AN INDONESIAN CHATBOT FOR DISEASE DIAGNOSIS USING RETRIEVAL-AUGMENTED GENERATION

CHATBOT INDONESIA UNTUK DIAGNOSIS PENYAKIT MENGGUNAKAN RETRIEVAL-AUGMENTED GENERATION

Muhammad Adrinta Abdurrazzaq[1], Edwin Lesmana Tjiong[2], Aulia Fasya[3], Michelle Hiu[4], Joses Tanuwidjaya[5]

Universitas Kalbis, Jl. Pulomas Selatan Kav. No.22, Jakarta Timur, DKI Jakarta, Indonesia
muhammad.abdurrazzaq@kalbis.ac.id [1], edwin.tjiong@kalbis.ac.id[2], 2023105499@student.kalbis.ac.id [3], 2023105488@student.kalbis.ac.id [4], 2023105513@student.kalbis.ac.id [5]

Abstract *– The rapid advancement of Large Language Models (LLMs) has enabled their use in medical information systems, although challenges such as hallucinations, domain mismatches, and the lack of a verified knowledge base remain significant, particularly in low-source languages like Indonesian. This study introduces an Indonesian-language medical chatbot based on the open-source GPT-OSS-20B model enhanced through a Retrieval-Augmented Generation (RAG) pipeline. The system combines semantic retrieval using jina-embeddings-v3, lexical re-ranking with the BM25 algorithm, and a lightweight Logistic Regression-based domain filter as an initial filter to prevent out-of-domain LLM usage. Evaluation using Indonesian medical articles and annotated patient-doctor conversations shows that the domain filter works well on synthetic data but results in misclassification of natural queries. A hybrid weighted reranker (FAISS L2 + BM25) performed the best with a Top-30 accuracy of 0.699. Black-box testing indicates that the system flow functions as designed, although the response quality has not been validated by clinical experts. These findings suggest that RAG-based open-source LLMs can improve access to Indonesian-language medical information, but still have important limitations such as the lack of clinical validation, potential errors in scraped data, and suboptimal robustness of domain filters.*

Keywords *- Retrieval-Augmented Generation, GPT-OSS, medical chatbot, information retrieval, hybrid ranking*

Abstrak - Kemajuan pesat Large Language Models (LLM) memungkinkan pemanfaatannya dalam sistem informasi medis, meskipun tantangan seperti halusinasi, ketidaksesuaian domain, dan kurangnya dasar pengetahuan yang terverifikasi masih signifikan, khususnya pada bahasa berdaya sumber rendah seperti Bahasa Indonesia. Penelitian ini memperkenalkan *chatbot* medis berbahasa Indonesia berbasis model open-source GPT-OSS-20B yang diperkuat melalui *pipeline Retrieval-Augmented Generation* (RAG). Sistem mengkombinasikan temu balik semantik menggunakan jina-embeddings-v3, *re-ranking* leksikal dengan algoritma BM25, serta *filter* domain ringan berbasis Logistic Regression sebagai *filter* awal untuk mencegah penggunaan LLM diluar domain. Evaluasi menggunakan artikel medis Indonesia serta percakapan pasien–dokter beranotasi menunjukkan bahwa *filter* domain bekerja sempurna pada data sintetis, namun menghasilkan misklasifikasi pada *query* alami. Reranker denga pembobotan *hybrid* (FAISS L2 + BM25) memberikan kinerja terbaik dengan akurasi Top-30 sebesar 0.699. *Blackbox testing* menunjukkan bahwa alur sistem berfungsi sesuai rancangan, meskipun kualitas respon belum divalidasi oleh pakar klinis. Temuan ini menunjukkan bahwa LLM *open-source* berbasis RAG dapat meningkatkan akses informasi medis berbahasa Indonesia, namun tetap memiliki keterbatasan penting seperti ketiadaan validasi klinis, potensi kesalahan data hasil *scraping*, serta ketahanan filter domain yang belum optimal.

Kata Kunci – Retrieval Augmented Generation, GPT-OSS, chatbot medis, pengambilan informasi, hybrid ranking.

I. PENDAHULUAN

Sistem pelayanan kesehatan di Indonesia masih memiliki tantangan signifikan, mulai dari keterbatasan fasilitas dan tenaga medis hingga hambatan geografis yang menyebabkan akses layanan kesehatan tidak merata [1], [2]. Dalam konteks ini, kemajuan kecerdasan buatan (AI) dan LLM berpotensi meningkatkan akses masyarakat terhadap informasi kesehatan atau medis melalui interaksi teks yang intuitif [3]. Namun, penerapan LLM dalam domain medis tidak bebas dari masalah, karena model umum dilatih menggunakan data non-medis berskala besar sehingga rawan bias dan halusinasi [4]. Penelitian sebelumnya mengenai *LLM-Based Information Retrieval for Disease Detection* menunjukkan bahwa analisis semantik mampu mengidentifikasi penyakit berbasis gejala, tetapi akurasinya masih terbatas karena tidak adanya re-ranking yang optimal dan belum adanya kontrol terhadap pertanyaan non-medis [5]. Sementara itu, pendekatan RAG menawarkan solusi lebih aman karena memungkinkan LLM merujuk pada dokumen medis yang relevan tanpa proses pelatihan ulang yang mahal [6]. Di sisi lain *chatbot* mampu memberikan respon yang lebih manusiawi, sehingga diharapkan mampu memberikan penjelasan yang mudah dimengerti oleh pengguna.

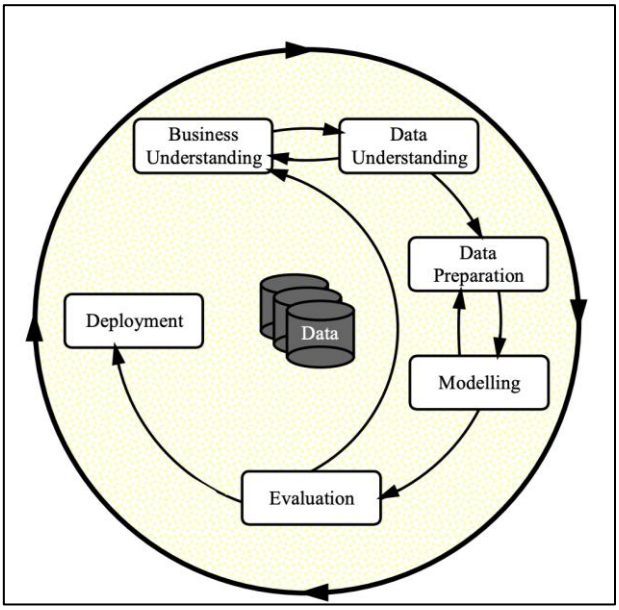
Dari berbagai temuan tersebut, *gap* penelitian yang muncul adalah ketiadaan sistem *chatbot* medis berbahasa Indonesia yang menyatukan *domain filtering*, *retrieval* semantik–leksikal secara hybrid, RAG sebagai penyedia konteks relevan, dan evaluasi *end-to-end* hingga tingkat interaksi pengguna. Kekurangan utama penelitian sebelumnya [5], yaitu belum adanya *domain filtering*, belum digunakannya *retrieval* semantik-leksikal, serta belum diintegrasikan dalam bentuk *chatbot* menjadi motivasi utama pengembangan arsitektur yang lebih komprehensif pada penelitian ini. Tujuan penelitian ini adalah merancang dan mengevaluasi chatbot medis Bahasa Indonesia berbasis GPT-OSS-20B dan RAG yang mampu menyediakan jawaban relevan dan terstruktur menggunakan basis pengetahuan medis terverifikasi. Rumusan pertanyaan penelitiannya adalah:

1. Bagaimana efektivitas domain filter dalam membedakan pertanyaan medis dan non-medis?
2. Bagaimana kinerja *retrieval* semantik-leksikal dibandingkan metode tunggal dalam skenario pencarian penyakit berbasis gejala?
3. Bagaimana hasil evaluasi *end-to-end* melalui *blackbox testing* dalam memvalidasi alur sistem *chatbot* medis?

II. SIGNIFIKANSI STUDI

Dalam penelitian ini digunakan model GPT-OSS, sebuah LLM *open-source* yang memungkinkan implementasi *on-premise* dengan biaya inferensi rendah dan tingkat transparansi tinggi karena bobot model sepenuhnya dapat diakses [7]. Transparansi ini penting dalam konteks medis karena memungkinkan audit independen terhadap bias serta evaluasi risiko penggunaan model non-klinis. Meskipun GPT-OSS menunjukkan kinerja kompetitif terhadap model *open-source* lain seperti LLaMA [8], DeepSeek-R1 [9], Gemma [10], dan Qwen [11], model ini tidak dilatih secara khusus pada domain medis. Keterbatasan tersebut menimbulkan risiko bias domain, ketidaktepatan penalaran medis, dan potensi halusinasi, sehingga penelitian ini menempatkan RAG sebagai mekanisme mitigasi risiko, bukan sebagai jaminan akurasi klinis. Untuk mendukung mitigasi tersebut, penelitian ini memanfaatkan jina-embeddings-v3 sebagai komponen pemetaan semantik yang efektif untuk Bahasa Indonesia [12], dan BM25 sebagai pelengkap berbasis leksikal dalam tahap re-ranking. Pendekatan *hybrid* ini menggabungkan keunggulan pemahaman kontekstual dan pencocokan kata kunci, yang telah terbukti meningkatkan kualitas temu balik dokumen pada sistem pencarian medis [13], [14]. Dengan mengintegrasikan dokumen yang telah diurutkan ulang ke dalam *pipeline* RAG, sistem dirancang agar respons LLM selalu merujuk pada sumber yang relevan guna menurunkan risiko kesalahan penalaran.

Penelitian ini juga menggunakan kerangka CRISP-DM untuk memastikan proses pengembangan sistem berlangsung terstruktur dan dapat direplikasi. Kerangka ini mencakup tahapan *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* [15], [16], yang secara metodologis mendukung konstruksi sistem *chatbot* berbasis LLM dan RAG tanpa mengaburkan fokus utama bagian signifikansi studi. Pada Gambar 1, dapat dilihat alur proses dari CRISP-DM.



Gambar 1. Proses, alur, serta fase – fase pada kerangka kerja CRISP-DM [16]

A. *Business Understanding*

Tahap ini bertujuan memahami permasalahan inti pada sistem layanan informasi medis di Indonesia, yaitu keterbatasan akses informasi medis berkualitas dan ketimpangan distribusi tenaga kesehatan. Sistem *chatbot* yang dibangun diarahkan untuk:

1. Memberikan informasi medis dasar dalam Bahasa Indonesia secara cepat, relevan, dan berbasis sumber pengetahuan terverifikasi.
2. Menggunakan pendekatan RAG untuk mengurangi risiko halusinasi LLM.

Tujuan operasional tahap ini adalah mendefinisikan kebutuhan sistem, alur kerja, serta kriteria evaluasi kinerja model.

B. *Data Understanding*

Data penelitian terdiri dari artikel penyakit hasil *web scraping* dari berbagai sumber medis daring terpercaya, di mana setiap artikel berbahasa Indonesia berisi informasi lengkap mengenai definisi, gejala, penyebab, faktor risiko, dan penanganan umum untuk satu entitas penyakit. Data uji diperoleh dari percakapan pasien–dokter pada forum kesehatan yang telah dilabeli oleh ekspertis dengan satu atau lebih nama penyakit relevan, dan digunakan untuk mengevaluasi kinerja temu balik informasi. Seluruh data berasal dari sumber publik pada situs kesehatan resmi dan portal medis umum, tidak mengandung identitas pribadi, serta tidak memerlukan proses de-identification; namun, penelitian ini tetap mengakui bahwa konten daring dapat mengandung variasi kualitas yang perlu diperhatikan dalam interpretasi hasil.

C. *Data Preparation*

Tahap ini mencakup pembersihan dan penyiapan data sebelum digunakan sebagai basis pengetahuan dan sebagai sampel pengujian.

1. Praproses Data

Praproses yang dilakukan pada artikel penyakit dan teks pertanyaan dari percakapan pasien-dokter adah permbersihan teks. Beberapa langkah praproses meliputi:

1) Menghapus elemen HTML yang tidak diperlukan.
2) Menghilangkan karakter non-alfabet.
3) Mengganti karakter baris baru menjadi karakter spasi tunggal.
4) Mengganti spasi ganda menjadi spasi tunggal
5) Normalisasi teks (pengubahan huruf menjadi lowercase).

2. Transformasi Teks Menjadi Vektor

Untuk membentuk basis data vektor, digunakan model jina-embeddings-v3. Data artikel penyakit yang bersih dikonversi menjadi vektor berdimensi 1024, kemudian disimpan pada basis data FAISS [17]. Sementara itu, data teks pertanyaan pasien yang bersih juga dikonversi menjadi vektor dengan dimensi yang sama dan digunakan pada tahap pengujian untuk dicocokkan dengan vektor data artikel penyakit.

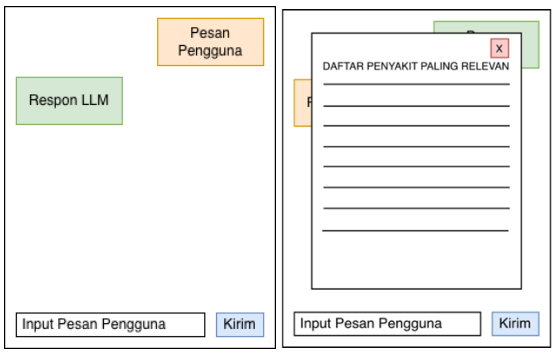
3. Pembentukan Label Pengujian

Data percakapan pasien digunakan sebagai *query*. Setiap percakapan telah memiliki label penyakit, sehingga dapat digunakan sebagai acuan evaluasi akurasi pada skenario *top-k retrieval*.

D. *Modeling*

Tahap ini membahas perancangan arsitektur sistem, pemodelan klasifikasi, proses RAG, dan penggunaan LLM. Secara keseluruhan, sistem terdiri dari beberapa komponen utama:

1. Pengembangan Antarmuka Aplikasi

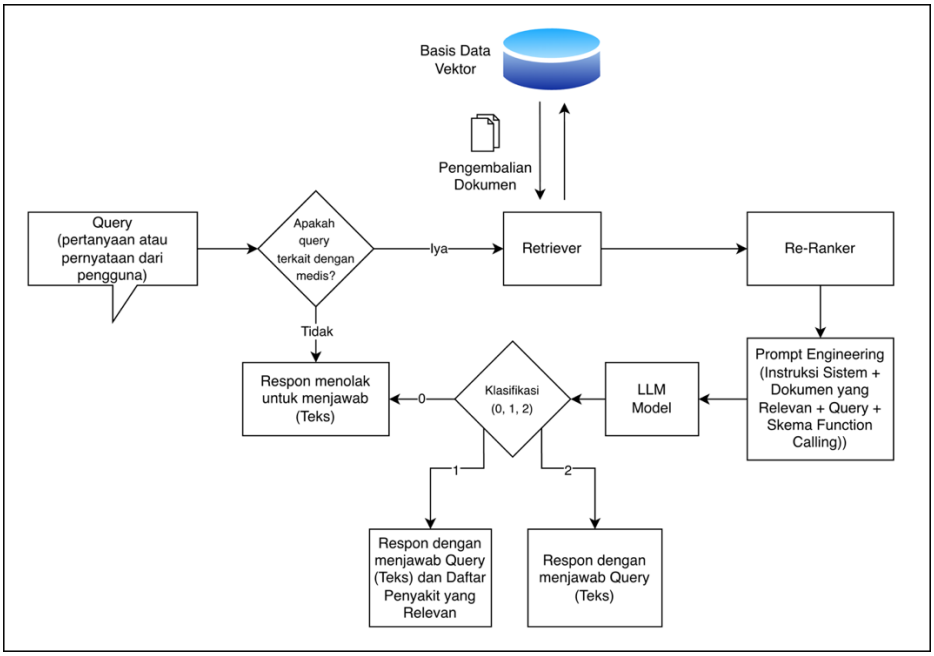


(a) (b)

Gambar 2. (a) Antarmuka aplikasi antar pengguna dengan bot (b) tampilan *pop-up* yang berisi penyakit yang relevan berdasarkan input pengguna

Antarmuka aplikasi dikembangkan berbasis web dengan menggunakan HTML, CSS, dan JavaScript. Pada Gambar 2, dapat dilihat mockup aplikasi. Antarmuka aplikasi akan menampilkan interkasi berupa teks antar pengguna dan bot. Selain itu jika pertanyaan pengguna terkait dengan diagnosis medis, maka aplikasi juga dapat menampilkan penyakit yang paling relevan dalam bentuk tampilan *pop-up*.

2. Pengembangan *Application Programming Interface* (API)



Gambar 3. Alur pemrosesan input pengguna dari awal hingga respon sistem pada input tersebut

Gambar 3 menunjukkan alur kerja sistem RAG untuk domain medis. Ketika pengguna mengajukan pertanyaan, sistem pertama-tama menerapkannya pada filter Logistic Regression yang dilatih menggunakan 100 data sintetis seimbang untuk memisahkan pertanyaan medis dan non-medis; pertanyaan non-medis langsung ditolak guna menghemat biaya komputasi. Jika pertanyaan diklasifikasi sebagai medis, sistem melakukan *retrieval* konteks menggunakan embedding jina-embeddings-v3 yang dicocokkan dengan basis data vektor FAISS untuk memperoleh dokumen paling relevan secara semantik. Dokumen-dokumen awal ini kemudian diurutkan ulang dengan BM25 agar urutan relevansinya lebih akurat berdasarkan kecocokan leksikal. Dokumen - dokumen teratas kemudian disusun ke dalam prompt berisi instruksi sistem, konteks medis, query pengguna, dan skema *function calling*, lalu diproses oleh GPT-OSS-20B untuk melakukan klasifikasi lanjutan menjadi tiga kategori: non-medis (0), medis berbasis gejala (1), dan medis non-gejala (2). Sistem kemudian menghasilkan respons terstruktur melalui fungsi *medical_assistant*: daftar kemungkinan penyakit untuk kategori gejala, penjelasan ringkas untuk kategori medis umum, dan penolakan untuk kategori non-medis. GPT-OSS-20B menggunakan parameter *default* yaitu temperature = 0.6, max_completions_token = 1024, dan response_format = text. Dengan alur ini, sistem dapat memberikan jawaban yang aman, relevan, dan terstruktur melalui kombinasi filter awal, *retrieval* semantik, serta *reranking* berbasis BM25.

Pada Gambar 4 dapat dilihat *prompt* yang digunakan pada penelitian ini. Prompt ini mengarahkan model untuk mengklasifikasi jenis pertanyaan, kemudian menghasilkan respons ringkas dalam Bahasa Indonesia dengan memanfaatkan konteks medis yang sudah disediakan. Semua keluaran wajib diberikan melalui fungsi *medical_assistant*, sehingga alur sistem tetap konsisten dan sesuai format yang diharapkan tanpa menghasilkan teks di luar struktur tersebut. Sementara pada Gambar 5 dapat dilihat skema *function calling* dari fungsi medical_assistant. Fungsi tersebut menerima dua parameter wajib: *classification* (kategori medis dengan gejala/medis umum/non-medis) dan *response* (jawaban singkat untuk pengguna). Dengan struktur ini, model harus mengembalikan output dalam format objek terstruktur yang memuat hasil klasifikasi serta respons akhir, sehingga sistem dapat memproses jawaban secara konsisten dan terstandarisasi.

```
You are a medical assistant specialized in processing queries and respond it in Bahasa Indonesia.
You performs classification and reasoning using one tool: `medical_assistant`.

Here is the relevant medical context retrieved from your knowledge base in Bahasa Indonesia:
{context}

You must:
1. Classify the user's question into one of:
- 1: medical (asks about symptoms)
- 2: medical (non-symptom)
- 0: non-medical

2. Generate a short, accurate response using the context above:
- If 1 → Suggest possible diseases or conditions (use retrieved context when possible).
- If 2 → Briefly explain the mentioned condition.
- If 0 → Politely say you can only answer medical questions.

Return your answer ONLY via the `medical_assistant` function.
```

Gambar 4. *Prompt* yang digunakan pada sistem, {context} tempat penyisipan pengetahuan dari hasil temu balik informasi

```
"type": "function",
"function": {
    "name": "medical_assistant",
    "description": (
        "Classify a user query in Bahasa Indonesia as medical/non-medical and provide an
         appropriate response. "
         "If medical and mentions symptoms, suggest possible diseases (RAG-style). "
         "If medical but not about symptoms, briefly explain the condition. "
         "If non-medical, politely reject."
    ),
    "parameters": {
        "type": "object",
        "properties": {
            "classification": {
                "type": "integer",
                "description": (
                    "Query classification: 1 = medical (symptoms), "
                    "2 = medical (non-symptoms), 0 = non-medical."
                ),
            },
            "response": {
                "type": "string",
                "description": "LLM's concise final response to the user.",
            },
        }
        "required": ["classification", "response"],
    },
}
```

Gambar 5. Struktur dari fungsi "*medical-assistant*" yang berfungsi untuk menginstruksikan LLM untuk melakukan klasifikasi dan memberikan jawaban yang sesuai.

E. *Evaluation*

Evaluasi dilakukan pada tiga komponen utama sistem, yaitu model Logistic Regression sebagai filter awal, modul *retrieval* (embedding *jina-embeddings-v3*, FAISS, dan reranker BM25), serta evaluasi akhir melalui *blackbox testing* terhadap keluaran LLM GPT-OSS-20B dan antarmuka aplikasi.

1. Evaluasi Filter Awal

Kinerja model filter awal diukur menggunakan metrik umum klasifikasi: *precision*, *recall*, *F1-score*, dan *accuracy* [18]. Mengingat konteks medis, recall pada kelas medis menjadi prioritas agar meminimalkan *false negative*. Berikut merupakan formula dari metrik evaluasi :

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$F1 = 2\,.\frac{Precision.Recall}{Precision\,.\ Recall} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Keterangan :
TP : *True positive*
TN : *True negative*
FP : *False positive*
FN : *False negative*

2. Evaluasi Sistem Temu Balik

Tahap evaluasi pada proses temu balik dilakukan dengan menilai kemampuan sistem dalam menemukan dokumen penyakit yang benar pada berbagai batas peringkat (*Top-K*). Karena setiap penyakit hanya direpresentasikan oleh satu dokumen, maka evaluasi dilakukan secara langsung melalui pengukuran *Top-K Accuracy* [19], yaitu persentase pengembalian dokumen yang cocok dengan *query* pada posisi ≤ K. Eksperimen dilakukan untuk nilai K = 1, 5, 10, 20, dan 30.

$$Top - K\ Accuracy = \frac{1}{N}\sum_{i=0}^{N} 1(rank_i \leq K) \quad (5)$$

Keterangan :
N : jumlah *query*
$rank_i$: posisi dokumen relevan untuk *query* ke-i
$1(.)$: fungsi indikato (bernilai 1 jika benar, 0 jika salah)

Metrik ini mencerminkan secara langsung keberhasilan model temu balik dalam menempatkan dokumen yang benar pada posisi terdepan. Akurasi Top-1 menilai apakah model berhasil menempatkan dokumen benar pada peringkat pertama; sedangkan Top-5, Top-10, Top-20, dan Top-30 mengukur toleransi sistem pada batas peringkat yang lebih lebar.

3. Evaluasi Sistem dan Antarmuka Aplikasi

Pengujain atau evaluasi pada sistem dan antarmuka aplikasi bertujuan untuk memastikan sistem dan antarmuka aplikasi bekerja sesuai dengan skenario yang telah ditentukan. Hal ini dilakukan dengan menggunakan metode *blackbox testing* [20]. Skenario pengujian dapa dilihat pada Tabel 1.

TABEL I
PENGUJIAN SISTEM DAN ANTARMUKA APLIKASI MELALUI BLACKBOX TESTING

Deskripsi Pengujian	Hasil yang Diharpkan	Indikator
Sistem menolak menjawab pada *query* yang tidak terkait dengan medis	Sistem menampilkan pesan penolakan menjawab *query* yang diberikan	Berhasil / Gagal
Sistem menjawab *query* medis terkait dengan gejala	Sistem menampilkan diagnosis awal terkait *query* dan memberikan daftar penyakit yang relevan	Berhasil / Gagal
Sistem menjawab *query* medis umum	Sistem menampilkan penjelasan terkait dengan *query* tanpa memberikan daftar penyakit	Berhasil / Gagal

III. HASIL DAN PEMBAHASAN

Bagian ini membahas hasil evaluasi terhadap kinerja sistem yang meliputi efektivitas dari kinerja filter awal, sistem temu balik, keseluruhan sistem dan antarmuka aplikasi dalam memberikan informasi medis. Hasil-hasil ini dibahas secara analitis untuk menilai reliabilitas, ketepatan, dan keterbatasan pendekatan yang digunakan dalam pengembangan chatbot medis berbasis RAG. Aplikasi dapat di akses di https://dyagnosa.com/chat.html.

A. Hasil Evaluasi Filter Awal

TABEL II
HASIL EVALUASI KLASIFIKASI MODEL LOGISTIC REGRESSION

Metrik	Nilai
Precision	1.0
Recall	1.0
Accuracy	1.0
F1	1.0

Filter awal berbasis Logistic Regression digunakan untuk menyaring pertanyaan non-medis sebelum masuk ke tahap RAG sehingga beban komputasi LLM dapat ditekan. Model dilatih menggunakan data sintetis berbahasa Indonesia dengan representasi TF-IDF (n-gram 1–2, 5000 fitur) dan dievaluasi menggunakan 5-fold cross-validation, menghasilkan metrik sempurna pada seluruh indikator seperti ditunjukkan pada Tabel 2. Namun, evaluasi lanjutan pada query dunia nyata menunjukkan bahwa kinerja tersebut tidak sepenuhnya mencerminkan kemampuan model terhadap variasi bahasa alami. Beberapa kalimat sapaan atau pertanyaan sosial seperti “Halo, apakabar? Saya harap anda sehat.” atau “Apakah benar jodoh, sakit, dan mati merupakan takdir dari yang maha kuasa?” tidak terdeteksi sebagai non-medis. Hal ini dapat dijelaskan oleh keterbatasan representasi TF-IDF yang hanya menghitung frekuensi kata tanpa memahami konteks, sehingga kemunculan kata sehat atau sakit secara dangkal dianggap sebagai indikator medis.

Keterbatasan tersebut memiliki dampak praktis, yaitu semakin banyak *query* non-medis yang tidak tersaring, semakin sering LLM harus melakukan inferensi, yang pada akhirnya meningkatkan biaya komputasi sistem. Meskipun demikian, Logistic Regression tetap dipertahankan karena menawarkan efisiensi tinggi dibandingkan model yang lebih kompleks seperti deep learning, sehingga masih sesuai dengan tujuan desain sistem yaitu menyediakan mekanisme filtrasi yang ringan, cepat, dan *cost-efficient*. Dengan demikian, kinerja sempurna pada dataset terkontrol perlu dipahami sebagai hasil pada kondisi ideal, sementara tantangan pada bahasa alami menunjukkan ruang perbaikan bagi filter domain di masa mendatang.

B. Hasil Evaluasi Sistem Temu Balik

Data sistem temu balik terdiri dari 1.366 artikel penyakit dan 135.276 percakapan dokter–pasien sebagai *query*. Proses *retrieval* dimulai dengan mengonversi dokumen dan *query* menjadi vektor menggunakan jina-embeddings-v3, lalu menghitung jarak L2 untuk memilih 100 dokumen awal. Dokumen tersebut kemudian diberi peringkat ulang menggunakan tiga metode: L2 berbasis semantik, BM25 berbasis kata kunci, dan Hybrid (L2+BM25) yang menggabungkan kedua pendekatan secara seimbang. Hasil evaluasi pada Tabel 3 menunjukkan perbedaan karakteristik yang jelas antara ketiganya: L2 unggul pada Top-1 dan Top-5 karena kemampuan menangkap kedekatan makna, BM25 menghasilkan skor rendah di Top-1 tetapi meningkat signifikan pada Top-30, sementara Hybrid menjadi metode paling stabil pada hampir seluruh nilai K—terutama Top-20 dan Top-30—dengan memanfaatkan kelebihan semantik dan leksikal sekaligus [13].

TABEL III
TOP-K ACCURACY DARI HASIL EVALUASI SISTEM TEMU BALIK PADA BERBAGAI METODE PENILAIAN RELEVANSI DOKUMEN

K	L2	Hybrid (L2+BM25)	BM25
1	**0.246**	0.239	0.118
5	**0.455**	0.454	0.292
10	0.550	**0.551**	0.396
20	0.643	**0.645**	0.518
30	0.695	**0.699**	0.596

Namun, akurasi Top-1 dan Top-5 yang masih rendah mengindikasikan bahwa sistem belum mampu secara konsisten menempatkan dokumen penyakit yang benar pada urutan teratas, terutama karena pertanyaan pasien bersifat ambigu dan jarang menyebutkan gejala secara eksplisit. Oleh karena itu, kinerja terbaik muncul pada rentang Top-20 hingga Top-30, di mana pencarian semantik dan sinyal leksikal saling melengkapi dalam menangani variasi bahasa sehari-hari. Temuan ini menunjukkan bahwa sistem lebih sesuai digunakan sebagai *informational support* untuk penelusuran pengetahuan medis dasar, bukan sebagai alat diagnosis klinis, sejalan dengan konsep *hybrid retrieval* yang dijelaskan oleh Lu dkk.

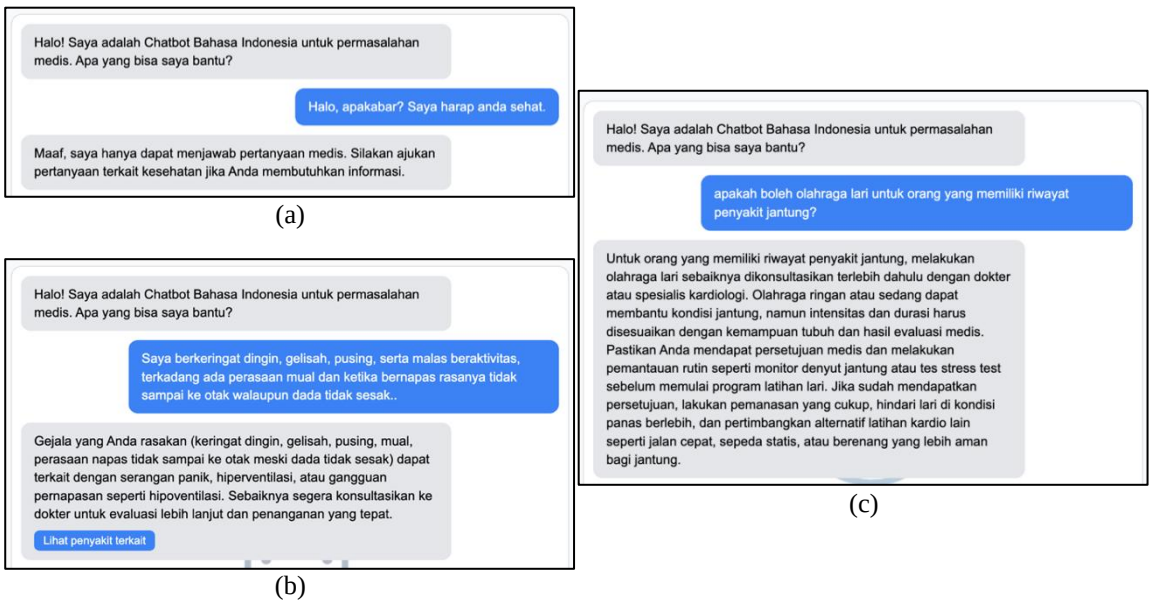
C. Hasil Evaluasi Sistem dan Antarmuka Aplikasi

TABEL IV
HASIL PENGUJIAN BLACKBOX TESTING

Deskripsi Pengujian	Indikator
Sistem menolak menjawab pada *query* yang tidak terkait dengan medis	Berhasil
Sistem menjawab *query* medis terkait dengan gejala	Berhasil
Sistem menjawab *query* medis umum	Berhasil

Pengujian alur sistem dan antarmuka dilakukan melalui blackbox testing untuk memeriksa kesesuaian output terhadap input pengguna, mencakup tiga skenario utama yaitu penolakan *query* non-medis, respons terhadap pertanyaan medis berbasis gejala, dan respons terhadap pertanyaan medis umum, yang seluruhnya memperoleh hasil "Berhasil" sebagaimana ditampilkan pada Tabel 4. Namun, cakupan skenario masih terbatas dan belum mewakili variasi input yang lebih kompleks, sehingga pengujian ini lebih berfungsi untuk memvalidasi alur kerja sistem daripada menilai akurasi medis atau aspek keamanan klinis.

Pada Gambar 6 dapat dilihat kesesuain respon dari sistem terhadap *query* yang diberikan.



Gambar 6. (a) Chatbot menolak menjawab pertanyaan non-medis, (b) chatbot menjawab pertanyaan medis terkait gejala dan mengembalikan penyakit yang relevan, dan (c) chatbot menjawab pertanyaan medis terkait non gejala

Pada Gambar 7 dapat terlihat daftar penyakit yang relevan terhadap *query* yang ada pada Gambar 6b.

Gambar 7. Tampilan daftar penyakit relevan ketika melakukan klik pada tombol “Lihat penyakit terkait” yang muncul pada kotak percakapan.

IV. KESIMPULAN

Bagian ini merangkum temuan utama penelitian sekaligus mengevaluasi pencapaian tujuan sistem yang dikembangkan. Selain itu, bagian ini menguraikan implikasi praktis dari hasil evaluasi serta batasan sistem yang perlu diperhatikan dalam penerapan chatbot medis berbasis RAG. Beberapa rekomendasi pengembangan juga disampaikan untuk meningkatkan keandalan dan keamanan sistem di masa mendatang.

A. Simpulan

Penelitian ini berhasil merancang prototipe chatbot medis berbahasa Indonesia berbasis LLM dan RAG dengan memanfaatkan GPT-OSS-20B, jina-embeddings-v3, dan BM25 sebagai reranker. Hasil evaluasi menunjukkan bahwa:

1. Filter awal Logistic Regression memperoleh nilai sempurna pada dataset sintetis dan membantu menurunkan beban inferensi LLM, meskipun masih belum robust terhadap variasi bahasa natural sehingga beberapa pertanyaan non-medis tetap lolos ke tahap LLM.
2. *Hybrid retrieval* (L2+BM25) memberikan kinerja terbaik dalam pemeringkatan dokumen, khususnya pada rentang Top-K yang lebih besar (Top-20 dan Top-30), dengan akurasi mencapai 0.699 pada K = 30. Namun, kinerja Top-1 dan Top-5 yang masih rendah menunjukkan bahwa sistem belum konsisten menemukan dokumen relevan pada posisi teratas.
3. Pengujian blackbox menunjukkan bahwa alur sistem berjalan sesuai desain dan respons dasar dapat diberikan pada *query* medis maupun non-medis. Namun, pengujian ini belum mencakup variasi kasus kompleks dan tidak menilai kualitas klinis respons.
4. Integrasi RAG membantu menyediakan konteks tambahan sehingga berpotensi mengurangi kecenderungan halusinasi, tetapi tidak menghilangkannya sepenuhnya. Basis pengetahuan yang berasal dari scraping juga dapat mengandung kesalahan, sehingga akurasi respons masih bergantung pada kualitas data sumber dan kelengkapan konteks yang berhasil di-retrieve.

Dengan demikian, arsitektur yang diusulkan dapat menjadi fondasi awal untuk sistem *chatbot* medis berbahasa Indonesia, namun belum dapat dianggap layak sebagai solusi klinis. Ketiadaan uji klinis, rendahnya Top-1 accuracy, belum adanya evaluasi pakar, serta keterbatasan domain filter menjadi faktor yang perlu diperbaiki sebelum sistem dapat digunakan dalam konteks layanan medis yang sensitif.

B. Saran

Beberapa rekomendasi berikut dapat dipertimbangkan untuk meningkatkan kinerja dan keamanan sistem pada pengembangan selanjutnya:

1. Mengganti atau melatih ulang filter awal dengan data asli dan lebih beragam agar ketahanan terhadap variasi bahasa natural meningkat, sekaligus mengurangi risiko pertanyaan non-medis mencapai tahap LLM.
2. Menerapkan reranker berbasis neural seperti cross-encoder apabila sumber daya mencukupi, untuk memperbaiki akurasi Top-1 yang kritis dalam sistem RAG.

3. Memperluas dan membersihkan basis pengetahuan dengan memasukkan artikel medis yang tervalidasi, berlisensi jelas, serta melalui proses kurasi untuk meminimalkan kesalahan akibat data scraping.
4. Melakukan human expert evaluation untuk menilai kesesuaian medis, keamanan informasi, dan potensi risiko klinis pada hasil respons LLM.
5. Menambahkan analisis risiko terkait bias data, batasan penalaran LLM non-klinis, serta prosedur mitigasi halusinasi sebelum sistem digunakan dalam skenario konsultasi medis publik.

REFERENSI

[1] Y. Mahendradhata *et al.*, "The Capacity of the Indonesian Healthcare System to Respond to COVID-19," *Front Public Health*, vol. 9, Jul. 2021, doi: 10.3389/fpubh.2021.649819.

[2] S. Wenang *et al.*, "Availability and Accessibility of Primary Care for the Remote, Rural, and Poor Population of Indonesia," *Front Public Health*, vol. 9, Sep. 2021, doi: 10.3389/fpubh.2021.721886.

[3] J. Yang *et al.*, "Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond," *ACM Trans Knowl Discov Data*, vol. 18, no. 6, pp. 1–32, Jul. 2024, doi: 10.1145/3649506.

[4] Y. Guo *et al.*, "Bias in Large Language Models: Origin, Evaluation, and Mitigation," Nov. 2024.

[5] M. A. Abdurrazzaq, E. L. Tjiong, and K. A. Wanady, "LLM-Based Information Retrieval for Disease Detection Using Semantic Similarity," *Jurnal Online Informatika*, vol. 10, no. 1, pp. 32–41, Apr. 2025.

[6] C. Njeh, H. Nakouri, and F. Jaafar, "Enhancing RAG-Retrieval to Improve LLMs Robustness and Resilience to Hallucinations," 2025, pp. 201–213. doi: 10.1007/978-3-031-74186-9_17.

[7] OpenAI *et al.*, "gpt-oss-120b & gpt-oss-20b Model Card," Aug. 2025.

[8] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023.

[9] DeepSeek-AI *et al.*, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," Jan. 2025.

[10] Gemma Team *et al.*, "Gemma: Open Models Based on Gemini Research and Technology," Apr. 2024.

[11] J. Bai *et al.*, "Qwen Technical Report," Sep. 2023.

[12] S. Sturua *et al.*, "jina-embeddings-v3: Multilingual Embeddings With Task LoRA," Sep. 2024.

[13] J. Lu, K. Hall, J. Ma, and J. Ni, "HYRR: Hybrid Infused Reranking for Passage Retrieval," Dec. 2022.

[14] L. Gao, Z. Dai, T. Chen, Z. Fan, B. Van Durme, and J. Callan, "Complementing Lexical Retrieval with Semantic Residual Embedding," Mar. 2021.

[15] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput Sci*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[16] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining."

[17] M. Douze *et al.*, "The Faiss library," Oct. 2025.

[18] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci Rep*, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[19] F. Petersen, H. Kuehne, C. Borgelt, and O. Deussen, "Differentiable Top-k Classification Learning," Jun. 2022.

[20] Aisya Tyanafisya *et al.*, "BLACK BOX TESTING USING EQUIVALENCE PARTITIONING TECHNIQUE ON BAKKAR WEBSITE," *Jurnal INSTEK (Informatika Sains dan Teknologi)*, vol. 10, no. 1, pp. 230–238, May 2025, doi: 10.24252/instek.v10i1.52951.

UCAPAN TERIMA KASIH

Penelitian ini didanai oleh Direktorat Penelitian dan Pengabdian Masyarakat, Direktorat Jenderal Riset dan Pengembangan, Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik Indonesia melalui pendanaan riset tahun anggaran 2025 dengan nomor kontrak induk 124/C3/DT.05.00/PM/2025 serta nomor kontrak turunan 0981/LL3/AL.04/2025 dan 006/LPPM-SRT/UK/VI/2025. Penulis mengucapkan terima kasih atas dukungan yang diberikan sehingga penelitian ini dapat terlaksana dengan baik.