digambarkan pada Gambar 4.



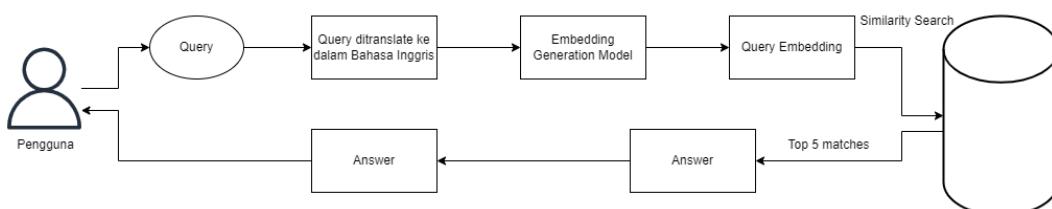
Gambar 4. Proses *Indexing*

Data dimuat dari file CSV dimulai dengan komponen *document loader*. Setelah dokumen dimuat, teknik pemisahan teks digunakan untuk memotong teks menjadi beberapa bagian berdasarkan panjang atau strurnya. Kemudian, fungsi *preprocess\_text* digunakan untuk memproses bagian teks ini. Fungsi ini membersihkan data dari karakter yang tidak penting seperti spasi berlebih, tanda kurung, tanda seru, tanda bintang, strip, dan tanda tanya. Setelah tahap praproses selesai, teks yang telah dibersihkan dikirim ke model *embedding* untuk diubah menjadi representasi vektor numerik yang akan disimpan di dalam *database* vektor. Dengan cara ini, sistem memiliki basis data representasi vektor yang dapat digunakan dalam tahap pencarian dokumen yang relevan dengan pertanyaan pengguna.

### D. Implementasi Sistem RAG

Setelah seluruh data referensi diindeks dalam basis data vektor, sistem siap untuk menjawab pertanyaan pengguna melalui dua tahapan utama: *retrieval* (pencarian) dan *generation* (pembuatan jawaban), seperti terlihat pada Gambar 5. Tahapan-tahapan ini membentuk inti dari metode *Retrieval Augmented Generation* (RAG) yang digunakan untuk membuat *Chatbot* konsultasi penyakit kulit.

Pada tahap pencarian (*retrieval*), sistem menggunakan model *embedding* yang sama dengan yang digunakan pada tahap indeksasi untuk mengonversi pertanyaan pengguna ke dalam representasi vektor. Sistem kemudian menghitung skor kemiripan antara vektor kueri dan vektor dokumen yang disimpan dalam *database* Qdrant. Hasil dari pencocokan ini menunjukkan bahwa lima dokumen yang menerima nilai tertinggi dipilih untuk digunakan sebagai konteks saat membuat jawaban.



Gambar 5. Proses *Retrieval* dan *Generation*

Tahap berikutnya adalah *generation*, di mana sistem membuat pesan dengan dokumen terpilih dan pertanyaan pengguna. Selanjutnya, model bahasa besar (LLaMA 3.1 8B) digunakan untuk memproses perintah ini untuk menghasilkan jawaban yang konsisten dan relevan. Strategi pembuatan jawaban ini dapat disesuaikan dengan tujuan tugas; jawaban dapat terbatas pada informasi yang ada dalam dokumen atau juga dapat mencakup pengetahuan internal dari model. Selain itu, *prompt* dapat menyertakan riwayat percakapan sebelumnya untuk mendukung interaksi *multi-turn* yang lebih alami dan kontekstual. Implementasi ini bertujuan untuk membuat *Chatbot* yang akurat, relevan, dan dapat dipertanggungjawabkan secara kontekstual dengan menggabungkan kemampuan pencarian berbasis semantik dan pemahaman bahasa alami.

Generasi adalah tahap berikutnya, di mana sistem membuat pesan dengan dokumen terpilih dan pertanyaan pengguna. Selanjutnya, model bahasa besar (LLaMA 3.1 8B) digunakan untuk memproses perintah ini untuk menghasilkan jawaban yang konsisten dan relevan. Strategi pembuatan jawaban ini dapat disesuaikan dengan tujuan tugas; jawaban dapat terbatas pada informasi yang ada dalam dokumen atau juga dapat mencakup pengetahuan internal dari model. Selain itu, *prompt* dapat menggabungkan riwayat percakapan sebelumnya untuk mendukung interaksi *multi-turn* yang lebih kontekstual dan alami [13].

Implementasi ini bertujuan untuk membuat *Chatbot* yang akurat, relevan, dan dapat dipertanggungjawabkan secara kontekstual dengan menggabungkan kemampuan pencarian berbasis semantik dan pemahaman bahasa alami.

#### E. Evaluasi

Evaluasi terhadap sistem *Chatbot* dilakukan untuk mengukur efektivitas penerapan pendekatan *Retrieval Augmented Generation* (RAG) dalam konteks konsultasi penyakit kulit. Pada tahap ini, *Retrieval Augmented Generation Assessment* (RAGAS) digunakan sebagai kerangka evaluasi. RAGAS merupakan metode penilaian otomatis yang dirancang khusus untuk mengevaluasi sistem berbasis RAG, dengan mengukur berbagai indikator keakuratan, termasuk kegunaan konteks yang diambil dan kualitas jawaban yang dihasilkan oleh model [14].

Evaluasi dilakukan dengan menggunakan tiga metrik utama, yaitu *faithfulness*, *LLMContextRecall*, dan *factual correctness*. Metrik *faithfulness* mengukur sejauh mana jawaban yang dihasilkan oleh model sesuai dengan konteks yang diberikan. *LLMContextRecall* mengevaluasi kemampuan model dalam memanfaatkan informasi dari dokumen hasil pencarian. Sementara itu, *factual correctness* digunakan untuk menilai akurasi fakta dalam jawaban yang dihasilkan, yang dihitung menggunakan ukuran *precision*, *recall*, dan *F1-score*.

Untuk melakukan evaluasi, skenario pengujian dengan lima pertanyaan acak yang berkaitan dengan penyakit kulit dibuat. Setiap pertanyaan dievaluasi secara kebahasaan dan, jika diperlukan, diterjemahkan ke dalam bahasa Inggris. Ini karena model LLM yang digunakan lebih cocok untuk *input* dalam bahasa Inggris. Selanjutnya, kueri tersebut digunakan untuk melakukan pencarian dokumen dalam basis data vektor Qdrant dengan menggunakan teknik pencocokan semantik berbasis *embedding*.

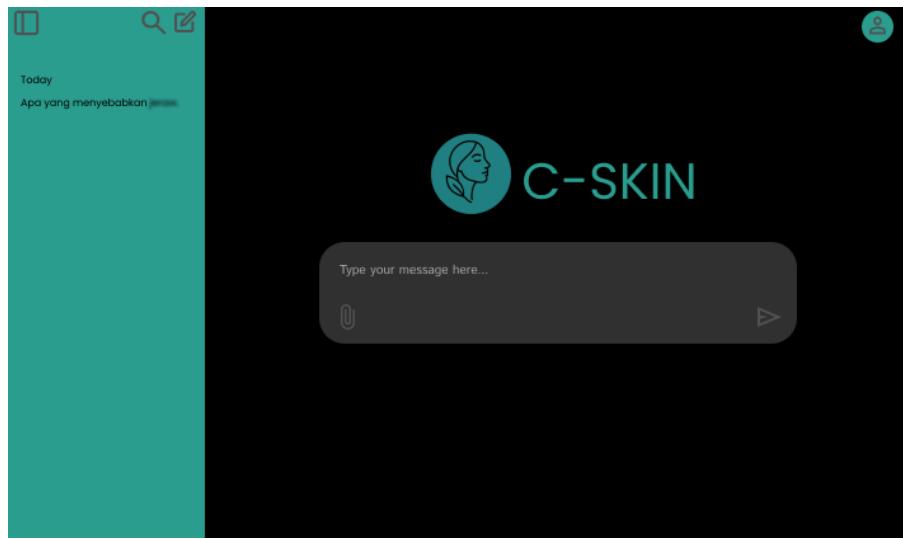
Setelah sistem menemukan dokumen yang relevan dengan tingkat kemiripan tertinggi, model LLM membuat jawaban berdasarkan konteks tersebut. Satu dataset evaluasi mengumpulkan seluruh proses interaksi, termasuk pertanyaan pengguna, dokumen hasil pencarian, dan jawaban yang dihasilkan. Selanjutnya, metrik RAGAS digunakan untuk menganalisis dataset ini untuk menghitung nilai kuantitatif untuk setiap metrik. Hasil evaluasi ini berfungsi sebagai dasar untuk menilai kinerja sistem dan memberikan saran yang relevan untuk langkah-langkah pengembangan sistem yang lebih lanjut. Khususnya, hal ini berkaitan dengan meningkatkan akurasi pencarian dokumen dan relevansi jawaban yang dihasilkan oleh model.

#### F. Integrasi Sistem

Mengintegrasikan sistem *Retrieval Augmented Generation* (RAG), yang telah dikembangkan dengan bantuan antarmuka pengguna berbasis web Chainlit, adalah langkah terakhir dalam metodologi penelitian ini. Tujuan integrasi ini adalah untuk memastikan bahwa sistem dapat digunakan secara interaktif oleh pengguna akhir, serta untuk memungkinkan visualisasi hasil pencarian dan jawaban dari model LLM secara *real-time*.

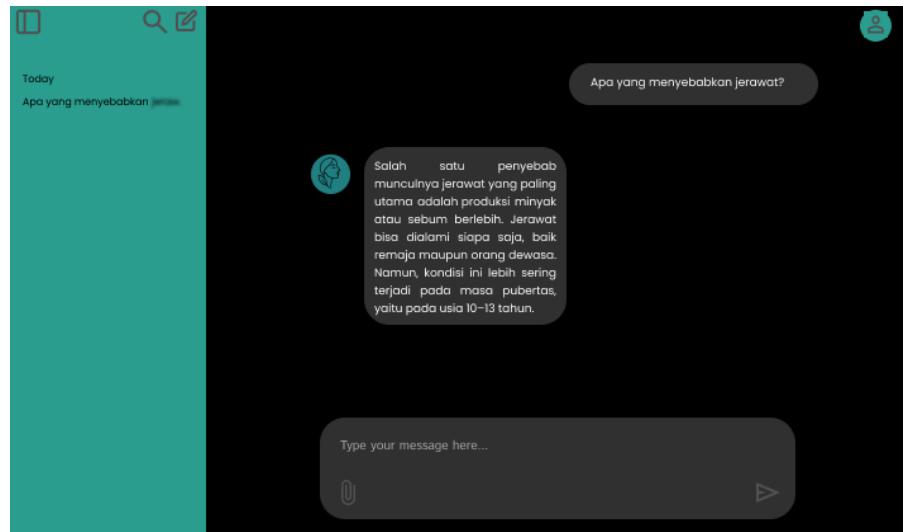
Chainlit, sebuah pustaka *open-source*, memungkinkan pengembangan cepat dan efektif aplikasi berbasis *Large Language Model* (LLM). Chainlit memungkinkan pengembang untuk menyematkan API ke dalam kode mereka yang sudah ada dan membuat antarmuka pengguna yang mirip ChatGPT. Selain itu, Chainlit memungkinkan pengembang untuk membuat aplikasi web interaktif untuk proyek pembelajaran mesin dan ilmu data tanpa harus memiliki keahlian khusus dalam pengembangan web [15].

Tiga tampilan utama termasuk dalam desain antarmuka pengguna: halaman awal dengan tema gelap (*dark mode*), halaman hasil pencarian dengan tema gelap, dan halaman hasil pencarian dengan tema terang (*light mode*). Sebelum pengguna memasukkan data, tampilan awal *dark mode* menampilkan *form* pertanyaan kosong, seperti yang ditunjukkan pada Gambar 6. Ini adalah awal interaksi *Chatbot* dengan pengguna.



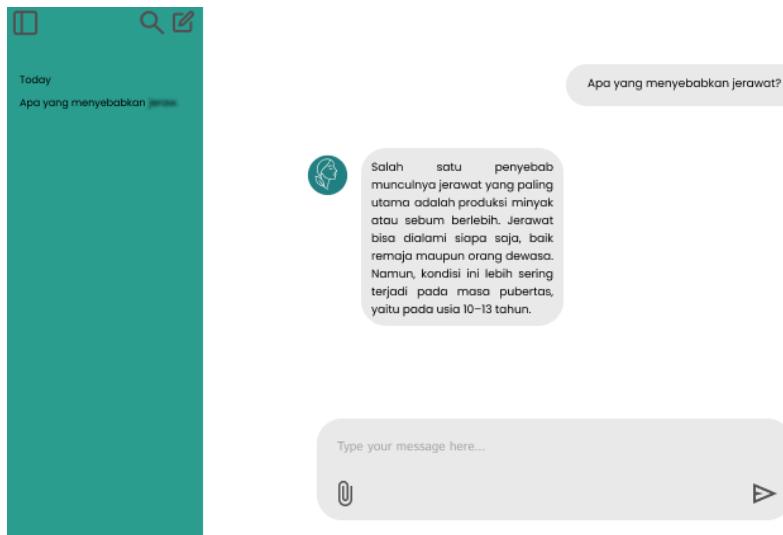
Gambar 6. Tampilan Awal

Seperti yang ditunjukkan pada Gambar 7, sistem akan menampilkan hasil jawaban dalam tampilan *dark mode* setelah pengguna mengajukan pertanyaan atau keluhan. Selain menjaga tampilan antarmuka yang konsisten, desain ini dimaksudkan untuk membuat mata nyaman, terutama dalam kondisi pencahayaan rendah.



Gambar 7. Tampilan Hasil Pencarian *Dark mode*

Sebaliknya, pengguna dapat memilih mode tampilan terang, atau mode terang, sesuai keinginan mereka. Hasil pencarian dengan tema terang menampilkan elemen yang serupa dengan mode gelap tetapi dengan skema warna yang lebih cerah untuk lingkungan dengan pencahayaan tinggi, seperti yang ditunjukkan pada Gambar 8.



Gambar 8. Tampilan Hasil Pencarian *Light mode*

Chainlit digunakan sebagai kerangka kerja antarmuka *frontend* untuk membantu sistem *backend* berbasis RAG berkomunikasi dengan pengguna. Chainlit dapat menampung *input* pengguna, menampilkan konteks, dan merespon secara langsung dalam antarmuka yang mudah dipahami. Untuk meningkatkan pengalaman pengguna, fitur baru telah ditambahkan, termasuk opsi personalisasi, pelacakan riwayat interaksi, dan penyesuaian konteks percakapan. Dengan tahap integrasi ini, sistem yang dibuat tidak hanya dapat melakukan proses pencarian dan pembuatan jawaban dengan baik, tetapi juga dapat digunakan dengan mudah melalui antarmuka web yang ramah pengguna.

### III. HASIL DAN PEMBAHASAN

Dalam penelitian ini, *Chatbot* berbasis *Retrieval Augmented Generation* (RAG) yang berfokus pada penyakit kulit dibuat. Sistem ini menggabungkan kemampuan pencarian informasi berbasis vektor (*retrieval*) menggunakan Qdrant dan *Sentence Transformer* serta kemampuan generatif dari model LLaMA 3.1 8B melalui pustaka Ollama. Studi ini menunjukkan bahwa, meskipun terdapat keterbatasan dalam hal efisiensi waktu proses generasi, pendekatan RAG dapat menghasilkan jawaban yang kontekstual, relevan, dan koheren untuk pertanyaan pengguna.

#### A. Dataset

Data awal dalam penelitian ini diperoleh dari situs *National Library of Medicine* - NIH (*National Institutes of Health*) [12]. Sumber ini dipilih karena kredibilitasnya dalam menyediakan informasi medis berbasis bukti (*evidence-based*), dan telah banyak digunakan sebagai rujukan dalam penelitian klinis dan pengembangan sistem medis kecerdasan buatan. Selanjutnya, *dataset* dikonversi dan diproses dalam format CSV dengan empat atribut utama: *question*, *answer*, *source*, dan *focus\_area*. Tabel 3 berikut menunjukkan distribusi jumlah data berdasarkan jenis penyakit kulit.

TABEL 3  
JUMLAH DATASET

| No | Data  | Jumlah |
|----|---|--------|
| 1  | <i>Carbuncle</i> atau Karbunkel                                       | 10     |
| 2  | <i>Acne Vulgaris</i> atau Jerawat                                     | 9      |
| 3  | <i>Atopic Dermatitis</i> atau Dermatitis Atopic                       | 10     |
| 4  | <i>Contact Dermatitis</i> atau Dermatitis Kontak                      | 8      |
| 5  | <i>Seborrheic Dermatitis</i> atau Dermatitis seboroik                 | 10     |
| 6  | <i>Nummular Dermatitis</i> atau Dermatitis numularis                  | 10     |
| 7  | <i>Diaper Dermatitis</i> ( <i>Diaper Rash</i> ) atau Dermatitis popok | 10     |
| 8  | <i>Varicella-Zoster Virus</i> ( <i>Chickenpox</i> ) atau Cacar Air    | 8      |
| 9  | <i>Shingles</i> atau <i>Herpes Zoster</i> atau Cacar Api              | 7      |
| 10 | <i>Mpox</i> ( <i>Monkeypox</i> ) atau Cacar Monyet                    | 10     |
| 11 | <i>Scabies</i> atau Kudis   | 10     |
| 12 | <i>Tinea Corporis</i> atau Kurap                                      | 10     |
| 13 | <i>Tinea Pedis</i> atau kutu air atau kurap kaki                      | 10     |

| No | Data  | Jumlah |
|----|---|--------|
| 14 | <i>Tinea Manuum</i> atau kurap tangan                                     | 10     |
| 15 | <i>Tinea Barbae</i> atau kurap jenggot                                    | 10     |
| 16 | <i>Tinea Capitis</i> atau <i>Ring Worm</i>                                | 10     |
| 17 | <i>Tinea Cruris</i> atau <i>jock itch</i>                                 | 10     |
| 18 | <i>Herpes Simplex Type 1</i>  | 8      |
| 19 | <i>Herpes Simplex Type 2</i>  | 10     |
| 20 | <i>Chronic Urticaria (Chronic Hives)</i> atau biduran kronis              | 10     |
| 21 | <i>Vitiligo</i>   | 10     |
| 22 | <i>Skin Cancer</i> atau Kanker Kulit                                      | 7      |
| 23 | <i>Psoriasis</i>  | 10     |
| 24 | <i>Impetigo</i>   | 9      |
| 25 | <i>Miliaria (Heat Rash, Prickly Heat, Sweat Rash)</i> atau biang keringat | 10     |
| 26 | <i>Rosacea</i>  | 9      |
| 27 | <i>Melasma</i>  | 8      |
| 28 | <i>Malignant Melanoma</i>   | 10     |
| 29 | <i>Tinea Versicolor</i> atau Panu   | 10     |
| 30 | <i>Leprosy (Hansen Disease)</i> atau Kusta atau lepra                     | 10     |

Penelitian ini menggunakan *dataset* yang digunakan untuk membangun sistem *Chatbot* konsultasi medis berbasis *Retrieval Augmented Generation* (RAG). *Dataset* ini dirancang secara sistematis untuk membantu proses *indexing* dan pencarian informasi berbasis vektor yang bergantung pada pendekatan semantik. Struktur setiap entri dalam dataset disesuaikan dengan arsitektur sistem pencarian berbasis *Large Language Model* (LLM), sehingga integrasi data dapat dilakukan dengan efektif dan relevan dengan konteks pertanyaan pengguna.

Empat komponen utama membentuk struktur data. Pertama, kolom *question* berisi berbagai jenis pertanyaan yang berkaitan dengan penyakit kulit. Ini mencerminkan variasi bahasa alami pengguna, yang memungkinkan sistem untuk menangani *input* yang beragam secara fleksibel. Kedua, kolom *answer* mengandung jawaban yang valid secara medis, informatif, dan disusun berdasarkan referensi ilmiah yang kredibel. Ketiga, *National Library of Medicine* (NIH) [12] adalah sumber utama jawaban, yang disebutkan dalam kolom *source* karena memastikan bahwa informasi yang diberikan akurat dan dapat diandalkan. Terakhir, setiap pertanyaan diberi label topik atau kategori utamanya dalam kolom *focus\_area*, misalnya, "*Acne Vulgaris*", "*Vitiligo*", atau "*Leprosy*." Kolom ini sangat penting untuk proses pencarian berbasis klasifikasi topikal selama tahapan *retrieval* selain berfungsi sebagai metadata untuk pengelompokan data.

#### G. Preprocessing Data

Dalam sistem *Chatbot* berbasis *Retrieval Augmented Generation* (RAG), tahap *preprocessing* sangat penting untuk menyiapkan data untuk proses *indexing* dan vektorisasi. Untuk memastikan bahwa format teks konsisten dan tersusun dengan benar, tujuan utamanya adalah memastikan bahwa proses *embedding* berjalan dengan lancar. Normalisasi spasi, penghilangan tanda baca seperti tanda kurung, tanda seru, tanda tanya, dan simbol non-alfabet lainnya adalah bagian dari proses *preprocessing*. Untuk menghindari perubahan representasi yang tidak perlu selama proses vektorisasi, langkah ini dilakukan untuk menyederhanakan struktur teks. Selain itu, untuk menghindari kesalahan teknis selama proses *encoding* dan penyimpanan data, string kosong digunakan untuk menggantikan nilai kosong dalam dataset.

Proses ini diterapkan secara menyeluruh pada keempat kolom utama *dataset*, yaitu *question*, *answer*, *source*, dan *focus\_area*. Setelah tahap pembersihan, data dikonversi ke dalam format *list*, yang memudahkan proses *batch* saat dimasukkan ke dalam basis data vektor Qdrant. Tabel 4 menunjukkan bahwa proses *preprocessing* berhasil membuat format teks lebih sederhana dengan menghilangkan huruf kapital, tanda baca, dan karakter khusus lainnya.

TABEL 4  
SAMPEL DATASET SEBELUM DAN SESUDAH PREPROCESSING

| No | Sebelum   | Sesudah   |
|----|---|---|
| 1  | <i>What is Carbuncle?</i>   | <i>what is Carbuncle</i>  |
| 2  | <i>A Carbuncle is a cluster of boils, called furuncles, that are connected to each other. This occurs when an infection starts in one or more hair follicles and spreads to the skin and deeper layers of tissue surrounding these follicles. Carbuncles usually appear as red, tender, and swollen lumps with several oozing openings or pus-filled bumps on the surface. They often come with general symptoms of</i> | <i>a Carbuncle is a cluster of boils, called furuncles, that are connected to each other. this occurs when an infection starts in one or more hair follicles and spreads to the skin and deeper layers of tissue surrounding these follicles. Carbuncles usually appear as red, tender, and swollen lumps with several oozing openings or pusfilled bumps on the surface.</i> |

| No | Sebelum  | Sesudah   |
|----|--|---|
|    | <i>illness, and swollen glands may also occur in the nearby area. While Carbuncles can appear anywhere on the body where there is hair, they are most commonly found on areas with thicker skin like the back of the neck, back, and thighs. If an infected hair follicle, known as folliculitis, doesn't get treated, it can turn into a boil, or furuncle. When these boils cluster together, they create a Carbuncle. You can have one Carbuncle or several at the same time.</i> | <i>they often come with general symptoms of illness, and swollen glands may also occur in the nearby area. while Carbuncles can appear anywhere on the body where there is hair, they are most commonly found on areas with thicker skin like the back of the neck, back, and thighs. if an infected hair follicle, known as folliculitis, doesn't get treated, it can turn into a boil, or furuncle. when these boils cluster together, they create a Carbuncle. you can have one Carbuncle or several at the same time.</i> |
| 3  | NIH (National Library of Medicine)   | nih National Library of Medicine  |
| 4  | Carbuncle atau Karbunkel   | Carbuncle atau karbunkel  |

Proses ini secara langsung meningkatkan konsistensi data, yang berkontribusi pada ketepatan proses vektorisasi serta relevansi hasil pencarian informasi dalam sistem. Dengan format teks yang seragam, model vektorisasi dapat lebih mudah menangkap makna semantik dari data *input*, yang meningkatkan akurasi sistem dalam menanggapi pertanyaan pengguna secara kontekstual.

#### H. Vektorisasi Data

Dalam pendekatan *Retrieval Augmented Generation* (RAG), tahap vektorisasi adalah dasar dari sistem pencarian berbasis semantik. Dalam proses ini, model transformasi frasa digunakan untuk mengubah data teks, terutama kolom jawaban, menjadi representasi numerik berdimensi tetap. Tujuan dari proses ini adalah untuk mengubah informasi linguistik menjadi vektor semantik yang bermakna sehingga dapat dibandingkan berdasarkan makna yang sebanding. Dalam proses ini, model *transformer* canggih BAAI/bge-m3 digunakan. Model ini dikembangkan oleh *Beijing Academy of Artificial Intelligence* (BAAI). Model ini ideal untuk sistem multibahasa dan multikonteks seperti *Chatbot* medis karena kemampuannya yang unggul dalam tiga aspek utama: multifungsi, multibahasa, dan multigranular. Sampel data dari kolom *answer*, yang merupakan komponen utama proses *encoding*, ditunjukkan pada Gambar 9.

```
['a carbuncle is a cluster of boils, called furuncles, that are connected to each other. this occurs when an infection starts in one or more hair follicles and spreads to the skin and deeper layers of tissue surrounding these follicles. carbuncles usually appear as red, tender, and swollen lumps with several oozing openings or pus-filled bumps on the surface. they often come with general symptoms of illness, and swollen glands may also occur in the nearby area. while carbuncles can appear anywhere on the body where there is hair, they are most commonly found on areas with thicker skin like the back of the neck, back, and thighs. if an infected hair follicle, known as folliculitis, doesn't get treated, it can turn into a boil, or furuncle. when these boils cluster together, they create a carbuncle. you can have one carbuncle or several at the same time.', 'carbuncles, which are infections in the hair follicle, typically occur due to bacteria. the main bacteria responsible for this is called staphylococcus aureus, which can include methicillin-resistant staphylococcus aureus, a more resistant type of this bacteria. in certain cases, especially when the infection keeps coming back or is affecting the genital and anal areas, it can be caused by a group of bacteria that don't need oxygen to survive, known as anaerobic bacteria. staphylococcus aureus is a type of bacteria that can generally be found living harmlessly on the skin, particularly in areas where skin folds over itself, like the groin, armpits, buttocks, and neck. it can also live inside your nose. if you scratch these areas, the bacteria can be transferred to other parts of your body. if the skin is broken or damaged, these bacteria can infect the hair follicle. once they've entered the follicle, they can multiply, leading to different conditions such as folliculitis inflammation of the follicles, furuncle a boil, or a carbuncle a cluster of boils.', 'carbuncles, while common, occur most often in young adults to middleaged adults. they are not usually seen in young children. males are more likely to develop carbuncles than females. several situations can make a person more likely to develop carbuncles. these include factors that weaken or break the skin and conditions that affect the person's overall health. eczema, a
```

Gambar 9. Sampel Data Kolom *Answers*

Setiap jawaban diubah menjadi vektor berdimensi 1024, sesuai dengan *output* dari model BAAI/bge-m3. Vektor-vektor ini mengandung informasi semantik dari teks, yang memungkinkan sistem menemukan kesamaan makna antar entri meskipun struktur kalimat berbeda. Gambar 10 menunjukkan bahwa hasil *encoding* menghasilkan kumpulan vektor yang menjelaskan makna setiap jawaban. Selanjutnya, metrik kemiripan *cosine*, yang efektif dalam membandingkan kemiripan arah antara dua vektor dalam ruang besar, digunakan untuk melakukan proses pencarian informasi.

```
array([[-0.01789068, -0.00459144, -0.05224792, ..., -0.00330604,
       0.04573935, -0.04024445],
      [-0.01483113,  0.00160293, -0.05856834, ...,  0.0047892 ,
       0.02060792, -0.04545713],
      [-0.01777068, -0.0063722 , -0.05909504, ...,  0.00555904,
       0.02845593, -0.03971741],
      ...,
      [-0.01006315,  0.02759071, -0.04798842, ..., -0.0429833 ,
       -0.03569194, -0.02799565],
      [-0.01015824,  0.0415001 , -0.03395165, ..., -0.02394886,
       -0.01462314, -0.02281419],
      [-0.04476942,  0.02435144, -0.05593533, ...,  0.00702196,
       0.03375359, -0.05196096]], dtype=float32)
```

Gambar 10. Vektorisasi Data

Penelitian telah menunjukkan bahwa penggunaan metrik ini dapat meningkatkan akurasi pencarian semantik karena berfokus pada makna daripada struktur literal. Selanjutnya, semua hasil *embedding* dimasukkan ke dalam *database* vektor Qdrant, pada koleksi berjerawat. Strategi *quantization* dengan format INT8 digunakan untuk meningkatkan efisiensi ruang dan kecepatan pencarian. Strategi ini mengurangi ukuran penyimpanan tanpa mengurangi performa secara signifikan.

### I. Pencarian Informasi (Retrieval)

Salah satu bagian penting dari sistem *Chatbot* berbasis *Retrieval Augmented Generation* (RAG) adalah proses pencarian informasi, yang bertujuan untuk menggunakan kemiripan semantik untuk menemukan jawaban terbaik untuk pertanyaan yang diajukan pengguna. Saat ini, sistem melakukan pencocokan semantik dengan baik dengan *database* vektor Qdrant. Pencarian dimulai dengan pertanyaan yang dimasukkan oleh pengguna. Sebagai langkah awal, sistem secara otomatis mendeteksi *input* dalam berbagai bahasa. Sistem internal menggunakan *Large Language Model* (LLM) untuk mengubah pertanyaan yang tidak dalam bahasa Inggris ke dalam bahasa Inggris. Sebelum teks diubah menjadi representasi vektor, proses ini sangat penting untuk memastikan konsistensi linguistik.

Setelah pertanyaan diterjemahkan, teks hasil translasi kemudian diproses menggunakan model BAAI/bge-m3, yaitu model *Sentence Transformer* yang sama digunakan pada proses vektorisasi data sebelumnya. Model ini mengubah pertanyaan menjadi vektor berdimensi tetap, yang kemudian digunakan sebagai vektor kueri. Gambar 11 menunjukkan contoh hasil dari prosedur untuk mendapatkan Qdrant. Dalam proses ini, metrik *Cosine similarity* digunakan untuk membandingkan vektor kueri dengan seluruh vektor yang telah disimpan dalam koleksi Penyakit Kulit. Hanya informasi yang relevan yang diberikan kepada pengguna, dengan hanya lima hasil teratas dan ambang batas skor kemiripan sebesar 0,4.

Q: what are the signs and symptoms of carbuncle  
A: a person with a carbuncle, which is a kind of skin infection, will often tell you about a growing tender bump on their body. they might mention that it first appeared as a simple "pimple" or a small pocket of pus, which they tried to pop. but, in a few days or weeks, it began to grow bigger and become more tender and squishy to touch. now, it's important to know that carbuncles can lead to fullbody symptoms, but their presence is not necessary for diagnosing the condition. if present, these fullbody symptoms could include fever, swollen lymph nodes, fatigue, and a general feeling of discomfort or illness. when you look at a carbuncle, what you'd typically see is a reddish, painful bump with several pusfilled bumps on top. often, these bumps break open due to small injuries, rubbing, or pressure, which can form a crusty layer on top. as for where on the body you're likely to find them, carbuncles can appear anywhere you have hair. that being said, they're more commonly found on certain parts of the body such as the back of the neck, face, back, buttocks, underarms, and groin. initially appeared as a "pimple" or a small pocket of pus grew bigger over a few days or weeks became tender and squishy to touch may have caused fullbody symptoms including fever, swollen lymph nodes, fatigue, and discomfort or illness appears as a reddish, painful bump with several pusfilled bumps on top bumps often break open to form a crusty layer on top can appear anywhere on the body with hair, with more common areas being the back of the neck, face, back, buttocks, underarms, and groin  
Source: nih national library of medicine  
Focus Area: carbuncle atau karbunkel

Gambar 11. Sampel Retrieve Qdrant

Hasil pencarian dikemas dalam struktur yang informatif, terdiri dari kolom *question*, *answer*, *source*, dan *focus\_area*, sehingga memberikan konteks yang cukup kepada model generatif untuk menghasilkan jawaban akhir. Proses ini memungkinkan sistem menelusuri informasi penting dengan akurat, meskipun *input* awal berasal dari bahasa selain Inggris.

Metode ini menunjukkan kemampuan sistem untuk melakukan pencarian informasi melalui pendekatan semantik lintas-bahasa. Metode ini meningkatkan fleksibilitas dan aksesibilitas *Chatbot* karena memungkinkan mereka menjembatani keragaman bahasa pengguna tanpa mengurangi kualitas hasil pencarian.

#### J. Generasi Jawaban (Generation)

Setelah proses pencarian informasi berbasis semantik dan hasil yang relevan diperoleh, tahap selanjutnya dalam sistem *Chatbot* ini adalah menghasilkan jawaban dalam bentuk cerita yang mudah dipahami oleh pengguna. Model LLaMA 3.1 8B, yang diintegrasikan melalui pustaka Ollama, digunakan untuk melakukan proses ini. Model ini membentuk C-Skin, *Chatbot* medis yang ramah tetapi profesional yang merespons pertanyaan dengan konteks yang tepat.

Dalam proses generasi jawaban, sistem yang dikembangkan menggunakan *prompt* template sebagai mekanisme utama untuk menjaga konsistensi struktur, kualitas isi, dan etika komunikasi. Template ini dirancang secara sistematis untuk membatasi ruang lingkup informasi yang digunakan oleh model bahasa dan memastikan bahwa semua jawaban yang dihasilkan hanya berasal dari hasil pencarian informasi yang relevan, seperti yang ditunjukkan pada Gambar 12. Oleh karena itu, proses generatif tidak hanya mengontrol isi tetapi juga keandalan dan ketepatan data yang dikirim.

Q: what are the signs and symptoms of carbuncle  
A: a person with a carbuncle, which is a kind of skin infection, will often tell you about a growing tender bump on their body. they might mention that it first appeared as a simple "pimple" or a small pocket of pus, which they tried to pop. but, in a few days or weeks, it began to grow bigger and become more tender and squishy to touch. now, it's important to know that carbuncles can lead to fullbody symptoms, but their presence is not necessary for diagnosing the condition. if present, these fullbody symptoms could include fever, swollen lymph nodes, fatigue, and a general feeling of discomfort or illness. when you look at a carbuncle, what you'd typically see is a reddish, painful bump with several pusfilled bumps on top. often, these bumps break open due to small injuries, rubbing, or pressure, which can form a crusty layer on top. as for where on the body you're likely to find them, carbuncles can appear anywhere you have hair. that being said, they're more commonly found on certain parts of the body such as the back of the neck, face, back, buttocks, underarms, and groin. initially appeared as a "pimple" or a small pocket of pus grew bigger over a few days or weeks became tender and squishy to touch may have caused fullbody symptoms including fever, swollen lymph nodes, fatigue, and discomfort or illness appears as a reddish, painful bump with several pusfilled bumps on top. bumps often break open to form a crusty layer on top can appear anywhere on the body with hair, with more common areas being the back of the neck, face, back, buttocks, underarms, and groin  
Source: nih national library of medicine  
Focus Area: carbuncle atau karbunkel

Gambar 12. *Prompt* Template

Struktur pemrosesan LLM *Chain*, yang mengintegrasikan model LLaMA dengan template *prompt* yang telah dirancang, digunakan untuk melaksanakan proses generasi jawaban. Komponen ini menerima dua masukan utama: hasil pencarian informasi (*context retrieved*) dan pertanyaan yang diajukan oleh pengguna. Selain itu, parameter seperti nada atau gaya komunikasi digunakan, yang diatur dalam sistem ini dengan nilai "profesional dan ramah". Tujuan penyesuaian nada komunikasi ini adalah untuk membuat interaksi yang tidak hanya informatif tetapi juga nyaman dan empatik, seperti layanan konsultasi medis berbasis AI. Contoh hasil akhir yang dihasilkan oleh model ditunjukkan pada Gambar 13.

Terima kasih telah berkonsultasi dengan C-Skin. Berdasarkan gejala yang Anda deskripsikan, lumpa di bawah ketiak Anda yang menyakitkan dan tampak seperti ada nanah di dalamnya dapat menunjukkan beberapa kemungkinan penyakit. Namun, perlu diingat bahwa hanya dokter yang dapat melakukan diagnosis akurat. Dari konteks yang tersedia, beberapa kemungkinan penyakit yang mungkin terkait dengan gejala Anda adalah: \* Karbunkel (carbuncle): karbunkel adalah infeksi kulit yang menyebabkan pembengkakan dan rasa sakit di daerah yang terinfeksi. Gejala lainnya termasuk demam, limfadenopati (peradangan kelenjar getah bening), kelelahan, dan perubahan warna kulit. \* Infeksi kulit lainnya: ada kemungkinan bahwa lumpa di bawah ketiak Anda disebabkan oleh infeksi kulit lainnya seperti impetigo atau ekitimanya. Untuk mendapatkan diagnosis yang akurat, sebaiknya Anda berkonsultasi dengan dokter. Dokter akan melakukan pemeriksaan fisik dan mungkin melakukan tes laboratorium untuk menentukan penyebab pasti gejala Anda. Saya harap informasi ini membantu. Jika Anda memiliki pertanyaan lain, jangan ragu untuk bertanya!

Gambar 13. Hasil Generasi

Format jawaban telah disesuaikan untuk sesuai dengan struktur *prompt template*. Itu dimulai dengan kalimat, "Terima kasih telah berkonsultasi dengan C-Skin," dan ditutup dengan pernyataan empati yang mendukung kenyamanan pengguna. Metode ini tidak hanya meningkatkan keakuratan jawaban yang sesuai dengan konteks, tetapi juga memastikan bahwa informasi disampaikan dengan cara yang moral, informatif, dan mudah diterima oleh khalayak luas. Selain itu, penggunaan cepat yang terorganisir memungkinkan LLM membuat jawaban yang konsisten dan tidak menyimpang dari referensi yang tersedia.

#### K. Evaluasi Model

Salah satu tujuan evaluasi sistem adalah untuk mengevaluasi kemampuan model bahasa besar (LLM) untuk memberikan tanggapan yang relevan dan tepat untuk situasi medis tertentu. Metode *Retrieval Augmented Generation Assessment* (RAGAS) adalah metode kuantitatif yang digunakan untuk menilai kinerja sistem berbasis RAG dengan menggunakan tiga metrik utama: Kesetiaan, *LLMContextRecall*, dan *Factual correctness*.

Untuk menilai, lima pertanyaan representatif tentang penyakit kulit (jerawat, karbunkel, *Vitiligo*, panu, dan kusta) diberikan dan dijawab dengan referensi. Semua data diperiksa dalam bahasa Inggris untuk memastikan konsistensi linguistik. Proses RAG terdiri dari dua langkah: mendapatkan dokumen yang relevan dari *database* Qdrant dan membuat jawaban menggunakan model LLM. Metrik Kesetiaan menilai kesesuaian jawaban dengan dokumen sumber, *LLMContextRecall* menilai pemanfaatan konteks, dan *F1-score* menilai ketepatan informasi dalam jawaban yang dihasilkan. Hasil evaluasi dapat dilihat pada Tabel 5.

TABEL 5  
HASIL EVALUASI MODEL

| No | Metrik Evaluasi                        | Skor   |
|----|--|--------|
| 1  | <i>Faithfulness</i>                    | 0,9429 |
| 2  | <i>LLMContextRecall</i>                | 0,9600 |
| 3  | <i>Factual correctness (F1-score)</i>  | 0,7080 |
| 4  | <i>Factual correctness (Recall)</i>    | 0,7640 |
| 5  | <i>Factual correctness (Precision)</i> | 0,4720 |

Berdasarkan hasil evaluasi, sistem menunjukkan kinerja yang baik dalam memanfaatkan konteks yang tersedia dan memastikan bahwa jawaban sesuai dengan dokumen sumber. Ini ditunjukkan oleh skor *Faithfulness* (0,429) dan *LLMContextRecall* (0,6960), tetapi nilai *Factual correctness*, terutama pada aspek *precision* (0,4720), menunjukkan bahwa beberapa informasi yang dihasilkan belum sepenuhnya tepat. Meskipun informasi relevan dapat ditemukan oleh sistem (*recall* = 0,7640), penyampaian konten masih perlu ditingkatkan. Oleh karena itu, untuk meningkatkan ketepatan jawaban, strategi *prompt engineering*, pemurnian konteks, dan kemungkinan integrasi modul verifikasi fakta semuanya diperlukan.

#### L. Penerapan Chainlit

*Chatbot* medis C-Skin dikembangkan menggunakan Chainlit sebagai bagian dari integrasi sistem untuk membuat antarmuka web yang interaktif dan mudah diakses. Tujuan implementasi ini adalah untuk membuat pengalaman pengguna lebih mudah dipahami dan membantu proses interaksi dalam situasi penggunaan nyata. Selain menyediakan berbagai fitur tambahan untuk mendukung kepraktisan pengguna dan evaluasi sistem, platform Chainlit memungkinkan visualisasi proses tanya-jawab dengan model LLM secara *real-time*. Sistem menggunakan akun GitHub dan Google untuk mendukung elemen autentikasi. Pendekatan ini tidak hanya membuat proses registrasi lebih mudah, tetapi juga membuat pengguna lebih nyaman menggunakan layanan tanpa perlu membuat akun baru secara manual. Gambar 14 menunjukkan bagaimana pengguna akan diarahkan ke halaman utama *Chatbot* setelah berhasil masuk.

Terima kasih telah berkonsultasi dengan C-Skin. Berdasarkan gejala yang Anda deskripsikan, lumpa di bawah ketiak Anda yang menyakitkan dan tampak seperti ada nanah di dalamnya dapat menunjukkan beberapa kemungkinan penyakit. Namun, perlu diingat bahwa hanya dokter yang dapat melakukan diagnosis akurat. Dari konteks yang tersedia, beberapa kemungkinan penyakit yang mungkin terkait dengan gejala Anda adalah: \* Karbunkel (carbuncle): karbunkel adalah infeksi kulit yang menyebabkan pembengkakan dan rasa sakit di daerah yang terinfeksi. Gejala lainnya termasuk demam, limfadenopati (peradangan kelenjar getah bening), kelelahan, dan perubahan warna kulit. \* Infeksi kulit lainnya: ada kemungkinan bahwa lumpa di bawah ketiak Anda disebabkan oleh infeksi kulit lainnya seperti impetigo atau ekimanya. Untuk mendapatkan diagnosis yang akurat, sebaiknya Anda berkonsultasi dengan dokter. Dokter akan melakukan pemeriksaan fisik dan mungkin melakukan tes laboratorium untuk menentukan penyebab pasti gejala Anda. Saya harap informasi ini membantu. Jika Anda memiliki pertanyaan lain, jangan ragu untuk bertanya!

Gambar 14. Tampilan *Chatbot* dan Riwayat Percakapan

Halaman ini memiliki fitur untuk menyimpan riwayat percakapan. Pengguna dapat meninjau kembali diskusi sebelumnya dan memanfaatkan jawaban yang telah diberikan terutama dalam kasus medis yang sebanding dengan riwayat ini. Fitur ini sangat membantu pengguna dalam proses pembelajaran mandiri dan evaluasi berulang kondisi kulit mereka.

Untuk memberi pengguna kemampuan untuk menilai relevansi dan kualitas jawaban *Chatbot*, fitur umpan balik juga dikenal sebagai *feedback* ditambahkan sebagai bagian dari upaya untuk meningkatkan interaktivitas dan evaluasi sistem. Dalam hal kinerja model bahasa dan akurasi pencarian informasi, masukan ini sangat penting dalam proses penyempurnaan sistem secara berkelanjutan. Secara keseluruhan, Chainlit berhasil menggabungkan teknologi RAG dan LLM ke dalam antarmuka *Chatbot* medis yang *responsif* dan mudah digunakan yang mendukung evaluasi berbasis pengguna secara *real-time* melalui platform web, yang dapat diakses melalui tautan berikut : <https://c-skin-934512542726.asia-southeast1.run.app>.

#### M. Pengujian Sistem

Pada tahap pengujian sistem, *Chatbot* dievaluasi untuk efisiensi dan akurasi saat menjawab pertanyaan. Digunakan tiga kategori pertanyaan: pendek, sedang, dan panjang. Kategori ini digunakan untuk menunjukkan bagaimana *input* pengguna berbeda dalam situasi penggunaan nyata. Setiap pertanyaan melalui tiga tahap pemrosesan utama: *encoding*, pencarian vektor Qdrant dari *database*, dan proses *generative* LLM.

Hasil pengujian menunjukkan proses pemulihan yang sangat cepat, dengan 0,08–0,29 detik. Ini menunjukkan bahwa Qdrant sangat efektif dalam menemukan data yang relevan. Namun, proses produksi membutuhkan waktu lebih lama, kira-kira 255–283 detik. Hal ini disebabkan oleh keterbatasan perangkat keras yang digunakan, yaitu server dengan konfigurasi n2-standard-8 (8 vCPU, 32 GB RAM) tanpa akselerator GPU. Karena proses inferensi model LLM dilakukan sepenuhnya di CPU, waktu yang diperlukan sangat lama.

Untuk relevansi, skor kemiripan (*similarity score*) untuk lima hasil teratas pada setiap pertanyaan berkisar antara 0,593 dan 0,696, menunjukkan bahwa sistem mampu mengidentifikasi dokumen dengan tingkat kesesuaian semantik yang cukup tinggi terhadap pertanyaan pengguna. Secara keseluruhan, pengujian menunjukkan bahwa sistem dapat memberikan jawaban yang akurat dan informatif untuk berbagai jenis pertanyaan. Namun, efisiensi proses generatif masih menjadi masalah yang dapat diatasi melalui peningkatan infrastruktur seperti penggunaan GPU untuk mempercepat proses inferensi dan mengurangi waktu tunggu pengguna.

#### IV. SIMPULAN

Penelitian ini membuktikan bahwa pendekatan *Retrieval Augmented Generation* (RAG) dengan integrasi LangChain dan model LLaMA mampu meningkatkan kualitas *responses Chatbot* dalam konsultasi penyakit kulit, sebagaimana ditunjukkan oleh nilai *Faithfulness* (0,9429) dan *LLMContextRecall* (0,9600) yang tinggi. Meskipun nilai *Recall* (0,7640) menunjukkan kemampuan sistem dalam menangkap informasi penting, rendahnya *Precision* (0,4720) dan *F1-score* (0,7080) mengindikasikan perlunya optimasi lebih lanjut untuk meningkatkan akurasi jawaban. Integrasi dengan *platform Chainlit* juga berhasil menciptakan antarmuka yang interaktif dan ramah pengguna melalui fitur login, riwayat percakapan, dan umpan balik. Dari sisi performa, kecepatan *retrieval* yang optimal menunjukkan efisiensi sistem, meskipun waktu generasi jawaban yang masih tinggi menandakan perlunya dukungan infrastruktur GPU. Ke depannya, penelitian ini dapat dilanjutkan dengan pengembangan model yang lebih efisien dan pemanfaatan umpan balik pengguna untuk meningkatkan akurasi dan pengalaman interaksi dalam layanan konsultasi medis berbasis *Chatbot*.

#### DAFTAR PUSTAKA

- [1] Y. W. A. Rustam, C. Chazar, and Moch. A. Ramdhani, “Aplikasi Diagnosa Penyakit Kulit Menggunakan dengan Menggunakan Metode Convolutional Neural Networks,” *INFORMASI (Jurnal Informatika dan Sistem Informasi)*, vol. 15, no. 2, pp. 208–224, 2023.
- [2] M. Ath-Thariq and T. N. Suharsono, “Deteksi Penyakit Kulit Serupa Pada Wajah Berbasis Mobile dengan Metode Convolutional Neural Network,” *Innovative: Journal Of Social Science Research*, vol. 3, no. 5, pp. 876–887, 2023.
- [3] A. P. Chaves and M. A. Gerosa, “How should my chatbot interact? A survey on human-chatbot interaction design,” *Int J Hum Comput Interact*, vol. 37, no. 8, pp. 729–758, 2020.
- [4] J. Kim and M. Min, “From RAG to QA-RAG: Integrating Generative AI for Pharmaceutical Regulatory Compliance Process,” 2024, [Online]. Available: <http://arxiv.org/abs/2402.01717>
- [5] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. G. Valencia, and W. Cheungpasitporn, “Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications,” *Medicina (B Aires)*, vol. 60, no. 3: 445, 2024.
- [6] L. N. Berezhnoy, “Creation of a conversational bot specialized in a specific domain,” University of Alicante, 2024. [Online]. Available: <https://rua.ua.es/server/api/core/bitstreams/d3939216-3844-4f6c-8b42-42209a69c8e3/content>
- [7] M. I. Syah, N. S. Harahap, Novriyanto, and S. Sanjaya, “Penerapan Retrieval Augmented Generation Menggunakan Langchain dalam Pengembangan Sistem Tanya Jawab Berbasis Web,” *ZONAsi Jurnal Sistem Informasi*, vol. 6, no. 2, pp. 370–379, 2024.
- [8] H. Jain, “A Medical Chatbot: Your Healthcare Assistance,” *International Research Journal of Modernization in Engineering Technology and Science*, vol. 6, no. 6, 2024.
- [9] W. Shi, Y. Zhuang, Y. Zhu, H. Iwinski, M. Wattenbarger, and M. D. Wang, “Retrieval-Augmented Large Language Models for Adolescent Idiopathic Scoliosis Patients in Shared Decision-Making,” in *ACM-BCB 2023 - 14th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, Association for Computing Machinery, Inc, 2023.
- [10] A. Y. Alan, E. Karaarslan, and O. Aydin, “A RAG-based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM,” 2025. [Online]. Available: <https://arxiv.org/abs/2401.15378>
- [11] S. Sarma and N. Pathak, “A Bi-LSTM and Attention Based Approach for Developing an Assamese AI Chatbot,” *Indian J Sci Technol*, vol. 17, no. 36, pp. 3811–3823, 2024.
- [12] “National Center for Biotechnology Information.” Accessed: Jun. 24, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/>
- [13] Y. Gao *et al.*, “Retrieval-Augmented Generation for Large Language Models: A Survey,” 2024, [Online]. Available: <http://arxiv.org/abs/2312.10997>
- [14] “LangChain - Ragas,” 2025. [Online]. Available: <https://docs.ragas.io/en/stable/howtos/integrations/langchain/>
- [15] A. G, “RAG based Chatbot using LLMs,” *International Journal of Scientific Research in Engineering and Management*, vol. 08, no. 06, pp. 1–5, 2024.