Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal: 150-159 e-ISSN: 2987-7016

Implementasi Arsitektur Hybrid RAG pada Chatbot Skrining Kesehatan Mental

M Dhimas Hadid [1], Nazruddin Safaat H [2], Muhammad Irsyad [3], Febi Yanto [4]
[1,2,3,4] Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau
[1]12250111231@students.uin-suska.ac.id, [2]nazruddin.safaat@uin-suska.ac.id*, [3]irsyadtech@uin-suska.ac.id, [4]febiyanto@uin-suska.ac.id

Abstract

Worker mental health is a critical global challenge. In Indonesia, only 57% of individuals with mental disorders receive mental health services, falling short of the national target of 90%. Conventional Retrieval-Augmented Generation (RAG) systems are not designed to simultaneously handle psychometric assessment data and clinical literature retrieval without increasing hallucination risk. This study proposes a Hybrid RAG architecture with a layered Query Router to integrate deterministic psychometric interpretation and semantically grounded clinical knowledge retrieval. A mental health screening chatbot was developed on Telegram using a dual-path architecture that separates relational processing through PostgreSQL from semantic retrieval through Qdrant, BM25, and Distribution-Based Score Fusion (DBSF). The system is governed by a four-layer Query Router and a Routing-Aware Quality Gate with a Fail-Closed policy. Five psychometric instruments were implemented: WHO-5, GAD-7, M-TBI, K10, and NAQ-R. Evaluation employed LaBSE BERTScore, E5 Cosine Similarity, and RAGAS LLM-as-a-Judge. Results show that the Query Router achieved 100% classification accuracy on 100 test queries. The Hybrid RAG configuration with DBSF obtained the highest LaBSE F1-score of 0.7039 and E5 Cosine Similarity of 0.9234. Furthermore, a Faithfulness score of 0.9350 indicates that most generated claims are supported by retrieved clinical documents. The moderate Answer Correctness score of 0.7447 is attributed to cross-lingual evaluation limitations rather than deficiencies in the proposed architecture.

Keywords: Chatbot, Hybrid RAG, Mental Health Screening, Query Router, Telegram

Abstrak

Kesehatan mental pekerja merupakan tantangan global yang mendesak. Di Indonesia, hanya 57% penyandang gangguan jiwa yang memperoleh layanan kesehatan mental, masih jauh dari target nasional sebesar 90%. Sistem Retrieval-Augmented Generation (RAG) konvensional belum dirancang untuk menangani data asesmen psikometri dan pengambilan literatur klinis secara bersamaan tanpa meningkatkan risiko halusinasi. Penelitian ini mengusulkan arsitektur Hybrid RAG dengan Query Router berlapis untuk mengintegrasikan interpretasi psikometri yang bersifat deterministik dan retrieval pengetahuan klinis yang relevan secara semantik. Sebuah chatbot skrining kesehatan mental dikembangkan pada Telegram menggunakan arsitektur dua jalur yang memisahkan pemrosesan relasional melalui PostgreSQL dan retrieval semantik melalui Qdrant, BM25, serta Distribution-Based Score Fusion (DBSF). Sistem dikendalikan oleh Query Router empat lapis dan Routing-Aware Quality Gate dengan kebijakan Fail-Closed. Lima instrumen psikometri diimplementasikan, yaitu WHO-5, GAD-7, M-TBI, K10, dan NAQ-R. Evaluasi dilakukan menggunakan LaBSE BERTScore, E5 Cosine Similarity, dan RAGAS LLM-as-a-Judge. Hasil pengujian menunjukkan bahwa Query Router mencapai akurasi klasifikasi 100% pada 100 kueri uji. Konfigurasi Hybrid RAG dengan DBSF memperoleh nilai LaBSE F1 tertinggi sebesar 0,7039 dan E5 Cosine Similarity sebesar 0,9234. Selain itu, skor Faithfulness sebesar 0,9350 menunjukkan bahwa sebagian besar klaim yang dihasilkan didukung oleh dokumen klinis yang berhasil diambil sistem. Nilai Answer Correctness sebesar 0,7447 dipengaruhi keterbatasan evaluasi lintas bahasa, bukan kelemahan arsitektur yang diusulkan.

Kata kunci: Chatbot, Hybrid RAG, Query Router, Skrining Kesehatan Mental, Telegram

.

1. Pendahuluan

Kesehatan mental merupakan kondisi kesejahteraan psikologis yang memungkinkan individu mengelola tekanan hidup dan berkontribusi secara produktif [1]. Pandemi COVID-19 meningkatkan prevalensi gangguan mental secara global sebesar 25 persen [2], sementara lebih dari satu miliar individu kini hidup dengan gangguan mental yang menimbulkan kerugian ekonomi sekitar US$1 triliun per tahun [3].

Beban ini paling nyata dirasakan pekerja usia muda. Survei Deloitte Gen Z and Millennial 2024 menemukan bahwa hanya 51% Gen Z dan 56% milenial menilai kesehatan mental mereka dalam kondisi baik [4]. Di Indonesia, kondisi ini diperparah oleh keterbatasan tenaga kesehatan jiwa dimana satu psikiater melayani 250.000 penduduk sehingga hanya 57% penyandang gangguan jiwa yang terlayani dari target nasional 90% [5][6].

Gap layanan ini tidak hanya soal jumlah tenaga ahli, tetapi juga soal aksesibilitas skrining yang responsif. Platform yang tersedia seperti SATUSEHAT Mobile menyajikan instrumen psikometri dalam format statis

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

JUPTIK

2026 Vol. 4 No.1 Hal: 150-159 e-ISSN: 2987-7016

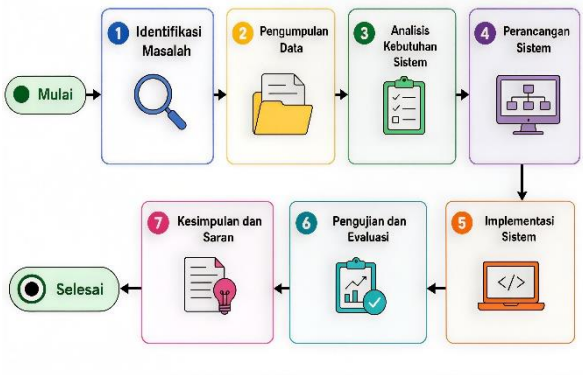
tanpa kemampuan tanya-jawab adaptif [7]. Penelitian menunjukkan bahwa format konversasional meningkatkan keterlibatan pengguna secara signifikan dibanding skrining web statis [8], namun chatbot berbasis aturan tidak mampu memberikan respons berbasis literatur klinis secara dinamis. Tinjauan sistematis Casu et al. [33] terhadap 15 studi intervensi chatbot AI untuk kesehatan mental mengonfirmasi manfaatnya dalam mengurangi gejala depresi dan kecemasan, namun mengidentifikasi tantangan integrasi dengan sistem layanan kesehatan sebagai hambatan utama adopsi skala luas.

Pendekatan Retrieval-Augmented Generation (RAG) menawarkan kemampuan tanya-jawab berbasis dokumen yang dinamis sebagai solusi atas keterbatasan tersebut [24]. Namun, sistem RAG konvensional menghadapi dua masalah yang selama ini belum diselesaikan secara bersamaan. Pertama, sistem tidak memisahkan data psikometri pengguna yang bersifat deterministik dari literatur klinis yang bersifat semantik, sehingga rentan menghasilkan interpretasi numerik yang salah [10]. Kedua, penerapan RAG pada bahasa Indonesia dengan domain spesifik menghadapi tantangan tersendiri, Suharyadi dan Saputra [34] menunjukkan bahwa pendekatan single-retrieval tidak cukup untuk sistem tanya-jawab domain spesifik berbahasa Indonesia karena variasi leksikal dan terminologi teknis tidak dapat ditangkap secara memadai oleh satu jalur retrieval saja; keterbatasan serupa berlaku pada domain klinis psikologi yang menjadi fokus penelitian ini.

Penelitian ini bertujuan membuktikan bahwa arsitektur Hybrid RAG dengan Query Router berlapis mampu menjawab kedua kebutuhan tersebut secara bersamaan, yaitu akurasi deterministik pada interpretasi data psikometri dan relevansi semantik pada pengambilan literatur klinis, melalui tiga ukuran: ketepatan klasifikasi rute, keunggulan strategi retrieval hybrid dibandingkan alternatifnya, dan minimnya halusinasi faktual pada jawaban yang dihasilkan.

2. Metode Penelitian

Tahapan penelitian terdiri atas tujuh tahap: (1) Identifikasi Masalah, (2) Pengumpulan Data, (3) Analisis Kebutuhan Sistem, (4) Perancangan Sistem, (5) Implementasi Sistem, (6) Pengujian dan Evaluasi, dan (7) Kesimpulan dan Saran. Pendekatan pengembangan yang digunakan adalah metode Prototyping yang bersifat iteratif [11]. Alur tahapan penelitian ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

Tahap identifikasi masalah dilakukan untuk mengenali keterbatasan layanan skrining kesehatan mental yang ada, khususnya bagi pekerja usia muda di Indonesia. Tahap pengumpulan data mencakup pengumpulan literatur klinis dari sumber otoritatif sebagai basis pengetahuan sistem. Tahap analisis menentukan kebutuhan fungsional dan nonfungsional, termasuk pemilihan instrumen psikometri dan strategi integrasi sistem. Tahap perancangan menghasilkan arsitektur Hybrid RAG yang diintegrasikan dengan antarmuka chatbot Telegram. Tahap implementasi merealisasikan arsitektur tersebut ke dalam kode menggunakan Python, FastAPI, Qdrant, dan PostgreSQL. Tahap pengujian menggunakan kombinasi tiga metrik evaluasi, yaitu BERTScore yang mengukur kualitas semantik teks menggunakan representasi BERT [12], E5 Cosine Similarity yang mengukur kemiripan semantik antara teks menggunakan model embedding multibahasa [13], serta RAGAS LLM-as-a-Judge yang mengevaluasi kualitas sistem RAG secara otomatis [14]. Tahap terakhir merangkum temuan dan merumuskan rekomendasi pengembangan.

2.1. Pengumpulan Data

Studi literatur mencakup teori kesehatan mental pekerja, instrumen psikometri, LLM, RAG, dan platform Telegram. Data penelitian terbagi dua, yaitu Basis Pengetahuan Klinis berupa literatur psikologi dan kesehatan untuk Jalur Semantik, dan Data Skrining Pengguna berupa skor kuesioner yang tersimpan di PostgreSQL untuk Jalur Relasional.

2.2. Analisis Kebutuhan Sistem

Analisis kebutuhan dilakukan untuk mengidentifikasi permasalahan yang menjadi dasar pengembangan sistem dan menentukan fungsi-fungsi yang harus dipenuhi.

Permasalahan utama yang ditemukan adalah belum tersedianya sistem pendukung skrining kesehatan mental yang mampu menggabungkan dua jenis data secara bersamaan: data hasil kuesioner psikometri yang bersifat terstruktur dan pasti, serta literatur klinis yang bersifat naratif dan memerlukan pemahaman konteks.

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal: 150-159 e-ISSN: 2987-7016

Sistem yang ada saat ini umumnya hanya menangani salah satu dari keduanya, sehingga tidak mampu memberikan respons yang akurat secara data dan relevan secara klinis [7][8].

Berdasarkan analisis kebutuhan di atas, sistem dirancang untuk mengukur kesehatan mental dari lima dimensi yang paling relevan dengan konteks pekerja di Indonesia. Pemilihan kelima instrumen berikut didasarkan pada validitas klinis yang sudah teruji, ketersediaan versi adaptasi bahasa Indonesia, dan kemampuan masing-masing instrumen dalam mengukur dimensi yang berbeda secara komplementer sehingga menghasilkan gambaran kondisi psikologis yang menyeluruh. Kelima instrumen tersebut ditunjukkan pada Tabel 1.

Tabel 1. Instrumen Skrining

Instrumen	Aspek	Butir
WHO-5	Kesejahteraan Psikologis	5
GAD-7	Kecemasan (Anxiety)	7
M-TBI	Burnout / Kelelahan Kerja	22
K10	Distres Psikologis	10
NAQ-R	Perundungan Kerja	22

Kelima instrumen dipilih berdasarkan tiga kriteria: validitas klinis yang telah teruji, ketersediaan versi adaptasi bahasa Indonesia, dan kemampuan masing-masing instrumen mengukur dimensi berbeda secara komplementer sehingga menghasilkan gambaran kondisi psikologis yang menyeluruh [15–22].

WHO-5 Well-Being Index mengukur kesejahteraan psikologis subjektif melalui lima butir pertanyaan yang mencakup suasana hati positif, ketenangan, vitalitas, dan keterlibatan dalam aktivitas sehari-hari. Instrumen ini dipilih karena berfungsi sebagai indikator awal kondisi mental secara umum; skor rendah pada WHO-5 menjadi sinyal untuk pemeriksaan lebih lanjut menggunakan instrumen lain yang lebih spesifik [15][16].

GAD-7 (Generalized Anxiety Disorder-7) mengukur tingkat kecemasan umum melalui tujuh butir yang merujuk pada frekuensi gejala kecemasan dalam dua minggu terakhir. Kecemasan merupakan gangguan mental yang paling umum dialami pekerja usia produktif, sehingga inklusi instrumen ini memiliki relevansi langsung dengan populasi sasaran sistem [17][18].

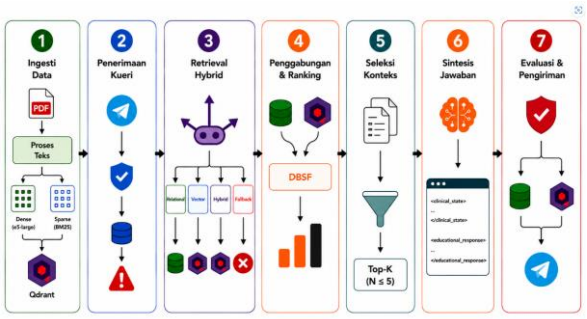
M-TBI (Maslach-Trisni Burnout Inventory) mengukur burnout atau kelelahan kerja melalui 22 butir yang mencakup tiga dimensi utama: kelelahan emosional, depersonalisasi, dan penurunan pencapaian pribadi. Instrumen ini dipilih secara khusus karena burnout merupakan kondisi yang secara eksklusif berakar pada konteks pekerjaan dan tidak tertangkap oleh instrumen kesejahteraan atau kecemasan umum [19].

K10 (Kessler Psychological Distress Scale) mengukur distres psikologis nonspesifik melalui 10 butir yang menilai tingkat kecemasan dan depresi secara bersamaan dalam 30 hari terakhir. Instrumen ini berperan sebagai skrining umum yang melengkapi GAD-7 pada dimensi depresif, sehingga kondisi komorbid kecemasan-depresi yang kerap terjadi pada pekerja dapat teridentifikasi [20].

NAQ-R (Negative Acts Questionnaire-Revised) mengukur paparan perundungan di tempat kerja melalui 22 butir yang mencakup tindakan negatif terkait pekerjaan, tindakan negatif personal, dan isolasi sosial. Perundungan kerja merupakan stresor lingkungan yang berdampak langsung pada kondisi mental pekerja namun tidak tercakup dalam instrumen klinis berbasis gejala, sehingga keberadaan NAQ-R menjadikan profil skrining sistem lebih komprehensif dan kontekstual [21][22].

2.3. Perancangan Sistem

Tahapan ini bertujuan merumuskan arsitektur logis sistem chatbot skrining kesehatan mental berbasis *Hybrid* RAG. Perancangan difokuskan pada pemisahan dua jalur pengambilan data yang bersifat berbeda: jalur deterministik untuk data psikometri terstruktur dan jalur semantik untuk literatur klinis, keduanya dikendalikan oleh Query Router yang menentukan jalur yang tepat secara otomatis berdasarkan jenis pertanyaan pengguna.



Gambar 2. Arsitektur Sistem Hybrid RAG Chatbot Skrining Kesehatan Mental

Arsitektur sistem ditunjukkan pada Gambar 2. Alur pemrosesan kueri pada sistem terdiri atas tujuh tahap yang berjalan secara sekuensial sebagai berikut.

(1) *Ingesti Data*. Sebelum sistem dapat melayani kueri pengguna, seluruh literatur klinis dalam format PDF diproses melalui tahap pra-pemrosesan data tekstual (text preprocessing) yang mencakup normalisasi teks, pembersihan derau (noise removal), dan pemecahan teks (chunking). Setiap chunk direpresentasikan dalam dua bentuk vektor secara bersamaan, yaitu dense vector menggunakan model multilingual-e5-large dan sparse vector menggunakan BM25, kemudian diindeks ke dalam koleksi Qdrant sebagai basis pengetahuan sistem.

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal: 150-159 e-ISSN: 2987-7016

(2) Penerimaan Kueri dan Pengamanan Awal. Pengguna mengirimkan kueri teks melalui antarmuka Telegram. Sistem melakukan verifikasi status onboarding pengguna dan kepemilikan sesi secara asinkron, kemudian menyimpan kueri ke database relasional untuk menjaga integritas riwayat transaksi. Sebelum proses retrieval dimulai, teks kueri dan riwayat obrolan dianalisis menggunakan pencocokan pola kata kunci untuk mendeteksi tendensi menyakiti diri sendiri atau bunuh diri. Apabila terdeteksi, pipeline RAG dilewati dan sistem langsung mengembalikan respons penanganan krisis standar beserta rujukan layanan kesehatan jiwa yang relevan.

(3) Hybrid Retrieval melalui Query Router dan Pengambilan Informasi. Kueri yang dinyatakan aman diproses melalui empat lapisan klasifikasi secara berurutan: (a) Layer 0a mendeteksi upaya injeksi prompt atau manipulasi instruksi sistem; (b) Layer 0b memeriksa relevansi domain, kueri yang didominasi topik non-psikologis tanpa kata kunci kesehatan mental langsung diarahkan ke jalur FALLBACK; (c) Layer 1 menggunakan pencocokan kamus kata kunci untuk mengklasifikasikan kueri ke salah satu dari tiga jalur utama, yaitu RELATIONAL untuk kueri tentang data skrining personal pengguna, VECTOR untuk kueri edukatif umum terkait literatur psikologi, dan HYBRID untuk kueri yang memerlukan kombinasi keduanya; (d) Layer 2 menggunakan LLM Classifier sebagai mekanisme cadangan apabila Layer 1 menghasilkan klasifikasi ambigu. Pada tahap ini juga diterapkan mekanisme soft query rewriting untuk kueri campuran guna mencegah dilusi semantik pada model embedding. Berdasarkan keputusan tersebut, modul retrieval mengambil informasi dari sumber yang sesuai: jalur RELATIONAL mengakses PostgreSQL melalui query SQL deterministik [23], jalur VECTOR melakukan pencarian pada Qdrant menggunakan dense vector 1024-dimensi yang dikombinasikan dengan sparse vector BM25, jalur HYBRID menjalankan kedua proses dengan tambahan metadata filtering berdasarkan jenis risiko kesehatan mental pengguna yang aktif, sedangkan jalur FALLBACK melewati tahap ini sepenuhnya.

(4) Penggabungan dan Pemeringkatan Hasil (Merge & Ranking). Hasil retrieval dari jalur dense dan sparse digabungkan menggunakan Distribution-Based Score Fusion (DBSF) secara natif di sisi vector database. Rescoring dan reranking selanjutnya dilakukan berdasarkan skor kemiripan kosinus untuk menghasilkan urutan chunk yang paling relevan.

(5) Seleksi Konteks (Top-K Context). Dari seluruh hasil reranking, hanya N chunk berkualitas tinggi dengan $N \leq 5$ yang berada di atas ambang batas skor yang diteruskan ke tahap generasi. Pembatasan ini diterapkan untuk mencegah dilusi konteks yang dapat menurunkan kualitas jawaban LLM.

(6) Sintesis Jawaban (LLM Generator). Konteks terpilih digabungkan dengan riwayat percakapan enam pesan terakhir dan kueri pengguna saat ini ke dalam prompt template terstruktur menggunakan strategi Relevance Sandwich. LLM kemudian menghasilkan respons dalam format yang dibatasi oleh tag XML khusus, yaitu tag <clinical_state> untuk analisis klinis internal dan tag <educational_response> untuk konten psikoedukasi yang ditujukan kepada pengguna.

(7) *Evaluasi Kualitas dan Pengiriman Jawaban.* Sebelum respons dikirimkan, sistem mengeksekusi modul Quality Gate. Untuk jalur RELATIONAL, diterapkan Exact Match deterministik antara data angka dalam tag <clinical_state> dengan rekaman asli di database. Untuk jalur VECTOR dan HYBRID, dihitung skor kemiripan semantik menggunakan model embedding multibahasa berbasis Language-Agnostic BERT Sentence Embedding (LaBSE) apabila dokumen referensi berbahasa Indonesia, atau E5 Cosine Similarity apabila berbahasa Inggris, guna menangkap kemiripan semantik kontekstual tingkat token secara presisi pada korpus multibahasa. Apabila skor berada di bawah ambang batas atau respons mengandung indikasi penolakan konteks, sistem menerapkan kebijakan mitigasi halusinasi model (fail-closed policy) yang melakukan penyaringan ketat (semantic quality gate) pada luaran generator sebelum disajikan kepada pengguna akhir, sehingga respons LLM dibuang dan digantikan dengan pesan kegagalan sistem standar. Tag XML kemudian dihapus, dan respons akhir beserta metadata evaluasi dikirimkan ke pengguna melalui antarmuka Telegram.

2.4. Implementasi Sistem

Sistem diimplementasikan menggunakan Python, FastAPI, Qdrant, dan PostgreSQL tanpa framework orkestrasi pihak ketiga, sehingga setiap komponen pipeline dapat dikontrol secara penuh. Setiap dokumen klinis diindeks dengan dua representasi vektor dalam satu koleksi Qdrant: dense vector menggunakan intfloat/multilingual-e5-large untuk pemahaman semantik lintas bahasa [13], dan sparse vector berbasis BM25 untuk pencocokan leksikal eksak [25]. Penggabungan hasil retrieval menggunakan Distribution-Based Score Fusion (DBSF) secara natif di sisi vector database [31].

Prompt ke model bahasa disusun menggunakan strategi Relevance Sandwich [26], yaitu menempatkan chunk paling relevan di awal dan akhir konteks untuk mengatasi fenomena lost-in-the-middle [27]. Antarmuka chatbot dibangun di atas Telegram Bot API melalui mekanisme Webhook [29] dengan enkripsi protokol MTProto [30]. Seluruh rekayasa pada lapis klasifikasi Query Router difokuskan untuk mencegah false trigger akibat ambiguitas token, sejalan dengan prinsip adaptive routing [28].

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

JUPTIK

2026 Vol. 4 No.1 Hal: 150-159 e-ISSN: 2987-7016

2.5. Pengujian Sistem

Pengujian dilakukan melalui evaluasi kualitas semantik respons menggunakan kerangka tripartit: LaBSE BERTScore, E5 Cosine Similarity, dan RAGAS LLM-as-a-Judge [14].

Pengujian menggunakan 20 pertanyaan uji representatif disusun secara manual untuk memastikan variasi topik dari kelima dimensi psikologis sistem: burnout, kecemasan, perundungan kerja, distres psikologis, dan kesejahteraan umum. Setiap pertanyaan dilengkapi dengan jawaban referensi (*golden answer*) yang dikurasi secara manual oleh peneliti berdasarkan dokumen sumber dalam basis pengetahuan klinis, mencakup kutipan konsep, ambang batas kategori, dan terminologi psikometri yang tepat. Proses ini memastikan bahwa evaluasi BERTScore mengukur kemiripan semantik terhadap referensi yang dapat diverifikasi kebenarannya, bukan terhadap output sistem lain, sehingga menghindari bias sirkular yang umum terjadi pada evaluasi tanpa *gold dataset*. Metode evaluasi dipilih secara adaptif berdasarkan bahasa referensi chunk yang diambil dari Qdrant. Untuk referensi berbahasa Indonesia, digunakan LaBSE BERTScore [32], sebuah model dual-encoder yang dilatih dengan translation ranking objective pada 109 bahasa sehingga menghasilkan representasi semantik lintas bahasa yang stabil. Untuk referensi berbahasa Inggris, digunakan E5 Cosine Similarity (intfloat/multilingual-e5-large) yang menghasilkan skor kemiripan ternormalisasi pada rentang [0, 1] [13].

Pemisahan metode ini diperlukan karena kondisi lintas bahasa yang inherent: knowledge base sebagian besar berbahasa Inggris, sementara jawaban chatbot selalu dalam bahasa Indonesia. Penerapan BERTScore token-level secara langsung akan menghasilkan skor mendekati nol karena greedy matching tidak dapat menyelaraskan token "incompetent" dengan "tidak kompeten" [13]. E5 mengatasi ini melalui contrastive learning yang memetakan kalimat bermakna sama dari dua bahasa ke koordinat vektor berdekatan [13]. Namun E5 rentan terhadap false positive akibat information dilution. Kalimat bergaya klinis yang meyakinkan dapat memperoleh skor tinggi meski faktanya salah, karena seluruh kalimat direpresentasikan sebagai satu titik vektor. Oleh karena itu, RAGAS LLM-as-a-Judge diterapkan secara paralel untuk memverifikasi setiap klaim faktual terhadap dokumen yang diambil [14].

RAGAS menghasilkan lima dimensi evaluasi: Faithfulness, Answer Relevance, Context Precision, Answer Correctness, dan Context Recall [14]. Ambang batas kelulusan ditetapkan F1 ≥ 0,60 untuk LaBSE, similarity ≥ 0,60 untuk E5, dan 1,0 untuk Exact Match jalur RELATIONAL. Model pembangkit jawaban (GPT-4o) dan evaluator RAGAS (GPT-4o-mini) sengaja dipisahkan untuk mengurangi self-evaluation bias sekaligus mengoptimalkan efisiensi biaya.

LaBSE BERTScore dihitung pada hidden states lapisan ke-9, menghasilkan Precision, Recall, dan F1-Score [12]:

$$P_{BERT} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \text{sim}(\hat{x}_j, x_i) \quad (1)$$

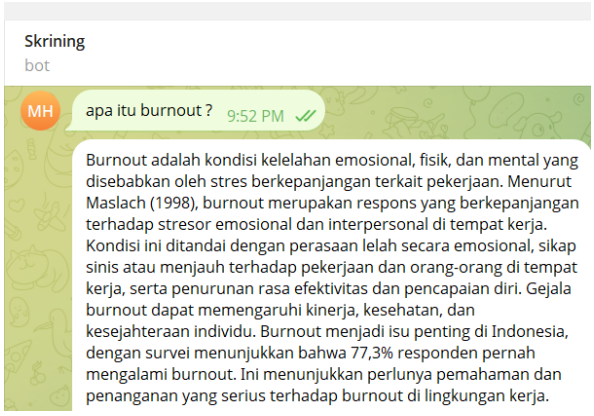
$$R_{BERT} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \text{sim}(x_i, \hat{x}_j) \quad (2)$$

$$F_{BERT} = 2 \cdot \frac{P_{BERT} \times R_{BERT}}{P_{BERT} + R_{BERT}} \quad (3)$$

dengan $\hat{x}$adalah token jawaban sistem, $x$adalah token referensi, dan sim(·)adalah cosine similarity antara embedding kontekstual yang dinormalisasi. E5 Cosine Similarity dihitung menggunakan intfloat/multilingual-e5-large dengan prefiks "query:" pada kedua teks, menghasilkan dot-product ternormalisasi L2 dalam rentang [0, 1] [13].

3. Hasil dan Pembahasan

Sistem chatbot Hybrid RAG berhasil dikembangkan dengan antarmuka percakapan alami pada Telegram. Modul Profiling mengadministrasikan lima instrumen psikometri secara sekuensial dan menyimpan skor terstruktur di PostgreSQL. Ketika pengguna menanyakan data pribadinya, sistem mengambil skor melalui jalur deterministik PostgreSQL tanpa proses embedding, sehingga menghindari risiko halusinasi pada interpretasi nilai numerik klinis. Tampilan antarmuka ditunjukkan pada Gambar 3.



Gambar 3. Tampilan Antarmuka Utama Sistem pada Telegram

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal: 150-159 e-ISSN: 2987-7016

3.1. Hasil Pengujian pada Query Router

Pengujian Query Router dilakukan pada sistem final menggunakan 100 kueri uji yang mencakup empat kelas rute: RELATIONAL, VECTOR, HYBRID, dan FALLBACK. Pengujian menggunakan dua mekanisme routing utama yang menjadi penentu klasifikasi akhir: Layer 1 berbasis pencocokan kamus kata kunci dan Layer 2 berbasis LLM Classifier sebagai cadangan, di mana Layer 0a dan Layer 0b yang menangani filter keamanan dan relevansi domain tidak diuji secara terpisah karena berfungsi sebagai pra-kondisi sebelum proses klasifikasi rute dimulai. Hasil pengujian ditampilkan pada Tabel 2.

Tabel 2. Hasil Pengujian pada Query Router

Jalur	Prec	Rec	F1
Relational	1.000	1.000	1.000
Vector	1.000	1.000	1.000
Hybrid	1.000	1.000	1.000
Fallback	1.000	1.000	1.000
Rata-rata	1.000	1.000	1.000

Query Router berhasil mengklasifikasikan seluruh 100 kueri uji secara benar (akurasi 100%), dengan Precision, Recall, dan F1 sempurna 1.000 pada keempat kelas rute. Pengujian pada 100 kueri uji yang mencakup keempat kelas rute dengan distribusi seimbang (25 kueri per jalur) memperkuat keyakinan bahwa akurasi sempurna ini tidak bersifat artefak dari ukuran sampel kecil, melainkan mencerminkan robustitas mekanisme routing berlapis pada variasi pola bahasa yang lebih luas. Hasil ini mengonfirmasi bahwa pemisahan jalur deterministik dan semantik bekerja secara tepat: kueri tentang data psikometri personal tidak pernah bocor ke jalur semantik, dan sebaliknya.

Akurasi sempurna ini secara langsung memvalidasi ukuran pertama dari tujuan penelitian, yaitu ketepatan klasifikasi rute, dan membuktikan bahwa mekanisme routing berlapis merupakan prasyarat yang dapat diandalkan untuk menjamin akurasi numerik klinis sekaligus relevansi literatur secara bersamaan. Tanpa ketepatan routing di tahap ini, baik halusinasi numerik pada jalur RELATIONAL maupun kontaminasi konteks pada jalur VECTOR tidak dapat dihindari secara konsisten.

3.2. Hasil Pengujian Ablasi Strategi *Retrieval*

Untuk memverifikasi bahwa strategi Hybrid RAG dengan DBSF yang digunakan dalam sistem produksi merupakan pilihan yang tepat, dilakukan studi ablasi tiga arah terhadap strategi retrieval. Tiga konfigurasi yang dibandingkan adalah Dense Only yang hanya mengandalkan vektor dense tanpa komponen sparse, *Hybrid* RAG dengan DBSF sebagai konfigurasi produksi yang menggabungkan dense dan sparse melalui normalisasi berbasis distribusi [31], serta Hybrid RAG dengan RRF sebagai alternatif metode penggabungan berbasis peringkat [25]. Pengujian dilakukan pada 20 pertanyaan uji yang sama dan dievaluasi menggunakan LaBSE BERTScore [12] dan E5 Cosine Similarity [13]. Hasil perbandingan rata-rata ketiga konfigurasi ditampilkan pada Tabel 3.

Tabel 3. Hasil Pengujian BERTScore dan E5 Cosine Similarity

Metode	Prec	Rec	F1	E5
Dense Only	0,6282	0,7381	0,6781	0,9198
Hybrid RAG (DBSF)	0,6463	0,7745	0,7039	0,9234
Hybrid RAG (RRF)	0,6233	0,7355	0,6740	0,9192

Berdasarkan Tabel 3, *Hybrid* RAG dengan DBSF mencatat rata-rata F1-Score tertinggi sebesar 0,7039, diikuti Dense Only sebesar 0,6781, dan *Hybrid* RAG dengan RRF sebesar 0,6740. Selisih F1 antara DBSF dan Dense Only adalah 0,0258, sedangkan selisih antara DBSF dan RRF lebih besar yaitu 0,0299. Dari sisi E5 Cosine Similarity, ketiga metode menghasilkan skor yang sangat berdekatan di mana Dense Only di angka 0,9198, Hybrid RAG dengan RRF 0,9192, Hybrid RAG dengan DBSF yang tertinggi di angka 0,9234.

Jika ditelusuri lebih dalam, perbedaan antar metode menjadi lebih nyata pada kasus-kasus yang membutuhkan kecocokan terminologi klinis spesifik. Pada kondisi tersebut, Dense Only cenderung menghasilkan F1 yang lebih rendah karena retrieval berbasis vektor dense saja kurang sensitif terhadap kata kunci eksak. Hal ini memiliki implikasi klinis langsung: terminologi psikometri yang presisi seperti nama instrumen, ambang batas kategori risiko (misalnya "skor GAD-7 lebih dari atau sama dengan 10 menunjukkan kecemasan sedang"), dan kode klasifikasi diagnostik tidak selalu tertangkap jalur semantik karena representasi vektor dense meratakan perbedaan leksikal tersebut ke dalam koordinat embedding yang berdekatan. Hybrid DBSF mampu menutup celah ini karena komponen BM25 dalam jalur sparse secara langsung menangkap kecocokan leksikal yang tidak tertangkap oleh jalur semantik [25], sehingga dokumen yang memuat terminologi klinis eksak memperoleh bobot lebih tinggi dalam proses fusion [31]. Temuan ini sejalan dengan hasil Mala et al. [9] yang menunjukkan bahwa hybrid retrieval secara konsisten mengungguli sparse maupun dense secara terpisah dalam menekan laju halusinasi pada sistem berbasis LLM, meskipun Mala et al. [9] menggunakan mekanisme RRF pada korpus berbahasa Inggris domain umum. Sebaliknya, pada pertanyaan yang bersifat konseptual dan tidak bergantung pada terminologi tertentu, ketiga metode menghasilkan skor yang hampir identik karena perbedaan antara jalur sparse dan dense menjadi tidak signifikan.

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal: 150-159 e-ISSN: 2987-7016

Oleh karena itu, konfigurasi Hybrid DBSF dipertahankan sebagai pipeline produksi. Hasil studi ablasi ini secara langsung membuktikan ukuran kedua dari tujuan penelitian: strategi retrieval hybrid unggul dibandingkan kedua alternatifnya dalam mempertahankan relevansi semantik literatur klinis, khususnya pada kueri yang membutuhkan presisi terminologi psikometri.

3.3. Hasil Pengujian RAGAS

Untuk memverifikasi akurasi faktual dan kualitas retrieval sistem Hybrid RAG secara komprehensif, dilakukan evaluasi menggunakan kerangka RAGAS dengan model GPT-4o-mini sebagai hakim otomatis. Evaluasi mencakup lima dimensi: Faithfulness yang mengukur sejauh mana klaim dalam jawaban didukung dokumen yang diambil, Answer Relevance yang mengukur relevansi jawaban terhadap pertanyaan, Context Precision yang mengukur proporsi konteks yang benar-benar relevan, Answer Correctness yang membandingkan jawaban dengan referensi emas, dan Context Recall yang mengukur kelengkapan dokumen yang berhasil diambil. Hasil evaluasi terhadap 20 kueri uji ditampilkan pada Tabel 4.

Tabel 4. Hasil Pengujian RAGAS

No	Faith	Ans. Rel	Ctx Prec	Ans. Corr	Ctx. Rec
1	1,0000	0,8871	0,8056	0,7597	0,6667
2	1,0000	0,8732	0,7500	0,9890	1,0000
3	0,8462	0,8813	0,5000	0,4406	1,0000
4	1,0000	0,9161	0,8333	0,9888	1,0000
5	0,6250	0,9033	1,0000	0,6423	0,5000
6	1,0000	0,8707	1,0000	0,9913	1,0000
7	0,5714	0,8326	0,6389	0,3392	0,5000
8	1,0000	0,9852	1,0000	0,5228	0,5000
9	1,0000	0,8802	0,8333	0,7390	0,6667
10	1,0000	0,8931	1,0000	0,7676	1,0000
11	0,8571	0,9742	1,0000	0,9667	1,0000
12	1,0000	0,8714	0,8042	0,7247	1,0000
13	1,0000	0,9144	1,0000	0,9850	1,0000
14	0,8000	0,8814	0,9500	0,7331	1,0000
15	1,0000	0,9103	1,0000	0,9792	1,0000
16	1,0000	0,8009	1,0000	0,6441	1,0000
17	1,0000	0,8574	1,0000	0,6552	1,0000
18	1,0000	0,9541	1,0000	0,6336	1,0000
19	1,0000	0,8257	1,0000	0,6293	1,0000
20	1,0000	0,9447	1,0000	0,7326	1,0000
Rata-rata	0,9350	0,8929	0,9058	0,7447	0,8917

Hasil evaluasi RAGAS menunjukkan performa yang baik secara keseluruhan dengan metrik Faithfulness mencapai rata-rata 0,9350, artinya lebih dari 93% klaim dalam jawaban chatbot didukung dokumen yang ditemukan dan sistem tidak menunjukkan tanda-tanda halusinasi berat. Context Recall rata-rata 0,8917 menunjukkan sistem retrieval berhasil menemukan dokumen yang relevan pada sebagian besar kueri. Answer Relevance rata-rata 0,8929 menunjukkan chatbot sebagian besar menjawab langsung ke inti pertanyaan. Context Precision rata-rata 0,9058 mengonfirmasi bahwa sebagian besar potongan teks yang diambil memang relevan dengan pertanyaan.

Dimensi Answer Correctness (0,7447) mencerminkan tantangan inheren evaluasi otomatis pada sistem lintas bahasa: chatbot menjawab dalam bahasa Indonesia sementara referensi emas sebagian besar berbahasa Inggris, sehingga RAGAS menilai kemiripan lebih rendah meskipun substansinya benar. Ini bukan semata kelemahan sistem, melainkan batas metrik yang juga dilaporkan pada evaluasi RAG lintas bahasa lainnya [10]. Indikator yang lebih representatif untuk menilai keandalan sistem ini adalah Faithfulness (0,9350), di mana lebih dari 93% klaim yang disampaikan chatbot dapat dilacak ke dokumen yang diambil, artinya sistem tidak mengarang informasi.

3.4. Pembahasan

Pada evaluasi kualitas jawaban, seluruh 20 kueri uji berhasil melampaui ambang batas kelulusan F1 $\geq$ 0,60 dengan rata-rata F1-Score BERTScore tertinggi diraih oleh Hybrid RAG (DBSF) sebesar 0,7039, diikuti Dense Only 0,6781 dan Hybrid RAG (RRF) 0,6740, serta E5 Cosine Similarity tertinggi juga diraih Hybrid RAG (DBSF) sebesar 0,9234. Perbedaan antara kedua skor ini dapat dijelaskan dari cara kerja masing-masing metrik. BERTScore mencocokkan setiap kata dalam jawaban chatbot berbahasa Indonesia dengan setiap kata dalam dokumen referensi berbahasa Inggris satu per satu, sehingga perbedaan kosakata seperti "perundungan" dengan "bullying" secara sistematis menekan nilai Precision dan Recall meskipun maknanya identik [12]. E5 tidak mengalami masalah ini karena menilai kemiripan dua teks secara keseluruhan, sehingga teks yang membahas topik sama akan menghasilkan skor tinggi terlepas dari perbedaan bahasa [13]. Nilai F1 rata-rata 0,7039 pada Hybrid RAG (DBSF) justru menunjukkan kemampuan model LaBSE dalam menjembatani perbedaan bahasa tersebut, selaras dengan karakteristik dual-encoder yang dilatih pada 109 bahasa dengan translation ranking objective [32]. Meski demikian, skor E5 yang tinggi tidak otomatis menjamin jawaban bebas halusinasi karena seluruh kalimat diringkas menjadi satu titik vektor tanpa verifikasi per klaim, sehingga jawaban yang terdengar meyakinkan tetap bisa memperoleh skor tinggi meski faktanya keliru [13].

Evaluasi RAGAS sebagai lapisan verifikasi faktual menunjukkan Faithfulness rata-rata 0,9350, yang berarti lebih dari 93% klaim dalam jawaban chatbot didukung oleh dokumen yang ditemukan dan sistem tidak menunjukkan tanda-tanda halusinasi berat [14]. Answer Relevance rata-rata 0,8929 menunjukkan chatbot sebagian besar menjawab langsung ke inti

pertanyaan. Context Recall rata-rata 0,8917 menunjukkan retrieval berhasil menemukan dokumen yang relevan pada sebagian besar kueri, dan Context Precision rata-rata 0,9058 mengonfirmasi bahwa sebagian besar potongan teks yang diambil memang relevan dengan pertanyaan.

Nilai Answer Correctness yang moderat (0,7447) pada beberapa kueri disebabkan oleh dua kondisi yang saling terkait. Pertama, chatbot terkadang menyampaikan substansi yang benar namun menggunakan terminologi atau susunan kalimat yang berbeda dari referensi emas, sehingga RAGAS menilai kemiripannya lebih rendah meski secara isi tidak keliru. Kedua, pada beberapa kueri terjadi ketidakcocokan sumber antara dokumen yang berhasil diambil retriever dengan sumber yang menjadi dasar referensi emas, sehingga chatbot menjawab berdasarkan konteks yang tersedia namun tidak identik dengan yang diharapkan. Kondisi ini sekaligus menjelaskan Faithfulness yang lebih rendah pada kueri tertentu, di mana sebagian klaim dalam jawaban tidak dapat ditelusuri ke dokumen yang diambil. Temuan ini memperkuat rekomendasi untuk memperluas knowledge base klinis berbahasa Indonesia agar cakupan sumber retrieval lebih selaras dengan referensi emas yang ditetapkan.

Secara keseluruhan, ketiga hasil pengujian saling mengonfirmasi satu sama lain: akurasi routing sempurna, keunggulan Hybrid DBSF dalam presisi terminologi klinis, dan Faithfulness di atas 93%. Ketiga ukuran yang ditetapkan dalam tujuan penelitian seluruhnya terpenuhi: ketepatan klasifikasi rute 100% membuktikan pemisahan jalur bekerja secara andal, keunggulan Hybrid DBSF mengonfirmasi relevansi semantik pada retrieval klinis multibahasa, dan Faithfulness rata-rata 0,9350 mengonfirmasi bahwa secara agregat lebih dari 90% klaim jawaban chatbot dapat ditelusuri ke dokumen klinis terverifikasi. Hasil ini menunjukkan sistem tidak menunjukkan tanda-tanda halusinasi berat secara keseluruhan, namun belum dapat diklaim bebas halusinasi secara konsisten pada seluruh skenario.

Arsitektur ini berhasil membuktikan bahwa pemisahan jalur deterministik dan semantik bukan hanya pilihan desain, melainkan kebutuhan fungsional. Tanpa pemisahan tersebut, baik halusinasi numerik maupun degradasi relevansi semantik tidak dapat dihindari secara bersamaan. Hasil ini selaras dengan temuan Gumma et al. [10] yang menunjukkan bahwa akurasi faktual sistem RAG pada domain klinis multibahasa secara umum masih rendah, khususnya pada kueri non-Inggris, sekaligus memperluas temuan tersebut dengan menunjukkan bahwa pemisahan jalur deterministik dan semantik secara eksplisit dapat menjadi salah satu mekanisme untuk mengatasi tantangan tersebut pada platform percakapan berbahasa Indonesia.

Suharyadi dan Saputra [34] mengembangkan sistem RAG hybrid untuk konsultasi hukum berbahasa Indonesia dengan menggabungkan BM25, FAISS, dan keyword boosting melalui kombinasi linear berbobot tetap (α=0,4; β=0,4; γ=0,2) yang disetel secara empiris menggunakan pencarian grid pada 50 kueri validasi. Evaluasi dilakukan secara manual oleh dua pakar hukum menggunakan skala Likert, dan konfigurasi ensemble terbaik memperoleh rata-rata skor 4,5 dari 5. Sistem pada penelitian ini menggunakan DBSF yang menormalisasi skor sparse dan dense berdasarkan distribusi aktual tiap komponen secara natif di sisi vector database [31], tanpa hyperparameter bobot yang harus disetel secara trial-and-error untuk setiap domain baru. Perbedaan arsitektural yang lebih mendasar adalah keberadaan Query Router empat lapis dan jalur relasional deterministik pada penelitian ini, yang memisahkan kueri data psikometri pengguna dari kueri literatur klinis secara eksplisit. Suharyadi dan Saputra [34] tidak mengimplementasikan pemisahan jalur semacam itu karena sistem konsultasi hukum mereka menangani seluruh kueri melalui retrieval semantik. Evaluasi pada penelitian ini menggunakan kerangka otomatis tripartit yang menghasilkan metrik kuantitatif reprodusibel tanpa bergantung pada ketersediaan pakar domain untuk setiap siklus pengujian.

Mala et al. [9] mengevaluasi hybrid retrieval menggunakan Reciprocal Rank Fusion (RRF) dengan query expansion berbasis WordNet pada dataset HaluBench yang seluruhnya berbahasa Inggris, mencakup domain umum seperti keuangan, biomedis, dan pengetahuan umum. Konfigurasi hybrid terbaik mereka mencapai accuracy 89,55% dan hallucination rate 9,36% pada subset anotasi. Studi ablasi tiga arah pada penelitian ini menunjukkan bahwa DBSF menghasilkan LaBSE F1 lebih tinggi (0,7039) dibandingkan RRF (0,6740) pada korpus klinis multibahasa. Perbedaan ini konsisten dengan perbedaan mekanisme kedua metode: DBSF mempertimbangkan besaran skor aktual dari masing-masing komponen retrieval dalam proses normalisasi distribusi [31], sementara RRF hanya menggunakan peringkat relatif tanpa mempertimbangkan selisih skor. Perbedaan kedua terletak pada cakupan arsitektur: Mala et al. [9] tidak mengimplementasikan pemisahan jalur deterministik untuk data terstruktur, sehingga seluruh respons bergantung pada retrieval semantik. Pada penelitian ini, data numerik psikometri pengguna diambil langsung dari PostgreSQL melalui kueri SQL deterministik [23], yang secara struktural mencegah risiko interpretasi angka klinis yang salah akibat keterbatasan retrieval berbasis vektor.

4. Kesimpulan

Penelitian ini berhasil membuktikan bahwa arsitektur Hybrid RAG dengan Query Router berlapis mampu memenuhi dua tuntutan yang selama ini sulit dipenuhi secara bersamaan oleh sistem RAG konvensional, yaitu akurasi deterministik pada interpretasi data psikometri numerik dan relevansi semantik pada pengambilan literatur klinis. Tiga hasil pengujian saling mengonfirmasi tujuan tersebut: (1) Query Router mencapai akurasi klasifikasi 100% pada 100 kueri uji; (2) Hybrid RAG dengan DBSF meraih LaBSE F1 tertinggi sebesar 0,7039 dan E5 Cosine Similarity tertinggi sebesar 0,9234 dalam studi ablasi tiga arah; dan (3) Faithfulness rata-rata 0,9350 mengonfirmasi bahwa lebih dari 90% klaim jawaban chatbot dapat ditelusuri ke dokumen klinis yang diambil. Nilai Answer Correctness yang moderat sebesar 0,7447 mencerminkan kendala metrik evaluasi otomatis pada kondisi lintas bahasa, bukan kegagalan sistem, karena chatbot menjawab dalam bahasa Indonesia sementara referensi emas sebagian besar tersedia dalam bahasa Inggris. Kontribusi utama penelitian ini adalah bukti empiris bahwa pemisahan jalur deterministik dan semantik yang dikendalikan mekanisme routing berlapis dengan kebijakan Fail-Closed merupakan solusi yang layak dan efektif untuk chatbot kesehatan pada domain klinis multibahasa.

Penelitian ini memiliki sejumlah keterbatasan yang perlu diakui. (1) Evaluator RAGAS bergantung pada LLM sebagai hakim otomatis yang membawa risiko bias inheren pada penilaian faktual. (2) Mekanisme Quality Gate berjalan secara sinkron sehingga menambah latensi respons dan berpotensi memblokir layanan apabila evaluator mengalami kegagalan. (3) Sistem rentan terhadap fenomena retrieval dilution pada kueri campuran yang mengandung banyak konteks personal, karena kebisingan informasi mengaburkan koordinat vektor sehingga dokumen literatur klinis utama tidak selalu berhasil diambil. (4) Penelitian ini tidak menyertakan User Acceptance Testing dengan partisipan nyata dari populasi sasaran, sehingga aspek pengalaman pengguna seperti kejelasan bahasa respons, kemudahan alur skrining, dan kepercayaan terhadap sistem belum dapat dinilai secara empiris.

Berdasarkan keterbatasan tersebut, beberapa arah pengembangan diusulkan. (1) Perluasan knowledge base klinis berbahasa Indonesia perlu diprioritaskan agar cakupan sumber retrieval lebih selaras dengan konteks pengguna lokal sekaligus berpotensi meningkatkan nilai Answer Correctness. (2) Penerapan mekanisme query reframing atau intent extraction otomatis sebelum kueri dikirim ke database vektor dapat mengurangi dampak retrieval dilution pada kueri campuran. (3) Penambahan bibliography filter dan optimasi threshold re-ranking direkomendasikan guna meningkatkan Context Precision. (4) Implementasi fallback evaluator atau evaluasi asinkron perlu dipertimbangkan untuk menjaga ketersediaan layanan tanpa mengorbankan jaminan kualitas informasi. (5) Pengujian lanjutan dengan melibatkan variasi gold dataset berbahasa Indonesia yang divalidasi oleh panel pakar psikologi klinis secara langsung akan meningkatkan validitas eksternal metrik Answer Correctness. (6) Perluasan knowledge base dengan mengintegrasikan literatur psikologi Islam juga layak dipertimbangkan sebagai langkah pengembangan lanjutan. Referensi yang membahas kesehatan mental dalam perspektif Islam, seperti kajian tentang nafs, sabar, tawakal, dan kesejahteraan spiritual, dapat diindeks sebagai korpus tersendiri dengan label kategori khusus di Qdrant. Langkah ini relevan mengingat sebagian besar pengguna sistem di Indonesia berada dalam konteks budaya dan nilai keislaman, sehingga respons yang berlandaskan perspektif tersebut berpotensi lebih diterima dan bermakna bagi pengguna secara personal.

Daftar Rujukan

[1] WHO, “Mental health.” Accessed: Mar. 04, 2026. [Online]. Available: https://www.who.int/health-topics/mental-health

[2] WHO, “COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide.” Accessed: May 03, 2026. [Online]. Available: https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

[3] WHO, “Over a billion people living with mental health conditions – services require urgent scale-up.” Accessed: Apr. 29, 2026. [Online]. Available: https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up

[4] Deloitte Global, “2025 Gen Z and Millennial Survey: Growth and the pursuit of money, meaning, and well-being” New York, 2025.

[5] S. Winurini, “Penanganan Kesehatan Mental di Indonesia,” Jakarta, 2023.

[6] Kementerian Kesehatan RI, “Laporan Kinerja Pembinaan Pelayanan Kesehatan Jiwa dan Kelompok Disabilitas serta Korban KTP/A Tahun 2025,” Jakarta, 2025.

[7] Kementerian Kesehatan RI, “Skrining Kesehatan Jiwa Gratis Lewat SATUSEHAT Mobile.” Accessed: Mar. 01, 2026. [Online]. Available: https://kemkes.go.id/id/skrining-kesehatan-jiwa-gratis-lewat-satusehat-mobile

[8] I. Hungerbuehler, K. Daley, and K. Cavanagh, “Chatbot-Based Assessment of Employees’Mental Health: Design Process and Pilot Implementation,” *JMIR Form Res*, vol. 5, no. 4, p. e21678, 2021, https://doi.org/10.2196/21678

[9] C. S. Mala, G. Gezici, and F. Giannotti, “Hybrid Retrieval for Hallucination Mitigation in Large Language Models : A Comparative Analysis,” *ArXiv*, Feb. 2025, doi: 10.48550/arXiv.2504.05324.

[10] V. Gumma, A. Raghunath, M. Jain, and S. Sitaram, “HEALTH-PARIKSHA: Assessing RAG Models for Health Chatbots in Real-World Multilingual Settings,” *ArXiv*, Oct. 2025, doi: 10.48550/arXiv.2410.13671.

[11] R. S. Pressman and B. R. Maxim, *Software Engineering: A Practitioner's Approach*, 9th ed. New York: McGraw-Hill Education, 2020.

[12] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation With Bert," in *International Conference on Learning Representations (ICLR)*, Feb. 2020, pp. 1–43. Accessed: Apr. 03, 2026. [Online]. Available: https://arxiv.org/abs/1904.09675

[13] L. Wang *et al.*, "Text Embeddings by Weakly-Supervised Contrastive Pre-training," *ArXiv*, Feb. 2024, doi: 10.48550/arXiv.2212.03533.

[14] S. Es, J. James, L. Espinosa-Anke, S. Schockaert, and E. Gradients, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Mar. 2024, pp. 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16

[15] C. W. Topp, S. Dinesen, and S. Søndergaard, "The WHO-5 Well-Being Index : A Systematic Review of the Literature," *Psychother Psychosom*, vol. 84, no. 3, pp. 167–176, 2015, https://doi.org/10.1159/000376585

[16] M. Siradjuddin, D. A. Perwitasari, L. M. Irham, H. Dania, and T. Herlina, "Validity and reliability of the world health organisation-five well being index (WHO-5) questionnaire in early detection of depression during Covid-19 pandemic in Yogyakarta," *Pharmaciana*, vol. 13, no. 2, pp. 204–210, Feb. 2023, https://doi.org/10.12928/pharmaciana.v13i2.24319

[17] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Lo, "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7," *Arch Intern Med*, vol. 166, no. 10, pp. 1092–1097, 2006, https://doi.org/10.1001/archinte.166.10.1092

[18] A. Budikayanti, A. Larasari, K. Malik, Z. Syeban, L. A. Indrawati, and F. Octaviana, "Screening of Generalized Anxiety Disorder in Patients with Epilepsy: Using a Valid and Reliable Indonesian Version of Generalized Anxiety Disorder-7 (GAD-7)," *Neurology Research International*, vol. 2019, p. 10, Jun. 2019, https://doi.org/10.1155/2019/5902610

[19] L. T. Widhianingtanti and G. van Luijtelaar, "The Maslach-Trisni Burnout Inventory : Adaptation for Indonesia," *JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)*, vol. 11, no. 1, pp. 1–21, 2022, https://doi.org/10.15408/jp3i.v11i1.24400

[20] T. Duc, F. Kaligis, T. Wiguna, and L. Willenberg, "Screening for depressive and anxiety disorders among adolescents in Indonesia: Formal validation of the centre for epidemiologic studies depression scale– revised and the Kessler psychological distress scale," *J Affect Disord*, vol. 246, pp. 189–194, 2019, https://doi.org/10.1016/j.jad.2018.12.042

[21] S. Einarsen, H. Hoel, and G. Notelaers, "Measuring exposure to bullying and harassment at work : Validity , factor structure and psychometric properties of the Negative Acts Questionnaire-Revised," vol. 23, no. 1, pp. 22–24, May 2009, doi: 10.1080/02678370902815673.

[22] D. Erwandi and A. Kadir, "Identification of Workplace Bullying : Reliability and Validity of Indonesian Version of the Negative Acts," *Environmental Research and Public Health*, vol. 18, no. 8, Apr. 2021, https://doi.org/10.3390/ijerph18083985

[23] E. F. Codd, "A Relational Model of Data Large Shared Data Banks," vol. 13, no. 6, Jun. 1970, doi: 10.1145/362384.362685.

[24] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, Vancouver: Curran Associates, Inc., 2020, pp. 9459–9474. doi: 10.48550/arXiv.2005.11401.

[25] Y. Luan, J. Eisenstein, K. Toutanova, and M. Collins, "Sparse, Dense, and Attentional Representations for Text Retrieval," *Trans Assoc Comput Linguist*, vol. 9, pp. 329–345, Apr. 2021, https://doi.org/10.1162/tacl_a_00369

[26] S. Schulhoff *et al.*, "The Prompt Report: A Systematic Survey of Prompt Engineering Techniques," *ArXiv*, Feb. 2025, doi: 10.48550/arXiv.2406.06608.

[27] N. F. Liu *et al.*, "Lost in the Middle: How Language Models Use Long Contexts," *Trans Assoc Comput Linguist*, vol. 12, pp. 157–173, 2024, https://doi.org/10.1162/tacl_a_00638

[28] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park, "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity," *ArXiv*, Mar. 2024, https://doi.org/10.18653/v1/2024.naacl-long.389

[29] Telegram, "Telegram Bot API," https://core.telegram.org/bots/api. Accessed: May 10, 2026. [Online]. Available: https://core.telegram.org/bots/api

[30] T. Sušánka and J. Kokes, "Security Analysis of the Telegram IM," no. 8, pp. 1–8, Nov. 2017, https://doi.org/10.1145/3150376.3150382

[31] Qdrant, "Hybrid Search with Distribution-Based Score Fusion." Accessed: Jun. 09, 2026. [Online]. Available: https://qdrant.tech/articles/hybrid-search/

[32] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT Sentence Embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin: Association for Computational Linguistics, May 2022, pp. 878–891. https://doi.org/10.18653/v1/2022.acl-long.62

[33] M. Casu, S. Triscari, S. Battiato, L. Guarnera, and P. Caponnetto, "AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility, and Applications," *MDPI*, vol. 14, no. 13, p. 5889, Jul. 2024, https://doi.org/10.3390/app14135889

[34] Suharyadi and I. Saputra, "Hybrid Ensemble Retrieval-Augmented Generation for Indonesian Legal Consultation with Keyword Boosting," *Journal of Novel Engineering Science and Technology*, vol. 4, no. 2, pp. 71–85, Aug. 2025, https://doi.org/10.56741/jnest.v4i02.1042