ANALISIS PEMANFAATAN LARGE LANGUAGE MODEL DALAM PENANGANAN KASUS KESEHATAN MENTAL

Achmad Muafi Taufiqurrochman, Arda Surya Editya, Ahmad Khoir Alhaq
Informatika, Universitas Nahdlatul Ulama Sidoarjo
Jl. Lingkar Timur KM 5,5, Sidoarjo, Indonesia
23422015.mhs@unusida.ac.id

ABSTRAK

Permasalahan kesehatan mental di Indonesia masih terhambat oleh terbatasnya akses layanan dan tenaga profesional. Chatbot berbasis Large Language Model (LLM) berpotensi menjadi alternatif dukungan psikologis awal. Namun, efektivitas chatbot sangat bergantung pada relevansi dan akurasi respons terhadap kebutuhan pengguna. Penelitian ini bertujuan mengevaluasi kinerja lima model LLM (GPT-4o, GPT-5, Claude Haiku 4.5, Gemini 3 Pro, dan Deepseek V3.2) dalam menghasilkan respons chatbot penanganan kesehatan mental. Metode yang digunakan menerapkan pendekatan Retrieval-Augmented Generation (RAG) dengan menambahkan konteks dari knowledge base eksternal. Evaluasi dilakukan terhadap 20 pertanyaan pengujian. Respons model kemudian dibandingkan dengan ground truth menggunakan metrik BLEU-4 dan ROUGE. Hasil penelitian menunjukkan penerapan RAG secara signifikan mampu meningkatkan kualitas respons chatbot. Di antara seluruh model, Deepseek v3.2 memperoleh performa terbaik dan tertinggi pada seluruh metrik evaluasi. Integrasi RAG dengan LLM yang tepat terbukti menjadi pendekatan menjanjikan dalam mengembangkan chatbot pendukung kesehatan mental.

Kata kunci : *Chatbot, Kesehatan Mental, LLM, RAG, Evaluasi*

1. PENDAHULUAN

Kesehatan mental merupakan salah satu isu kesehatan yang semakin mendapat perhatian di Indonesia. Berdasarkan hasil Asia Care Survey 2024 yang dilakukan oleh Manulife, masyarakat Indonesia tidak hanya menunjukkan kekhawatiran terhadap penyakit fisik, tetapi juga terhadap berbagai gangguan kesehatan mental [1]. Survei yang dilakukan pada Januari 2024 terhadap lebih dari 1.000 responden tersebut menunjukkan bahwa stres dan burnout menjadi gangguan yang paling dikhawatirkan dengan persentase sebesar 56%, diikuti oleh gangguan tidur (42,6%), kecemasan (28,2%), kesepian (24,9%), depresi (20,7%), dan gangguan kognitif (9,1%). Tingginya urgensi ini juga tercermin pada kelompok usia remaja, di mana Survei Nasional Kesehatan Jiwa Remaja 2022 mencatat sekitar 5,5% remaja Indonesia atau setara dengan 2,45 juta jiwa mengalami gangguan mental dalam satu tahun terakhir [2]. Temuan tersebut menegaskan bahwa permasalahan ini sangat luas dan memerlukan penanganan serius agar tidak berdampak pada penurunan kualitas hidup, produktivitas, hingga kondisi kesehatan fisik individu.

Meskipun urgensinya semakin tinggi, penanganan kesehatan mental di Indonesia masih menghadapi kendala utama berupa keterbatasan akses layanan dan rendahnya ketersediaan tenaga profesional. Rasio psikiater di Indonesia tercatat sekitar 0,6 per 100.000 penduduk, yang masih berada di bawah rata-rata negara-negara ASEAN [3]. Kondisi ini menyebabkan dukungan psikologis awal belum dapat dijangkau secara merata. Sebagai alternatif, teknologi kecerdasan buatan melalui *chatbot* berbasis *Large Language Model* (LLM) membuka peluang baru untuk menyediakan pendampingan awal yang cepat, natural, dan adaptif. Namun, pemanfaatan LLM memunculkan permasalahan tersendiri; respons yang dihasilkan sering kali mengalami ketidaksesuaian konteks, kurang relevan terhadap pertanyaan pengguna, atau belum cukup akurat untuk mendukung kebutuhan komunikasi pada domain kesehatan mental [4]. Pada konteks yang sensitif ini, ketidakakuratan respons sangat berisiko menimbulkan kesalahpahaman dan menurunkan efektivitas sistem.

Oleh karena itu, penelitian ini bertujuan untuk menawarkan evaluasi komparatif terhadap kinerja beberapa model LLM—yaitu GPT-4o, GPT-5, Claude Haiku 4.5, Gemini 3 Pro, dan Deepseek V3.2—dalam menghasilkan respons *chatbot* untuk penanganan kesehatan mental. Sebagai rencana penyelesaian masalah dalam mengatasi kelemahan konteks model, penelitian ini memanfaatkan model *embedding* untuk memastikan informasi yang diberikan selaras dengan kebutuhan pengguna. Untuk menilai kualitas respons secara objektif, evaluasi dilakukan menggunakan metrik *Natural Language Processing* berupa BLEU-4, ROUGE-1, ROUGE-2, dan ROUGE-L. Metrik tersebut digunakan untuk mengukur tingkat kesamaan antara respons yang dihasilkan model dengan *ground truth* atau referensi jawaban yang telah ditentukan [4]. Hasil penelitian ini diharapkan dapat menjadi rujukan dalam pengembangan *chatbot* kesehatan mental yang relevan, akurat, dan sesuai konteks untuk mendukung masyarakat.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terkait

Beberapa penelitian sebelumnya telah mengkaji penerapan *Large Language Model* (LLM) dan teknologi *chatbot* dalam ranah kesehatan mental, baik

dari segi implementasi teknis, efektivitas intervensi, maupun batasan aplikasinya. Saffiera dkk. [5] mengembangkan *chatbot* empatik bernama "SeudaraBunda" yang mengintegrasikan LLM dengan arsitektur *Retrieval-Augmented Generation* (RAG) untuk skrining awal depresi pasca melahirkan di Aceh. Pengujian sistem yang diadaptasi dengan budaya lokal ini mencetak skor *System Usability Scale* (SUS) rata-rata 85,63, membuktikan bahwa pendekatan RAG efektif menghasilkan respons percakapan yang adaptif, aman, dan relevan.

Kajian literatur sistematis oleh Aima [6] mengevaluasi efektivitas *chatbot* AI dalam intervensi kesehatan mental digital. Melalui analisis terhadap delapan studi klinis, ditemukan bahwa interaksi *chatbot* secara signifikan efektif menurunkan gejala depresi dan kecemasan. Namun, *chatbot* masih memiliki keterbatasan dalam hal tingkat empati dan risiko respons yang di luar konteks, sehingga pengembangannya memerlukan panduan regulasi etika yang ketat.

Sementara itu, Yang dkk. [7] meninjau batasan dan risiko global penerapan LLM dalam kesehatan mental. Penelitian ini menyoroti adanya tren pergeseran menuju spesialisasi model domain khusus guna mengejar akurasi klinis, sekaligus memperingatkan risiko fatal dari "halusinasi" model dan kurangnya adaptasi budaya pada data non-Barat. Oleh karena itu, penerapan LLM menuntut integrasi *multimodal* dan spesialisasi psikologis yang lebih mendalam.

2.2. Chatbot Kesehatan Mental

Chatbot kesehatan mental merupakan sistem berbasis kecerdasan buatan yang dirancang untuk memberikan dukungan psikologis melalui interaksi percakapan. Teknologi ini memanfaatkan AI berbasis LLM untuk memahami *input* pengguna dan menghasilkan respons yang relevan, sehingga dapat digunakan sebagai alat bantu dalam identifikasi awal gangguan kesehatan mental serta pemberian dukungan emosional [8]. Seiring meningkatnya kebutuhan layanan kesehatan mental yang mudah diakses, *chatbot* menjadi solusi alternatif yang mampu menyediakan layanan secara fleksibel tanpa batasan waktu dan tempat.

Penggunaan *chatbot* berbasis AI menunjukkan potensi yang signifikan dalam bidang kesehatan mental, terutama dalam membantu deteksi dini melalui analisis percakapan pengguna menggunakan teknik *Natural Language Processing* (NLP) dan *sentiment analysis*. Selain itu, berbagai penelitian menunjukkan bahwa *chatbot* seperti Woebot, Wysa, dan Tess mampu membantu mengurangi gejala depresi dan kecemasan melalui pendekatan terapi berbasis percakapan. Secara umum, potensi dan tantangan *chatbot* dalam kesehatan mental sebagai berikut:

a. Meningkatkan akses layanan kesehatan mental secara luas dan *real-time*,
b. Mampu memberikan dukungan awal serta rekomendasi penanganan
c. Menghadapi keterbatasan dalam memahami konteks emosional, isu privasi data, serta perbedaan budaya dan literasi digital.

2.3. Large Language Model (LLM)

Large Language Model (LLM) merupakan model berbasis kecerdasan buatan yang digunakan dalam berbagai aplikasi *Natural Language Processing* (NLP) karena kemampuannya dalam mempelajari pola bahasa dari data tekstual dalam jumlah besar [9]. Perkembangan LLM didukung oleh arsitektur Transformer, yang memungkinkan model memahami konteks dan menghasilkan teks yang mendekati bahasa manusia [10]. LLM bekerja dengan memprediksi *token* berikutnya berdasarkan konteks sebelumnya, sehingga mampu menghasilkan respons yang koheren dan relevan dalam berbagai tugas seperti analisis sentimen, penerjemahan bahasa, dan sistem tanya jawab [9].

2.4. Embedding

Embedding merupakan teknik dalam *Natural Language Processing* (NLP) yang digunakan untuk merepresentasikan teks dalam bentuk vektor numerik sehingga dapat diproses oleh model pembelajaran mesin [11]. Berbeda dengan representasi tradisional seperti *bag of words*, *embedding* mampu menangkap makna semantik dan hubungan antar kata dalam suatu konteks. Dengan pendekatan ini, kata atau kalimat yang memiliki makna serupa akan memiliki representasi vektor yang berdekatan dalam ruang multidimensi.

2.5. Cosine Similarity

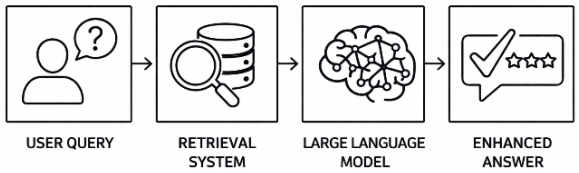
Cosine similarity merupakan metode yang digunakan untuk mengukur tingkat kesamaan antara dua vektor dalam ruang multidimensi, yang dalam konteks *Natural Language Processing* (NLP) biasanya digunakan untuk membandingkan representasi teks hasil *embedding* [12]. Metode ini bekerja dengan menghitung nilai *cosinus* dari sudut antara dua vektor, sehingga tidak dipengaruhi oleh panjang vektor melainkan oleh arah vektor tersebut. Semakin kecil sudut antara dua vektor, maka semakin tinggi tingkat kesamaannya. Secara matematis, *cosine similarity* dirumuskan sebagai berikut:

$$cosine\ similarity = \frac{A \cdot B}{\vee A \vee\vee B \vee} \qquad (1)$$

2.6. Retrieval Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) merupakan pendekatan dalam *Natural Language Processing* (NLP) yang menggabungkan proses pengambilan informasi (*retrieval*) dengan pembangkitan teks (*generation*) menggunakan LLM [13]. Pendekatan ini bertujuan untuk meningkatkan kualitas respons dengan memanfaatkan sumber pengetahuan eksternal, sehingga model tidak hanya

bergantung pada pengetahuan yang diperoleh selama proses pelatihan.



Gambar 1. Arsitektur RAG

Berdasarkan Gambar 1, proses RAG dimulai dari *user query* yang diberikan oleh pengguna, kemudian diproses oleh *retrieval system* untuk mencari informasi yang relevan dari basis pengetahuan. Proses ini memanfaatkan teknik *embedding* untuk merepresentasikan teks dalam bentuk vektor serta metode kesamaan seperti *cosine similarity* untuk menentukan informasi yang paling relevan. Informasi hasil *retrieval* selanjutnya digunakan sebagai konteks tambahan bagi LLM dalam menghasilkan respons.

2.7. Metode Evaluasi RAG

Evaluasi dalam NLP merupakan proses untuk mengukur kualitas teks yang dihasilkan oleh model dengan membandingkannya terhadap teks referensi. Dalam penelitian ini, evaluasi dilakukan untuk menilai kualitas respons chatbot yang dihasilkan oleh Large Language Model. Metode yang digunakan adalah BLEU dan ROUGE, yang merupakan metrik evaluasi berbasis kesamaan teks (*text similarity*) dan banyak digunakan dalam tugas generasi teks seperti *machine translation*, *text summarization*, dan sistem dialog [14]. Penggunaan kedua metode ini bertujuan untuk mengukur tingkat kesesuaian antara respon yang dihasilkan model dengan jawaban referensi yang telah ditentukan.

2.8. BLEU

BLEU (*Bilingual Evaluation Understudy*) merupakan metode evaluasi yang digunakan untuk mengukur tingkat kesamaan antara teks hasil generasi model dengan teks referensi berdasarkan kesamaan n-gram. BLEU berfokus pada *precision*, yaitu seberapa banyak kata atau frasa dalam teks hasil model yang sesuai dengan teks referensi[15]. Dalam penelitian ini digunakan BLEU-4, yang mempertimbangkan hingga 4-gram untuk menangkap kesesuaian konteks yang lebih panjang. BLEU dirumuskan sebagai berikut:

$$BLEU = BP \times exp(\sum w_n log p_n) \qquad (2)$$

2.9. ROUGE

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) merupakan metode evaluasi yang digunakan untuk mengukur tingkat kesamaan antara teks hasil model dengan teks referensi berdasarkan jumlah kesamaan unit teks (*overlap*) [13]. Oleh karena itu, ROUGE banyak digunakan dalam evaluasi tugas generasi teks seperti t*ext summarization* dan sistem dialog.

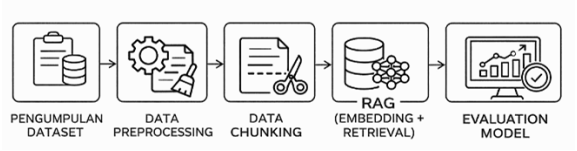
Dalam perhitungannya, ROUGE menghasilkan tiga metrik utama yaitu *precision*, *recall*, dan F1-score.

$$F1 - score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \qquad (3)$$

Dalam penelitian ini digunakan ROUGE-1, ROUGE-2, dan ROUGE-L. ROUGE-1 mengukur kesamaan berdasarkan *unigram*, ROUGE-2 berdasarkan *bigram* sehingga lebih sensitif terhadap urutan kata, dan ROUGE-L menggunakan *Longest Common Subsequence* (LCS) untuk menilai kesamaan urutan kata. Nilai yang digunakan sebagai acuan utama adalah F1-score karena memberikan keseimbangan antara *precision* dan *recall* dalam mengevaluasi kualitas respons model.

3. METODE PENELITIAN

Penelitian ini dirancang menggunakan pendekatan eksperimen untuk menganalisis dan mengevaluasi performa beberapa *Large Language Model* (LLM) dalam menghasilkan respons *chatbot* pada konteks kesehatan mental.



Gambar 2. Metodologi Penelitan

Alur penelitian secara keseluruhan ditunjukkan pada Gambar 2, yang mencakup tahapan mulai dari pengumpulan *dataset*, *preprocessing* data, proses *chunking*, penerapan metode *Retrieval-Augmented Generation* (RAG), hingga evaluasi kualitas respons model menggunakan metrik BLEU dan ROUGE.

3.1. Pengumpulan Dataset

Dataset yang digunakan dalam penelitian ini berupa *knowledge base* yang berisi kumpulan dokumen mengenai berbagai topik kesehatan mental, seperti kecemasan, depresi, bipolar, dan skizofrenia. Dokumen-dokumen tersebut diperoleh dari berbagai sumber terpercaya, termasuk organisasi kesehatan internasional dan institusi resmi, sehingga kualitas serta kredibilitas informasinya dapat dipertanggungjawabkan. Dalam penelitian ini, *dataset* berfungsi sebagai sumber pengetahuan utama pada sistem *Retrieval-Augmented Generation* (RAG), di mana setiap dokumen diproses untuk menyediakan konteks yang relevan pada tahap *retrieval* guna mendukung pembangkitan respons oleh model.

Tabel 1. Dataset knowlegde base

Kategori	Nama Dokumen	Sumber
Kecemasan	Anxiety Disorders	https://www.psychology.org.au/for-the-public/Psychology-topics/Anxiety
	Anxiety: symptoms, treatment and causes.	https://www.healthdirect.gov.au/anxiety
Depresi	What is depression? American Psychiatric Association.	https://www.psychiatry.org/patients-families/depression/what-is-depression
	Beyond Blue. (n.d.). Psychological treatments for depression.	https://www.beyondblue.org.au/the-facts/depression/treatments-for-depression/psychological-treatments-for-depression
Kesehatan Mental	World Health Organization. Mental Health: a state of well-being	http://www.who.int/features/factfiles/mental_health/en/
Self Harm	Kondisi Kesehatan Mental Self-harm atau Menyakiti Diri	https://www.seributujuan.id/id/self-harm#apa-itu-self-harm-menyakiti-diri
Skizofrenia	Schizophrenia Symptoms, signs and coping tips.	https://www.helpguide.org/articles/mental-disorders/schizophrenia-signs-and-symptoms.html

Tabel 1 menyajikan rincian *dataset knowledge base* yang digunakan, meliputi kategori topik kesehatan mental, nama dokumen, serta sumber dokumen yang telah diterjemahkan. Keberagaman kategori tersebut dirancang untuk memberikan cakupan informasi yang lebih luas sehingga sistem mampu menghasilkan respons yang lebih relevan, informatif, dan sesuai dengan konteks pertanyaan pengguna.

3.2. Data Preprocessing

Tahap data *preprocessing* dilakukan untuk menyiapkan dokumen pada *knowledge base* agar memiliki format yang seragam, bersih, dan siap digunakan dalam proses *retrieval* pada sistem *Retrieval-Augmented Generation* (RAG). Tahapan ini penting karena kualitas data masukan sangat memengaruhi kemampuan sistem dalam menemukan informasi yang relevan dan menghasilkan respons yang akurat. Dokumen yang diperoleh dari berbagai sumber umumnya memiliki struktur penulisan, format, dan karakteristik teks yang berbeda, sehingga perlu dilakukan penyesuaian sebelum diproses lebih lanjut.

Proses *preprocessing* diawali dengan pembersihan data (data *cleaning*), yaitu menghapus karakter-karakter yang tidak diperlukan, seperti simbol tertentu, spasi berlebih, tanda baca yang tidak relevan, serta elemen teks lain yang berpotensi mengganggu proses analisis. Tahap ini bertujuan untuk mengurangi *noise* pada dokumen sehingga isi teks menjadi lebih fokus pada informasi utama yang berkaitan dengan topik kesehatan mental.

Setelah proses pembersihan, dilakukan normalisasi teks untuk menyeragamkan bentuk penulisan data. Normalisasi ini mencakup penyesuaian huruf besar dan huruf kecil, perapian struktur kalimat, serta penyamaan format penulisan agar seluruh dokumen berada dalam bentuk yang konsisten. Langkah ini diperlukan agar perbedaan gaya penulisan antar dokumen tidak memengaruhi proses pencocokan informasi pada tahap *retrieval*.

Tahap berikutnya adalah penyesuaian format data agar dokumen menjadi lebih terstruktur dan mudah diproses oleh sistem. Pada penelitian ini, setiap dokumen disusun dalam format teks yang seragam sehingga dapat dipecah menjadi bagian-bagian informasi yang lebih kecil pada tahap *chunking*. Dengan struktur data yang lebih rapi, sistem dapat mengelola isi dokumen secara lebih efisien dan meningkatkan ketepatan dalam mengambil konteks yang sesuai dengan pertanyaan pengguna.

Melalui tahapan data *preprocessing* ini, *dataset* yang semula berasal dari berbagai sumber heterogen diubah menjadi data teks yang lebih bersih, konsisten, dan terorganisasi. Hasil dari proses ini menjadi dasar penting bagi tahapan berikutnya, yaitu *chunking*, pembentukan *embedding*, dan proses *retrieval*, sehingga keseluruhan sistem *chatbot* dapat bekerja secara lebih optimal dalam menghasilkan respons yang relevan dan kontekstual.

3.3. Data Chunking

Setelah melalui tahap *preprocessing*, dokumen pada *knowledge base* selanjutnya diproses pada tahap data *chunking*. Tahap ini dilakukan dengan membagi dokumen yang telah dibersihkan dan dinormalisasi menjadi bagian-bagian teks yang lebih kecil (*chunk*) agar informasi di dalam dokumen dapat dikelola secara lebih terstruktur. Proses ini penting dalam sistem *Retrieval-Augmented Generation* (RAG) karena dokumen yang terlalu panjang dapat menyulitkan proses pencarian konteks yang relevan dan berpotensi menurunkan efisiensi sistem.

Pembagian dokumen menjadi *chunk* bertujuan agar setiap potongan teks memuat informasi yang lebih terfokus dan mudah diproses pada tahap *retrieval*. Dengan ukuran teks yang lebih kecil dan terkontrol, sistem dapat melakukan pencarian informasi secara lebih presisi terhadap pertanyaan pengguna. Selain itu, *chunking* juga membantu mengurangi risiko pengambilan konteks yang terlalu luas atau kurang spesifik, sehingga respons yang dihasilkan model menjadi lebih relevan terhadap kebutuhan percakapan.

Pada penelitian ini, proses *chunking* dilakukan dengan membagi setiap dokumen berdasarkan panjang teks atau jumlah *token* tertentu. Setiap *chunk* disusun sedemikian rupa agar tetap mempertahankan

keterkaitan makna dari isi dokumen asli, sehingga informasi penting tidak terputus secara ekstrem. Pendekatan ini memungkinkan setiap bagian teks tetap merepresentasikan konteks topik kesehatan mental yang dibahas, seperti definisi, gejala, faktor risiko, maupun penanganan awal dari masing-masing kondisi.

Tabel 2. Proses Chunking Data

Nama Dokumen	Data Mentah	Setelah Chunking
Kategori	Kecemasan	Kecemasan
Referensi	Anxiety Disorders full page (1-3)	Anxiety Disorders page 1
Context	Apa itu ansietas atau kecemasan? Kecemasan adalah sebuah reaksi singkat alami terhadap kejadian yang membuat ..... atau saat masih ragu untuk mengambil langkah berikutnya.	Apa itu ansietas atau kecemasan? Kecemasan adalah sebuah reaksi singkat alami terhadap kejadian yang membuat .....mengatasi ketidakpastian kini telah dihubungkan dengan kecemasan.
Metadata	{ "dokumen": "Anxiety Disorders", "ref": { "page": "1-3"} }	{ "dokumen": "Anxiety Disorders", "ref": { "page": "1", "chunk": "1" } }

Tabel 2 menunjukkan contoh hasil proses *chunking* dari dokumen *knowledge base* yang digunakan dalam penelitian ini. Setiap dokumen dipecah menjadi beberapa bagian teks dengan ukuran yang lebih terstruktur, sehingga menghasilkan unit-unit informasi yang lebih spesifik dan mudah diproses. Hasil *chunking* tersebut selanjutnya digunakan pada tahap pembentukan *embedding* dan disimpan ke dalam basis data vektor. Penyimpanan dalam bentuk vektor ini bertujuan untuk mendukung proses pencarian semantik, sehingga sistem dapat menemukan *chunk* yang paling relevan terhadap pertanyaan pengguna berdasarkan kedekatan makna, bukan hanya kesamaan kata secara *literal*.

Dengan demikian, tahap data *chunking* berperan penting dalam meningkatkan efisiensi dan efektivitas sistem RAG. Melalui pembagian dokumen menjadi unit informasi yang lebih kecil, sistem dapat mengakses konteks yang lebih tepat, mempercepat proses pencarian, dan mendukung pembangkitan respons *chatbot* yang lebih akurat, relevan, dan sesuai dengan konteks pertanyaan pada domain kesehatan mental.

3.4. RAG

Pada tahap ini, metode *Retrieval-Augmented Generation* (RAG) diimplementasikan untuk meningkatkan kualitas respons *chatbot* dengan menambahkan informasi kontekstual yang diambil dari *knowledge base*. Pendekatan ini digunakan karena model *Large Language Model* (LLM) yang berdiri sendiri belum tentu selalu mampu menghasilkan jawaban yang relevan dan akurat, khususnya pada domain yang sensitif seperti kesehatan mental. Melalui RAG, sistem tidak hanya mengandalkan pengetahuan internal model, tetapi juga memanfaatkan informasi eksternal yang telah disiapkan pada basis pengetahuan sebagai referensi tambahan dalam proses pembangkitan respons.

Implementasi RAG pada penelitian ini melibatkan tiga proses utama, yaitu *embedding*, *retrieval*, dan *generation*. Pada tahap *embedding*, setiap *chunk* hasil pemrosesan dokumen diubah ke dalam bentuk representasi vektor agar dapat diproses secara semantik oleh sistem. Representasi ini memungkinkan teks dengan makna yang serupa ditempatkan pada ruang vektor yang berdekatan, sehingga pencarian informasi tidak hanya bergantung pada kecocokan kata, tetapi juga pada kedekatan makna.

Selanjutnya, pada tahap *retrieval*, sistem melakukan pencarian terhadap *chunk* yang paling relevan dengan pertanyaan pengguna. Untuk mengukur tingkat kedekatan antara vektor pertanyaan dan vektor dokumen, penelitian ini menggunakan metode *cosine similarity*. Metode ini dipilih karena mampu mengukur kesamaan arah antar vektor secara efektif, sehingga cocok digunakan dalam pencarian semantik pada basis data vektor. Berdasarkan hasil perhitungan tersebut, sistem akan mengambil sejumlah *chunk* dengan nilai kesamaan tertinggi (top-k) untuk dijadikan konteks tambahan.

Tahap berikutnya adalah *generation*, yaitu proses pembangkitan respons oleh model LLM dengan memanfaatkan pertanyaan pengguna beserta konteks hasil *retrieval*. Dengan adanya informasi tambahan dari *knowledge base*, model diharapkan mampu menghasilkan jawaban yang lebih relevan, informatif, dan sesuai dengan konteks pertanyaan. Pendekatan ini juga membantu mengurangi kemungkinan model menghasilkan respons yang terlalu umum atau tidak sesuai dengan domain pembahasan.

Tabel 3. Parameter RAG

Parameter	Nilai
Model Embedding	text-embedding-3-large
Metode Similarity	Cosine Similarity
Jumlah Retrieval (Top-k)	3
Model LLM	GPT-4o, GPT-5, Deepseek 3.2, Gemini 3 Pro, Claude Haiku 4.5
Sumber Data	Knowledge Base (Database Vector)
Output	Response dari LLM

Tabel 3 menunjukkan parameter yang digunakan dalam penerapan metode RAG pada penelitian ini. Model *embedding* text-embedding-3-large digunakan

untuk mengubah teks menjadi representasi vektor, sedangkan metode *cosine similarity* diterapkan untuk menghitung tingkat kesamaan antar vektor dalam proses *retrieval*. Sistem kemudian mengambil sejumlah *chunk* dari basis data vektor berdasarkan nilai kemiripan tertinggi (top-k) untuk digunakan sebagai konteks tambahan yang diberikan kepada model LLM. Penentuan parameter-parameter tersebut bertujuan untuk memastikan sistem RAG dapat bekerja secara optimal dalam menemukan informasi yang relevan dari *knowledge base* dan mendukung pembangkitan respons *chatbot* yang lebih kontekstual.

3.5. Evaluation Model

Tahap evaluasi dilakukan untuk menilai kualitas respons yang dihasilkan oleh setiap *Large Language Model* (LLM) dalam sistem *chatbot* kesehatan mental. Evaluasi ini bertujuan untuk mengetahui sejauh mana model mampu menghasilkan jawaban yang relevan, akurat, dan sesuai dengan konteks pertanyaan yang diberikan. Dalam penelitian ini, proses evaluasi dilakukan dengan membandingkan respons yang dihasilkan model terhadap jawaban referensi (*ground truth*) yang telah disusun sebelumnya. Pendekatan ini digunakan untuk memberikan ukuran yang lebih objektif terhadap performa masing-masing model dalam menjawab pertanyaan pada domain kesehatan mental.

Proses evaluasi dilakukan menggunakan sejumlah pertanyaan pengujian yang mewakili berbagai topik dalam *knowledge base*. Setiap pertanyaan diberikan kepada model, kemudian respons yang dihasilkan dibandingkan dengan jawaban referensi untuk melihat tingkat kesamaan isi dan struktur informasi. Melalui pendekatan ini, evaluasi tidak hanya menilai apakah model dapat memberikan jawaban, tetapi juga menilai kualitas jawaban tersebut dalam menyampaikan informasi yang sesuai dengan kebutuhan pengguna.

Metode evaluasi yang digunakan dalam penelitian ini meliputi BLEU-4, ROUGE-1, ROUGE-2, dan ROUGE-L. BLEU-4 digunakan untuk mengukur tingkat kesamaan respons model dengan jawaban referensi berdasarkan *precision*, yaitu seberapa banyak kata atau frasa pada respons model yang sesuai dengan referensi. Sementara itu, ROUGE digunakan untuk mengukur kualitas respons berdasarkan tingkat kemiripan isi dari sudut pandang *recall* maupun F1-score, sehingga dapat menunjukkan sejauh mana informasi penting dalam jawaban referensi berhasil tercakup dalam respons model.

Penggunaan kombinasi BLEU dan ROUGE memberikan evaluasi yang lebih komprehensif. BLEU berfokus pada ketepatan pemilihan kata atau frasa terhadap jawaban referensi, sedangkan ROUGE menekankan pada kelengkapan informasi yang berhasil ditangkap oleh model. Dengan demikian, kedua metrik ini saling melengkapi dalam memberikan gambaran menyeluruh mengenai kualitas respons yang dihasilkan oleh masing-masing model.

Hasil evaluasi tiap model dibandingkan untuk menentukan model dengan performa terbaik dalam menghasilkan respons *chatbot* kesehatan mental. Model dengan nilai metrik tertinggi dianggap paling mampu memberikan jawaban yang relevan dan mendekati *ground truth*, sehingga menjadi dasar dalam menilai efektivitas LLM pada *chatbot* berbasis RAG.

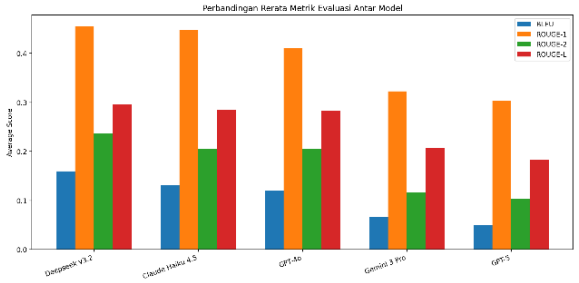
4. HASIL DAN PEMBAHASAN

Penelitian ini dilakukan dengan menyiapkan *dataset* pengujian yang terdiri dari 20 pertanyaan terkait kesehatan mental yang telah dilengkapi dengan *ground truth* sebagai jawaban referensi. Setiap data uji diberikan kode identitas (ID Test) untuk mempermudah proses analisis. Tabel 4 menunjukkan *dataset* pengujian yang digunakan, yang berisi pertanyaan serta jawaban referensi.

Tabel 4. Dataset Pengujian

ID Test	Pertanyaan	Ground Truth
T1	Apa yang dimaksud dengan masalah kesehatan mental?	Masalah kesehatan mental adalah serangkaian kondisi yang berdampak pada kesehatan mental, mengganggu suasana hati, perilaku, pemikiran, atau cara berinteraksi, serta ditentukan dari dampaknya terhadap fungsi harian (contoh: depresi, kecemasan, bipolar, skizofrenia).
T2	Bagaimana WHO mendefinisikan kesehatan secara umum?	WHO mendefinisikan kesehatan sebagai keadaan fisik, mental, dan sosial yang lengkap, bukan hanya sekadar ketiadaan penyakit atau kelemahan.
….	……	……
T20	Mengapa seseorang memilih untuk menyakiti dirinya sendiri (*self-harm*)?	Seseorang menyakiti diri (*self-harm*) sering kali digunakan sebagai respons untuk mengatasi, melepaskan, mengendalikan, atau mengungkapkan pikiran intens dan rasa sakit emosional yang terlalu sulit untuk diucapkan.

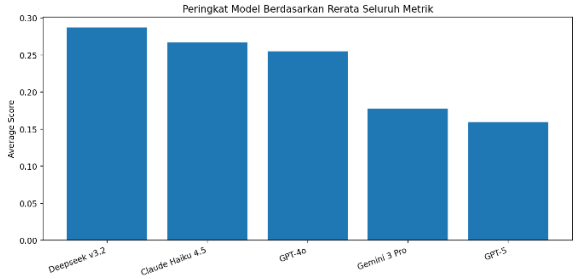
Pemberian kode ID Test bertujuan untuk mempermudah perbandingan hasil evaluasi. Setelah *dataset* disiapkan, setiap pertanyaan diproses menggunakan metode RAG, yaitu dengan melakukan *retrieval* dari *knowledge base* dan generasi respons oleh LLM. Hasil respons kemudian dievaluasi menggunakan BLEU-4 dan ROUGE (F1-score) untuk mengukur kesesuaian dengan *ground truth*. Hasil pengujian ditampilkan pada Gambar 3.



Gambar 3. Grafik Perbandingan Rerata Metrik Model

Gambar 3 menunjukkan perbandingan performa lima *Large Language Model* (LLM) berdasarkan nilai rata-rata BLEU, ROUGE-1, ROUGE-2, dan ROUGE-L dari seluruh data uji. Secara umum, Deepseek v3.2 menunjukkan performa paling tinggi dengan rerata BLEU sebesar 0,1595, ROUGE-1 sebesar 0,4556, ROUGE-2 sebesar 0,2372, dan ROUGE-L sebesar 0,2958.

Posisi berikutnya ditempati oleh Claude Haiku 4.5 dengan performa yang juga relatif tinggi pada hampir seluruh metrik, kemudian diikuti oleh GPT-4o. Sementara itu, Gemini 3 Pro dan GPT-5 memperoleh rerata yang lebih rendah dibandingkan ketiga model lainnya. Temuan ini menunjukkan bahwa berdasarkan rata-rata seluruh metrik evaluasi, model Deepseek v3.2 memiliki kemampuan yang paling baik dalam menghasilkan respons yang paling mendekati *ground truth*, baik dari segi ketepatan kata maupun kesesuaian isi jawaban. Dengan demikian, grafik ini memberikan gambaran bahwa perbedaan performa antar model cukup terlihat, terutama pada metrik ROUGE-1 dan ROUGE-L yang menunjukkan kualitas cakupan informasi dan konsistensi respons yang dihasilkan masing-masing model.



Gambar 4. Grafik Peringkat Metrik Antar Model

Grafik peringkat model berdasarkan rerata seluruh metrik menyajikan perbandingan performa lima *Large Language Model* (LLM) secara keseluruhan dengan menggabungkan nilai rata-rata BLEU, ROUGE-1, ROUGE-2, dan ROUGE-L. Grafik ini memberikan gambaran umum mengenai kemampuan masing-masing model dalam menghasilkan respons *chatbot* kesehatan mental yang mendekati *ground truth*. Berdasarkan hasil perhitungan, Deepseek v3.2 menempati peringkat pertama dengan rerata keseluruhan sebesar 0,2870. Posisi berikutnya ditempati oleh Claude Haiku 4.5 dengan nilai 0,2672, kemudian GPT-4o sebesar 0,2550, Gemini 3 Pro sebesar 0,1778, dan GPT-5 sebesar 0,1598. Hasil ini menunjukkan adanya perbedaan performa yang cukup jelas di antara model-model yang diuji.

Deepseek v3.2 menunjukkan keunggulan karena memiliki keseimbangan terbaik dalam ketepatan penggunaan kata dan kelengkapan informasi, sehingga menghasilkan respons yang lebih relevan dan akurat dibandingkan model lain. Claude Haiku 4.5 dan GPT-4o juga memiliki performa yang cukup kompetitif, sedangkan Gemini 3 Pro dan GPT-5 menunjukkan hasil yang lebih rendah dalam pengujian ini.

Secara keseluruhan, evaluasi berbasis BLEU dan ROUGE dapat menjadi indikator komprehensif dalam menentukan model terbaik. Berdasarkan hasil tersebut, Deepseek v3.2 dinyatakan sebagai model dengan performa terbaik, sekaligus menegaskan bahwa pemilihan LLM sangat berpengaruh terhadap kualitas *chatbot* kesehatan mental yang lebih efektif dan andal.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, penerapan metode *Retrieval-Augmented Generation* (RAG) terbukti mampu meningkatkan kualitas respons *chatbot* kesehatan mental melalui pemanfaatan *knowledge base* sebagai sumber konteks tambahan. Evaluasi menggunakan metrik BLEU-4, ROUGE-1, ROUGE-2, dan ROUGE-L menunjukkan adanya perbedaan performa yang cukup signifikan antar model. Berdasarkan rerata nilai evaluasi, Deepseek v3.2 memiliki performa terbaik, diikuti oleh Claude Haiku 4.5 dan GPT-4o. Hal ini menunjukkan bahwa kombinasi metode RAG dan pemilihan model LLM yang tepat berperan penting dalam menghasilkan respons yang lebih relevan, akurat, serta mendekati *ground truth*.

Untuk penelitian selanjutnya, disarankan agar *dataset* diperluas dengan jumlah pertanyaan yang lebih banyak dan cakupan topik kesehatan mental yang lebih beragam agar hasil evaluasi menjadi lebih representatif. Selain itu, penambahan metode evaluasi berbasis penilaian manusia (*human evaluation*) perlu dilakukan untuk mengukur aspek kualitas yang tidak dapat ditangkap sepenuhnya oleh metrik otomatis, seperti empati, kebermanfaatan, dan keamanan respons. Optimalisasi pada proses *retrieval*, termasuk strategi *chunking*, pemilihan nilai top-k, serta penggunaan model *embedding* yang berbeda, juga penting untuk meningkatkan akurasi sistem. Di samping itu, eksplorasi terhadap model LLM lain yang lebih beragam dapat dilakukan guna memperoleh konfigurasi *chatbot* yang lebih optimal dan andal.

DAFTAR PUSTAKA

[1] I. K. M. Putra, K. A. T. Dewi, and G. S. W. Yasa, "Pengaruh lingkungan kerja, job insecurity, beban kerja terhadap stres kerja karyawan Ramayana Suites & Resort," *Ganec Swara*, vol. 19, no. 3, pp. 955–963, Sep. 2025.

[2] F. Rahmawaty et al., "Faktor-faktor yang mempengaruhi kesehatan mental pada remaja: Factors affecting mental health in adolescents," *Jurnal Surya Medika (JSM)*, vol. 8, no. 3, pp. 276–281, Dec. 2022, doi: 10.33084/jsm.v8i3.4522.

[3] S. Idaiani and E. I. Riyadi, "Sistem kesehatan jiwa di Indonesia: Tantangan untuk memenuhi kebutuhan," *Jurnal Penelitian dan Pengembangan Pelayanan Kesehatan*, vol. 2, no. 2, pp. 70–80, Aug. 2018, doi: 10.22435/jpppk.v2i2.134.

[4] L. Judijanto et al., Optimalisasi ChatGPT: Panduan dan Penerapan untuk Belajar, Mengajar, dan Membuat Konten Tanpa Batas. Yogyakarta: PT Green Pustaka Indonesia, 2025.

[5] A. F. N. Husna et al., "Penggunaan intelligence doll untuk mengatasi kesehatan mental pada Gen-Z," *Repository Universitas Airlangga, Surabaya, Jan*. 2026.

[6] C. A. Saffiera, F. I. E. Sari, N. A. Hasma, and R. Akbar, "Retrieval-Augmented LLM-Based Empathetic Chatbot for Early Postpartum Depression Screening in Aceh," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 10, no. 2, pp. 1003–1013, Apr. 2026, doi: 10.33395/sinkron.v10i2.15841.

[7] H. H. Aima, "Literature Review: Pemanfaatan Chatbot Berbasis Kecerdasan Buatan Dalam Mendukung Kesehatan Mental," *Syntax Literate: Jurnal Ilmiah Indonesi*a, vol. 10, no. 12, pp. 12280–12292, Dec. 2025.

[8] J. Yang et al., "Exploring the application boundaries of LLMs in mental health: a systematic scoping review," *Front. Psychol*., vol. 16, Art. no. 1715306, Feb. 2026, doi: 10.3389/fpsyg.2025.1715306.

[9] F. Sulianta, ChatGPT: Memberdayakan Large Language Model untuk Berbagai Kebutuhan. Bandung: Feri Sulianta, 2024.

[10] Y. A. et al., Penerapan Large Language Model (LLM) dalam Chatbot. Ponorogo: Uwais Inspirasi Indonesia, 2025.

[11] A. S. Editya, F. Anam, H. Ismanto, A. Christanti, and L. Muzdalifah, "Chunk Segmentation for Optimizing Long Document Retrieval Augmented Generation on Islamic Scriptures," *in 2025 15th International Conference on Information & Communication Technology and System (ICTS), Surabaya*, 2025, doi: 10.1109/ICTS67612.2025.11369502.N.

[12] N. Rokhman, P. A. Maulan, and N. A. Wirahuda, "Analisis penilaian esai secara otomatis menggunakan natural language processing (NLP) dan cosine similarity," *Go Infotech: Jurnal Ilmiah STMIK AUB*, vol. 31, no. 1, pp. 41–52, Jun. 2025, doi: 10.36309/goi.v31i1.359.

[13] A. S. Wiradinata and V. C. Mawardi, "Abstractive Text Summarization Berita Bahasa Indonesia Menggunakan Retrieval-Augmented Generation," *Jurnal Ilmu Komputer dan Sistem Informasi*, vol. 13, no. 1, 2025, doi: 10.24912/jiksi.v13i1.32861.

[14] E. T. Eman, T. N. Fatyanosa, and A. F. Aji, "Analisis perbandingan metode chunking dalam chatbot berbasis retrieval-augmented generation rekomendasi terapi nutrisi medis pasien," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 1, pp. 1–10, Jan. 2026.

[15] M. Ghassemiazghandi, "An evaluation of ChatGPT's translation accuracy using BLEU score," *Theory and Practice in Language Studies*, vol. 14, no. 4, pp. 985–994, Apr. 2024, doi: 10.17507/TPLS.1404.07.

[16] M. Barbella and G. Tortora, "ROUGE metric evaluation for text summarization techniques," *SSRN, Rochester, NY, May* 26, 2022, doi: 10.2139/ssrn.4120317