Jurnal Sains Informatika Terapan (JSIT)
E-ISSN: 2828-1659, Volume: 5, Issue: 2, Month: Juni, Year: 2026
Page: 609 - 615, Available online at: https://rcf-indonesia.org/home/

ANALISIS PERFORMA MODEL *EMBEDDING BGE SMALL* DAN *MINILM-L6* TERHADAP KUALITAS RETRIEVAL MENGGUNAKAN METRIK RAGAS

Ahmad Ibrahim Maqbul[1)], Anggun Fergina[2)]
[1,2] Program Studi Teknik Informatika, Universitas Nusa Putra
Corresponding Author: [1] ahmad.ibrahim_ti22@nusaputra.ac.id

Article Info

Article history:

Received: Jun 16, 2026
Revised: Jun 18, 2026
Accepted: Jun 24, 2026
Published: Jun 26, 2026

Keywords:

Chatbot;
Chunking;
Model Embedding;
RAGAS;
Retrieval-Augmented Generation;

ABSTRACT

The application of Large Language Models in the medical domain is often hampered by issues of hallucination and limited up-to-date knowledge. Retrieval-Augmented Generation offers a solution for connecting LLM with factual data, but the quality of RAG output is highly dependent on the accuracy of the information retrieval process. This study aims to analyze the effect of chunk size and embedding model variations on retrieval quality in a medical chatbot system at the Nusa Putra Farmedika General Clinic. The method used is a comparative experiment by testing three chunk size variations (256, 512, and 1024 tokens) and comparing the performance of two embedding models, BGE Small and MiniLM-L6. The evaluation was conducted automatically using the RAGAS framework, focusing on the Context Recall and Context Precision metrics. These findings were implemented into a medical chatbot prototype as a form of functional validation. The results showed an inverse relationship between chunk size and retrieval quality, with a chunk size of 512 tokens producing the best level of information granularity. The BGE Small model proved to be slightly superior to MiniLM-L6 in capturing the semantics of clinical text. The most optimal configuration was achieved by combining the BGE Small model with a chunk size of 512, which produced the highest average score of 0.59, Context Recall of 0.45, and Context Precision of 0.74. This study recommends this configuration as a technical standard for the development of medical chatbot as a foundational step to improve context relevance and mitigate the potential for hallucinations.

1. PENDAHULUAN

Perkembangan teknologi kecerdasan buatan telah mempercepat transformasi digital di berbagai sektor, termasuk bidang kesehatan [1]. Di institusi kesehatan seperti Klinik Umum Nusa Putra Farmedika Pratama [2], penggunaan chatbot telah menjadi langkah strategis untuk mengedukasi masyarakat mengenai pengobatan mandiri untuk keluhan ringan serta mengurangi beban administratif tenaga medis [3]. Namun, chatbot konvensional, yang mengandalkan *Natural Language Processing* [4], pendekatan berbasis aturan tradisional, dan pencarian kata kunci, sering kali terbatas dalam memahami konteks percakapan yang luas. Pendekatan ini cenderung menghasilkan respons yang kaku dan tidak relevan serta tidak mampu menjawab pertanyaan medis yang kompleks [5].

Munculnya *Large Language Models* (LLM) menawarkan potensi revolusi dalam pengembangan sistem konsultasi medis berkat kemampuannya memahami konteks dan menghasilkan teks yang alami dan mirip manusia [6]. Meskipun memiliki keunggulan, efektivitas LLM sering terhambat oleh tantangan kritis berupa "*hallucinations*", yaitu kecenderungan model untuk menghasilkan informasi yang tampak meyakinkan namun secara faktual salah [7].

Di bidang kedokteran, informasi yang keliru dari chatbot tidak hanya merusak kepercayaan pengguna, tetapi juga dapat menimbulkan konsekuensi fatal bagi keselamatan pasien. Oleh karena itu, diperlukan mekanisme teknis untuk memitigasi risiko-risiko ini agar hasil keluaran model tetap sesuai dengan Pedoman Praktik Klinis. *Retrieval-Augmented Generation* (RAG) menawarkan solusi arsitektural untuk mengatasi masalah "halusinasi" dengan mengintegrasikan kemampuan generatif LLM dengan basis pengetahuan eksternal yang tervalidasi [8].

Namun, kualitas jawaban yang dihasilkan oleh sistem RAG sangat bergantung pada akurasi proses

pengambilan informasi [9], [10]. Kegagalan pada fase penentuan konteks ini akan secara langsung menyebabkan sistem gagal memberikan saran medis yang tepat. Kualitas pengambilan informasi ini dipengaruhi oleh dua parameter fundamental pada tahap prapemrosesan data: ukuran *chunk* [11], dan model representasi vektor (model *embedding*) [12].

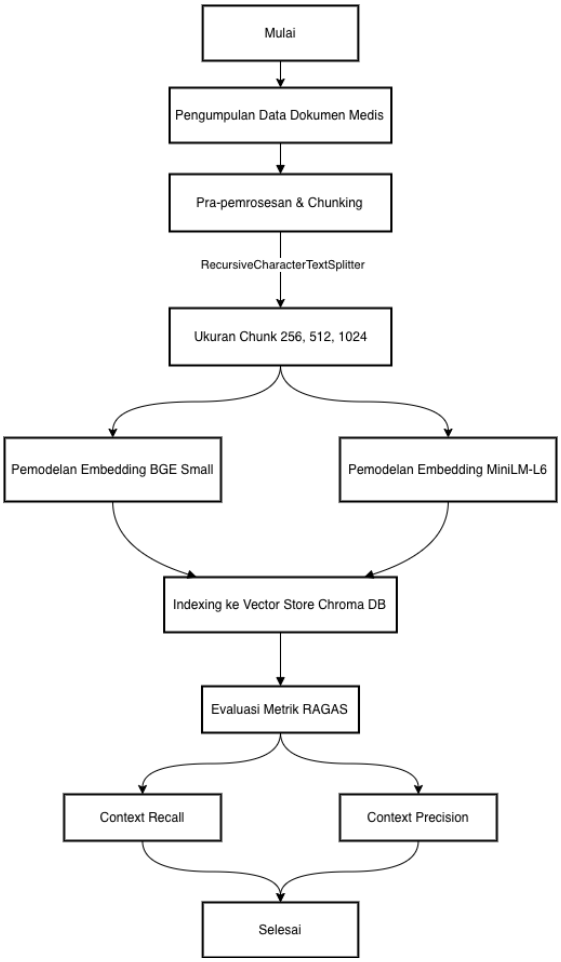
Menentukan ukuran *chunk* yang optimal dalam sistem RAG merupakan tantangan karena hal ini melibatkan keseimbangan antara akurasi pencarian, retensi konteks, dan efisiensi komputasi [13]. *Chunk* yang terlalu kecil berisiko menghilangkan konteks semantik yang penting, sementara chunk yang terlalu besar dapat memasukkan informasi yang tidak relevan (noise) ke dalam hasil pencarian [14]. Selain ukuran teks, tahap pengambilan data juga bergantung pada model *embedding* yang digunakan untuk mengubah teks medis menjadi representasi vektor numerik [15].

Memilih model yang tepat sangat penting untuk memetakan kata-kata dengan makna medis yang serupa ke dalam ruang dimensi vektor secara akurat [16]. Untuk memastikan efektivitas sistem secara objektif, evaluasi kinerja pengambilan data perlu diukur menggunakan kerangka kerja otomatis seperti *RAGAS* [17], tanpa hanya mengandalkan anotasi manual.

Berdasarkan masalah-masalah yang teridentifikasi ini, penelitian ini bertujuan untuk menganalisis pengaruh variasi ukuran *chunk* (256, 512, dan 1024 token) serta membandingkan model embedding BGE Small dan MiniLM-L6 terhadap kualitas pengambilan dokumen medis. Selain itu, penelitian ini secara eksplisit bertujuan untuk menentukan konfigurasi parameter data yang optimal berdasarkan metrik *Context Recall* dan *Context Precision*, serta mengimplementasikannya ke dalam prototipe chatbot medis sebagai bentuk validasi akhir untuk meminimalkan risiko halusinasi sistem.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan eksperimental komparatif untuk menguji enam skenario kinerja sistem *Retrieval-Augmented Generation* (RAG) di bidang kedokteran. Proses pengembangan disusun menggunakan bahasa pemrograman *Python*, dengan memanfaatkan kerangka kerja *LangChain* untuk mengintegrasikan semua komponen [18], mulai dari pra-pemrosesan data hingga tahap evaluasi. Secara keseluruhan, tahapan eksperimental sistem RAG dalam penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

Tahap pertama penelitian ini berfokus pada pengumpulan dokumen yang berfungsi sebagai basis pengetahuan utama bagi sistem chatbot. Basis pengetahuan ini disusun dari dokumen Pedoman Praktis Klinis yang bersumber dari lembaga tepercaya seperti National Center for Biotechnology Information (NCBI) [19], Organisasi Kesehatan Dunia (WHO) [20], Kementerian Kesehatan Republik Indonesia [21], Alodoq [22], Halodoq [23], dan Biofarma [24].

Dokumen-dokumen ini mencakup informasi spesifik mengenai pengobatan mandiri dan penanganan penyakit ringan seperti flu, diare, sakit kepala, sakit tenggorokan, dan maag, yang merupakan domain fokus layanan konsultasi di Klinik Umum Pratama Nusa Putra Farmedika. Semua data teks yang dikumpulkan dari lembaga-lembaga tersebut kemudian diproses pada tahap pengelompokan teks untuk mendukung penerapan arsitektur Retrieval-Augmented Generation, guna memastikan bahwa informasi yang diberikan oleh chatbot akurat dan relevan dengan konteks klinis.

2.2 Prapemrosesan & Chunking

Dokumen digital diekstraksi ke dalam memori sistem menggunakan pustaka *PyPDFLoader*. Teks mentah tersebut kemudian dipotong-potong menjadi segmen informasi yang lebih kecil (*chunking*) menggunakan teknik *RecursiveCharacterTextSplitter* [25]. Eksperimen dirancang dengan tiga variasi ukuran potongan, yaitu 256, 512, dan 1024 token. Pada setiap variasi, nilai tumpang tindih potongan ditetapkan konstan pada 50 token untuk menjaga kelangsungan informasi klinis. Perkiraan jumlah potongan (N) yang dihasilkan secara teoritis dihitung menggunakan persamaan matematika berikut:

$$N = \frac{TotalKarakter - Overlap}{ChunkSize - Overlap} \quad (1)$$

2.3 Model Embedding BGE Small

Segmen teks medis yang telah melalui tahap chunking kemudian diubah menjadi representasi vektor numerik. Pemodelan pertama menggunakan model *BGE Small*, yang diterapkan karena memiliki tingkat efisiensi tinggi dalam menangkap relevansi kontekstual dalam data teks [26]. Penggunaan model ini bertujuan untuk mengubah data medis dari klinik menjadi representasi vektor numerik, yang memungkinkan penempatan kata-kata dengan makna serupa sehingga saling berdekatan dalam ruang dimensi vektor. Melalui kemampuan pemetaan semantik ini, model *BGE Small* diharapkan dapat menyederhanakan proses pencarian dan pencocokan data sehingga hasil pengambilan informasi menjadi lebih akurat.

2.4 Model Embedding MiniLM-L6

Sebagai skenario perbandingan, pemodelan kedua menggunakan model *MiniLM-L6*. Model ini bekerja dengan menangkap makna atau fitur semantik dari potongan teks pedoman klinis dan mengubahnya menjadi format vektor berdimensi tinggi. Model *MiniLM-L6* digunakan karena menawarkan keunggulan berupa keseimbangan optimal antara kecepatan komputasi pemrosesan dan akurasi pemetaan vektor [27]. Pemilihan model ini secara khusus bertujuan untuk mengidentifikasi sejauh mana arsitektur model yang lebih ringan dapat mendukung fungsi tahap pencarian dalam menemukan informasi yang relevan bagi pengguna chatbot medis.

2.5 Pengindeksan Penyimpanan Vectok Chroma DB

Vektor numerik yang dihasilkan dari kedua model embedding tersebut secara otomatis diindeks menggunakan fungsi *Chroma.from_documents* [28]. Proses ini menyimpan representasi vektor secara permanen ke dalam direktori penyimpanan permanen di penyimpanan vektor *Chroma DB*. Hal ini memungkinkan basis pengetahuan dapat diakses secara efisien untuk pencarian semantik pada tahap selanjutnya.

2.6 Pengujian Metrik RAGAS

Tahap terakhir adalah mengevaluasi kualitas hasil pencarian menggunakan kerangka kerja *RAGAS* yang memanfaatkan LLM Llama 3.1 sebagai alat evaluasi otomatis [29], [30]. Pengujian berfokus pada dua metrik utama, yaitu *Context Recall* untuk menilai kelengkapan informasi, dan *Context Precision* untuk menilai keakuratan relevansi informasi. Dalam protokol evaluasi, pengujian dilakukan menggunakan dataset evaluasi yang terdiri dari 10 pasang pertanyaan medis dan jawaban referensi (*ground truth*) yang diambil dari pedoman klinis. Evaluasi ini memproses pengambilan data terhadap 1 dokumen sumber terkompilasi yang telah diindeks. Konfigurasi *RAGAS* menggunakan LLM Llama 3.1 sebagai evaluator (*LLM-as-a-judge*), dengan parameter pengambilan data diatur ke top-k=4 untuk mengambil konteks yang relevan. Untuk mengidentifikasi konfigurasi yang paling optimal, kinerja sistem diukur dengan menghitung nilai rata-rata (*Average Score*) menggunakan persamaan berikut:

$$average\ score = \frac{Context\ Recall + Context\ Precision}{2} \quad (2)$$

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengumpulan Data Dokumen Medis

Proses pengumpulan data berhasil menyusun basis pengetahuan yang bersumber dari dokumen kesehatan resmi, termasuk pedoman penanganan penyakit ringan seperti flu, diare, sakit kepala, sakit tenggorokan, maag, demam, dan pilek. Seluruh teks dari dokumen-dokumen tersebut berhasil digabungkan dan disiapkan di ruang kerja sistem sebagai data mentah yang siap diekstraksi.

Tabel 1. Hasil Dataset

No	Pertanyaan	Jawaban
1	Apa itu penyakit flu atau influenza?	Flu atau influenza adalah penyakit infeksi saluran pernapasan akut yang menyerang hidung, tenggorokan, dan paru-paru yang disebabkan oleh virus influenza tipe A, B, atau C. Penyakit ini sangat umum terjadi di musim pancaroba dan sangat mudah menular.
2	Apa saja gejala umum yang dirasakan penderita flu?	Gejala umum flu meliputi demam, otot sakit, menggigil dan berkeringat, sakit kepala, batuk kering

		kronis, serta sesak napas. Selain itu, penderita mungkin mengalami kelelahan, hidung tersumbat, dan sakit tenggorokan.
3	Apa rekomendasi obat untuk meredakan gejala flu?	Beberapa rekomendasi obat flu adalah Vicks Formula 44, Rhinos SR, Intunal F, Panadol Cold & Flu, Nalgestan, Anadex, Demacolin, dan Alpara. Obat-obatan ini berfungsi meredakan gejala seperti demam, hidung tersumbat, dan batuk.
4	Apa definisi penyakit diare?	Diare didefinisikan sebagai kondisi peningkatan frekuensi buang air besar dengan feses yang cair atau encer. Kondisi ini sering disertai mual, muntah, kram perut, dan terkadang penurunan berat badan.
5	Apa saja faktor penyebab terjadinya diare?	Penyebab diare meliputi infeksi usus (virus, bakteri, atau parasit), intoleransi makanan (seperti laktosa), efek samping obat-obatan tertentu, kondisi pencernaan kronis (IBS atau Crohn), serta faktor psikologis seperti stres dan kecemasan.
6	Bagaimana langkah awal pengobatan diare yang bisa dilakukan di rumah?	Langkah utama adalah rehidrasi dengan minum banyak cairan seperti air putih dan minuman isotonik untuk mengganti cairan tubuh yang hilang. Selain itu, konsumsilah makanan yang lembut dan mudah dicerna seperti bubur atau pisang, serta hindari makanan pedas dan berlemak.
7	Apa yang dimaksud dengan radang tenggorokan?	Radang tenggorokan adalah gangguan yang menimbulkan rasa sakit, gatal, dan nyeri pada tenggorokan, terutama saat menelan makanan atau minuman.
8	Apa penyebab paling umum dari sakit tenggorokan?	Penyebab paling umum adalah infeksi virus seperti virus influenza (flu) atau virus pilek. Namun, radang tenggorokan juga dapat dipicu oleh infeksi bakteri kelompok Streptococcus.
9	Apa saja gejala yang menandakan seseorang terkena radang tenggorokan?	Selain rasa nyeri saat menelan, gejala lainnya meliputi hidung beringus, sering bersin, mual, demam, kelelahan, nyeri otot, batuk, hingga amandel yang membengkak.
10	Apa komplikasi yang mungkin muncul jika radang tenggorokan tidak ditangani dengan tepat?	Komplikasi yang mungkin terjadi meliputi gangguan tidur (sleep apnea), dehidrasi karena sulit menelan, gangguan pernapasan seperti asma atau sinusitis, epiglotitis (pembengkakan tulang rawan tenggorokan), serta terbentuknya abses atau kantong nanah di sekitar amandel.

3.2 Hasil Prapemrosesan Data & Chunking

Teks mentah yang berhasil dimuat menggunakan *PyPDFLoader* kemudian dieksekusi oleh algoritma *RecursiveCharacterTextSplitter*. Proses pemotongan ini memprioritaskan integritas semantik dengan tidak memotong kalimat di tengah-tengah, sehingga jumlah potongan yang sebenarnya dapat berubah untuk mengakomodasi integritas paragraf medis. Rincian distribusi jumlah potongan teks untuk setiap variasi ukuran disajikan pada Tabel 2.

Tabel 2. Hasil Chunking

No	Variasi Chunk	Hasil Potongan
1	Chunk Size 256	429 Chunk
2	Chunk Size 512	206 Chunk
3	Chunk Size 1024	108 Chunk

Berdasarkan tabel 2, data ini menunjukkan hubungan berbanding terbalik di mana semakin kecil batasan token yang ditetapkan, semakin banyak jumlah segmen teks yang dihasilkan oleh sistem. Secara spesifik, penggunaan ukuran chunk 256 menghasilkan tingkat fragmentasi informasi yang paling tinggi dengan total 429 potongan. Hal ini mengindikasikan bahwa dokumen panduan klinis dipecah menjadi unit-unit informasi yang sangat granular atau tingkat ke-detailan data sangat rinci. Dalam konteks data medis, granularitas yang tinggi ini

berpotensi memisahkan satu topik spesifik seperti gejala penyakit tertentu dari topik lainnya, sehingga meminimalkan risiko tercampurnya konteks yang berbeda dalam satu blok vektor.

Sebaliknya, pada ukuran chunk 1024 yang hanya menghasilkan 108 potongan, informasi tersimpan dalam blok-blok yang jauh lebih besar. Meskipun hal ini mengurangi jumlah total data yang harus diindeks, ukuran yang terlalu besar berisiko menggabungkan beberapa topik atau konteks yang berbeda seperti definisi penyakit, gejala, dan pengobatan secara sekaligus ke dalam satu representasi, yang dapat mempengaruhi ketepatan saat sistem melakukan pencarian informasi spesifik.

3.3 Hasil Model BGE Small

Potongan teks dari ketiga variasi di atas diubah menjadi representasi vektor menggunakan model *BGE Small*. Pengukuran kualitas pencarian untuk model ini menunjukkan kinerja yang sangat mengesankan, terutama pada ukuran teks sedang. Hasil metrik *RAGAS* khusus untuk model *BGE Small* diuraikan dalam Tabel 3.

Tabel 3. Kinerja Kualitas Pencarian Model *BGE Small*

Variasi Chunk	Context Recall	Context Precision	Average Score
256 Chunk	0.22	0.56	0.39
512 Chunk	0.45	0.74	0.59
1024 Chunk	0.15	0.58	0.37

Berdasarkan tabel 3, model *BGE Small* mencapai akurasi tertinggi pada ukuran potongan teks sebesar 512, dengan nilai rata-rata 0,59. Nilai *Context Precision* yang mengesankan (0,74) menunjukkan bahwa model ini sangat akurat dalam memetakan terminologi klinis lokal, memastikan bahwa sebagian besar dokumen yang ditemukan sangat relevan dengan pertanyaan yang diajukan.

3.4 Hasil Model MiniLM-L6

Sebagai perbandingan, model *MiniLM-L6* juga memproses variasi teks yang sama. Meskipun model ini menawarkan kecepatan komputasi yang efisien, kemampuan penangkapan semantiknya pada teks medis berbahasa Indonesia tidak sekuat *BGE Small*. Hasil metrik untuk model *MiniLM-L6* disajikan pada Tabel 4.

Tabel 4. Kinerja Kualitas Pencarian Model *MiniLM-L6*

Variasi Chunk	Context Recall	Context Precision	Average Score
256 Chunk	0.06	0.49	0.27
512 Chunk	0.17	0.67	0.42
1024 Chunk	0.19	0.36	0.28

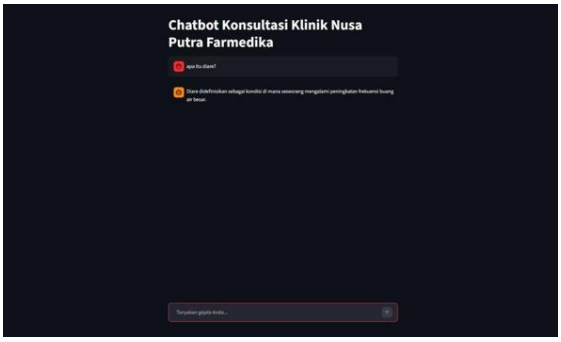
Berdasarkan pada Tabel 4 menunjukkan bahwa *MiniLM-L6* mencapai kinerja puncaknya pada ukuran chunk 512, dengan skor rata-rata 0,42. Namun, pada ukuran ekstrem, kinerjanya menurun secara signifikan. Pada ukuran *chunk* 256, meskipun *Context Precision* mencapai 0,49, Context Recall anjlok ke titik terendah absolut 0,06, sehingga menghasilkan skor rata-rata 0,27. Demikian pula, pada 1024 token, skor rata-rata berada di angka 0,28. Hal ini menunjukkan bahwa arsitektur yang lebih ringan ini sangat rentan terhadap efek fragmentasi teks ekstrem dan gangguan informasi, sehingga gagal mengambil sebagian besar informasi yang relevan.

3.5 Pengindeksan Hasil ke Penyimpanan Vektor Chroma DB

Representasi numerik multidimensi dari kedua model di atas berhasil diindeks dan disimpan secara permanen di penyimpanan vektor *Chroma DB*. Proses pengindeksan permanen ini menghilangkan kebutuhan akan perhitungan vektor berulang setiap kali ada kueri masuk. Direktori vektor ini terbukti stabil dan merespons pemanggilan kueri uji dengan latensi yang sesuai untuk fase evaluasi.

3.6 Perbandingan Evaluasi Metrik RAGAS dan Validasi Prototipe

Analisis gabungan dari semua hasil eksperimen menegaskan bahwa kombinasi model *BGE Small* dengan 512 chunk merupakan konfigurasi optimal mutlak. Konfigurasi ini mendominasi baik *Context Recall* maupun *Context Precision*. Sebagai bentuk validasi fungsional dari metrik-metrik ini, parameter-parameter optimal tersebut kemudian diintegrasikan langsung ke dalam prototipe aplikasi web chatbot medis Klinik Nusa Putra Farmedika. Sebagaimana diilustrasikan pada Gambar 2, chatbot berhasil mengambil dokumen pedoman klinis dengan presisi dan meneruskannya ke LLM untuk menghasilkan jawaban pengobatan mandiri yang terstruktur dan faktual yang meminimalkan risiko halusinasi.



Gambar 2. Tampilan Chatbot

4. KESIMPULAN

Berdasarkan hasil pengujian metrik evaluasi *RAGAS*, kinerja sistem *Retrieval-Augmented Generation* (RAG) pada domain medis menunjukkan hasil yang sangat bergantung pada ukuran chunk dan model embedding yang digunakan. Hasil eksperimen membuktikan adanya pola non-linear pada

pemotongan teks, di mana ukuran 512 token berhasil memberikan keseimbangan retensi konteks klinis terbaik dibandingkan ukuran 256 token (yang menyebabkan fragmentasi ekstrem) dan 1024 token (yang memasukkan noise). Pada komparasi representasi ruang vektor, model *BGE Small* secara konsisten mendominasi dan mengungguli performa model *MiniLM-L6* pada seluruh variasi ukuran teks.

Puncak performa dari model *MiniLM-L6* hanya mampu mencapai skor rata-rata 0,42. Sebaliknya, konfigurasi data yang terbukti paling optimal secara absolut dicapai oleh integrasi model *BGE Small* dengan ukuran *chunk* 512 token. Konfigurasi terbaik ini menghasilkan skor rata-rata tertinggi sebesar 0,59, yang didorong oleh tingginya nilai ketepatan relevansi informasi (*Context Precision*) sebesar 0,74 dan kelengkapan penemuan informasi (*Context Recall*) sebesar 0,45. Hasil validasi fungsional dari temuan angka tersebut pada purwarupa chatbot medis membuktikan bahwa konfigurasi ini berhasil meneruskan dokumen pedoman klinis dengan presisi, meminimalkan risiko halusinasi.

REFERENCES

[1] Komdigi, "Transformasi Digital Bersama Kementrian," Kementrian Komunikasi dan Digital. Accessed: Oct. 22, 2025. [Online]. Available: https://www.komdigi.go.id/transformasi-digital

[2] K. U. P. N. P. Farmedika, "Klinik Umum Pratama Nusa Putra Farmedika." [Online]. Available: https://clinic.nusaputra.ac.id/

[3] M. Farwati, I. T. Salsabila, K. R. Navira, and T. Sutabri, "ANALISA PENGARUH TEKNOLOGI ARTIFICIAL INTELLIGENCE (AI) DALAM KEHIDUPAN SEHARI-HARI," *Jurnal Sistem Informasi dan Manajemen*, vol. 11, no. 1, pp. 39–45, 2023, doi: https://doi.org/10.47024/js.v11i1.563.

[4] K. D. Ningtyas, R. Kurniawan, and A. Armansyah, "Penerapan Natural Language Processing Pada Aplikasi Chatbot Info Layanan Kantor Menggunakan Naive Bayes Algorithm," *Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD*, vol. 6, no. 2, pp. 266–271, 2023, doi: https://doi.org/10.53513/jsk.v6i1.7413.

[5] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, pp. 2–9, 2021.

[6] C. Karaniya Wigayha, "Analisis Sistematis Penggunaan Large Language Models (Llms) Dan Artificial Intelligence (Ai) Untuk Peningkatan Literasi Digital Pada Jenjang Pendidikan Tinggi," *PT Dinamika Publishing International*, vol. 1, no. 1, pp. 1–7, 2025, [Online]. Available: https://jurnal.dinamikapublika.id/index.php/EXGEN/article/view/10

[7] R. Ahadi, N. Safaat Harahap, M. Fikry, and F. Kurnia, "Retrieval-Augmented Generation in a Web-Based Question Answering System for Fiqh Books," *Journal of Artificial Intelligence and Software Engineering*, vol. 5, no. 2, pp. 626–635, 2025, doi: 10.30811/jaise.v5i2.7005.

[8] I. Fanani, "IMPLEMENTASI RETRIEVAL AUGMENTED GENERATION UNTUK EVALUASI PROPOSAL TUGAS AKHIR MAHASISWA," *Jurnal Teknologi Komputer dan Informatika*, vol. 3, no. 2, 2025, doi: https://doi.org/10.59820/tekomin.v3i2.336.

[9] R. MARLINA, "SISTEM TANYA JAWAB PERNIKAHAN DALAM ISLAM BERBASIS WEB," *Jurnal Informatika, Manajemen dan Komputer*, pp. 1–11, 2024.

[10] E. A. Prasetyo, *Chatbot untuk Informasi Pembangunan Wilayah Kota Semarang menggunakan Metode Retrieval Augmented Generation (RAG)*. 2024. [Online]. Available: http://ecampus.poltekkes-medan.ac.id/jspui/handle/123456789/1726

[11] J. RISAKOTTA, "Penerapan Chunking Strategy Untuk Meningkatkan Kemampuan Memahami Teks Dalam Bahasa Inggris Pada Smk Kesehatan Nusaniwe Ambon," *VOCATIONAL: Jurnal Inovasi Pendidikan Kejuruan*, vol. 2, no. 4, pp. 327–334, 2023, doi: 10.51878/vocational.v2i4.1751.

[12] M. Susanty and S. Sukardi, "Perbandingan Pre-trained Word Embedding dan Embedding Layer untuk Named-Entity Recognition Bahasa Indonesia," *Petir*, vol. 14, no. 2, pp. 247–257, 2021, doi: 10.33322/petir.v14i2.1164.

[13] 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. 1–6, 2026.

[14] C. and C. with Navnit, "Day 13 : Master RAG: How to Find the Ideal Chunk Size for Better AI Retrieval," 2026. [Online]. Available: https://youtu.be/lTdGt3vIhsY?si=F_cR9uLKgMBdorYN

[15] M. A. Mersha, M. Gemeda Yigezu, and J. Kalita, "Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms," *Procedia Comput.*

Sci., vol. 244, pp. 121–132, 2024, doi: 10.1016/j.procs.2024.10.185.
[16] I. L. Kharisma, M. S. Hidayat, Somantri, and Kamdan, "Implementasi Retrieval Augmented Generation dalam Sistem Chatbot Dermatologi Berbasis Website," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 11, no. 3, pp. 448–462, 2025, doi: https://doi.org/10.28932/jutisi.v11i3.12258.
[17] S. Aliphadji Talaohu, R. Soekarta, and M. Surahmanto, "Implementasi LLM Pada Chatbot PMB Universitas Muhammadiyah Sorong Menggunakan Metode RAG Berbasis Website," *Jurnal Ilmu Komputer dan Informatika*, vol. 03, no. 02, pp. 1–11, 2025.
[18] O. Topsakal and T. C. Akinci, "Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast," *International Conference on Applied Engineering and Natural Sciences*, vol. 1, no. 1, pp. 1050–1056, 2023, doi: 10.59287/icaens.1127.
[19] NCBI, "National Center for Biotechnology Information," National Center for Biotechnology Information. Accessed: Nov. 10, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/
[20] WHO, "World Health Organization," World Health Organization. Accessed: Nov. 10, 2025. [Online]. Available: https://www.who.int/
[21] K. RI, "Kementrian Kesehatan Republik Indonesia," Kementrian Kesehatan Republik Indonesia. Accessed: Nov. 11, 2025. [Online]. Available: https://www.kemkes.go.id/id/home
[22] Alodokter, "Alodokter," Alodokter. Accessed: Nov. 13, 2025. [Online]. Available: https://www.alodokter.com/
[23] Halodoq, "Halodoq." Accessed: Nov. 15, 2025. [Online]. Available: https://www.halodoc.com/?srsltid=AfmBOoq3etemB8IKchp1ukv0m9OZ40TqD28FdX-znWNZLNt_oZuRhAcY
[24] Biofarma, "PT Bio Farma," Biofarma Group. Accessed: Nov. 15, 2025. [Online]. Available: https://www.biofarma.co.id/id/pt-bio-farma-persero
[25] S. Latif, H. Ameer, M. H. Akram, and M. Fatima, "The Chunking Paradigm: Recursive Semantic for RAG Optimization," *Association for Computational Linguistics*, pp. 137–145, 2025, [Online]. Available: https://aclanthology.org/2025.icnlsp-1.15/
[26] S. Lai *et al.*, "Enhancing Technical Documents Retrieval for RAG," in *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2025, pp. 1176–1181. doi: 10.1109/APSIPAASC65261.2025.11249099.
[27] R. Sajja, Y. Sermet, and I. Demir, "Domain-specific embedding models for hydrology and environmental sciences: enhancing semantic retrieval and question answering," *Water Science and Technology*, vol. 92, no. 9, pp. 1328–1342, 2025, doi: 10.2166/wst.2025.156.
[28] A. Z. Abidin and M. M. Engel, "Comparative Analysis of Performance Aspects Between Chroma and Pgvector as a Vector Database," *bit-Tech*, vol. 8, no. 2, pp. 2079–2090, 2025, doi: 10.32877/bt.v8i2.3198.
[29] M. Antal and K. Buza, "Evaluating Open-Source LLMs in RAG Systems: A Benchmark on Diploma Theses Abstracts Using Ragas," *Acta Universitatis Sapientiae, Informatica*, vol. 17, no. 1, pp. 1–15, 2025, doi: 10.1007/s44427-025-00006-3.
[30] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations*, pp. 150–158, 2024, doi: 10.18653/v1/2024.eacl-demo.16.