JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 181~188
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Implementasi *Retrieval Augmented Generation* (RAG) Dalam Perancangan *Chatbot* Kesehatan Pencernaan

[1]Gufranaka Samudra, [2] Ahmad Turmudi Zy, [3]Ermanto
[1,2,3]Universitas Pelita Bangsa, Indonesia
[1]gufranakasamudra348@gmail.com; [2]turmudi@pelitabangsa.ac.id; [3]ermanto@pelitabangsa.ac.id;

Article Info

Article history:

Received, 2025-01-14
Revised, 2025-01-18
Accepted, 2025-01-22

Kata Kunci:

Retrieval-Augmented Generation
Chatbot
kesehatan pencernaan
Large Language Models

ABSTRAK

Perkembangan teknologi kecerdasan buatan, khususnya dalam pengembangan chatbot, telah membawa kemajuan signifikan, terutama di bidang kesehatan. Namun, tantangan utama dalam penggunaan model bahasa besar (*Large Language Models*/LLM) adalah potensi bias dan kurangnya akurasi dalam memberikan informasi, terutama pada topik kritis seperti kesehatan pencernaan. Penelitian ini bertujuan untuk mengimplementasikan *Retrieval-Augmented Generation* (RAG) dalam perancangan chatbot kesehatan pencernaan guna meningkatkan akurasi dan relevansi informasi yang disampaikan. Metode RAG mengintegrasikan model generatif dengan sistem *retrieval* berbasis dokumen untuk memberikan jawaban yang lebih terpercaya dan berbasis bukti. Proses penelitian melibatkan pengumpulan dataset kesehatan pencernaan melalui *scraping* data dari Alodokter, serta pengolahan data melalui tahap preprocessing, embedding menggunakan model bahasa Indonesia (firqaaa/indo-sentence-bert-base), dan pengolahan data menggunakan *database* vektor dengan *index HNSW*. Model Llama 3.1:8b digunakan untuk menghasilkan respons generatif. Hasil penelitian menunjukkan bahwa penerapan RAG dapat mengurangi bias model dan meningkatkan kualitas respons *chatbot*. Evaluasi menggunakan metrik seperti *Mean Reciprocal Rank* (MRR) 93%, *Faithfullness* 62%, *Answer Relevancy* 57%, dan *Semantic Similarity* 81% menunjukkan kinerja yang baik dalam memberikan jawaban yang akurat dan relevan sesuai konteks. Dengan pendekatan ini, *chatbot* mampu memberikan informasi yang lebih akurat dan kontekstual sesuai dengan kebutuhan pengguna, serta dapat mengurangi risiko halusinasi dalam informasi yang diberikan. Penelitian ini memberikan kontribusi dalam pengembangan teknologi *chatbot* kesehatan yang lebih andal, khususnya di domain kesehatan pencernaan, dan membuka peluang untuk pengaplikasian lebih lanjut pada bidang kesehatan lainnya.

Keywords:

Retrieval-Augmented Generation
chatbot
digestive health
Large Language Models

ABSTRACT

The development of artificial intelligence technology, especially in the development of chatbots, has brought significant progress, especially in the health sector. However, the main challenge in using large language models (LLM) is the potential for bias and lack of accuracy in providing information, especially on critical topics such as digestive health. This study aims to implement Retrieval-Augmented Generation (RAG) in designing a digestive health chatbot to improve the accuracy and relevance of the information delivered. The RAG method integrates a generative model with a document-based retrieval system to provide more reliable and evidence-based answers. The research process involves collecting digestive health datasets through data scraping from Alodokter, as well as data processing through the preprocessing stage, embedding using the Indonesian language model (firqaaa/indo-sentence-bert-base), and data processing using a vector database with the HNSW index. The Llama 3.1:8b model is used to generate generative responses. The results of the study show that the application of RAG can reduce model bias and improve the quality of chatbot responses. Evaluation using metrics such as Mean Reciprocal Rank (MRR) 93%, Faithfullness 62%, Answer Relevancy 57%, and Semantic Similarity 81% showed good performance in providing accurate and relevant answers according to context. With this approach, chatbots are able to provide more accurate and contextual information according to user needs, and can reduce the risk of hallucinations in the information provided. This research contributes to the development of more reliable health chatbot technology, especially in the digestive health domain, and opens up opportunities for further application in other health fields

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 181~188
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Penulis Korespondensi:

Gufranaka Samudra,
Program Studi Teknik Informatika,
Universitas Pelita Bangsa,
Email: gufranakasamudra348@gmail.com

1. PENDAHULUAN

Kemajuan luar biasa pada *Large Language Models* (*LLMs*) telah membuka peluang besar dalam pengembangan kecerdasan buatan, khususnya di bidang kesehatan dan chatbot [1]Salah satu pendekatan yang semakin populer adalah *Retrieval-Augmented Generation* (*RAG*), yang banyak digunakan dalam pengembangan chatbot berbasis pemrosesan bahasa alami (*Natural Language Processing*/*NLP*). *RAG* bertujuan untuk meningkatkan hasil model generatif *LLM* dengan memanfaatkan data eksternal, sehingga beberapa perusahaan mulai mengadopsi metode ini sebagai alternatif yang lebih efisien dibandingkan dengan *fine-tuning LLM* untuk mengintegrasikan pengetahuan domain khusus [2], [3].

RAG menjadi solusi andal untuk menghasilkan respons berbasis konteks, terutama dalam menangani tugas-tugas kompleks yang membutuhkan pengetahuan mendalam. Keunggulannya terletak pada efisiensi sumber daya komputasi, yang jauh lebih ringan dibandingkan metode *fine-tuning* [4], [5], [6]. Dalam pendekatan ini, *RAG* menggunakan data eksternal yang disimpan dalam basis data untuk mendukung proses pengambilan informasi. Dengan demikian, implementasi *RAG* memungkinkan penyimpanan informasi dalam jumlah besar, yang memastikan bahwa jawaban yang diberikan bersumber langsung dari data yang tersedia dalam basis data. Hal ini membuat *RAG* sangat cocok untuk penelitian ini, yang berfokus pada perancangan chatbot kesehatan pencernaan [7], [8].

Melalui pemanfaatan data kesehatan eksternal, *RAG* membantu *LLM* memberikan jawaban berdasarkan konteks yang relevan, sehingga dapat mengurangi risiko halusinasi pada *LLM*. Pendekatan ini sangat penting karena data kesehatan merupakan jenis data yang sensitif, sehingga diperlukan rancangan chatbot yang efisien secara komputasi namun tetap akurat dalam memberikan informasi [9], [10].

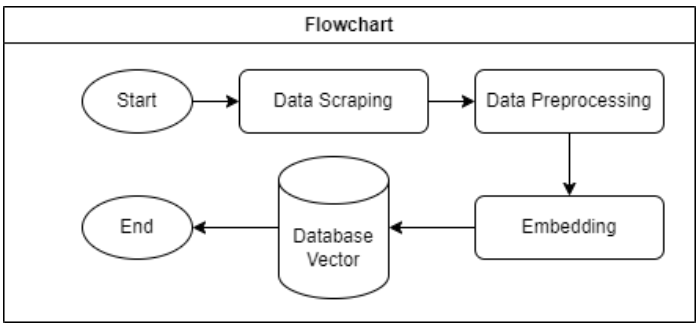
Selain itu, terdapat teknik lanjutan seperti *OpenRAG*, *single-hop*, dan *multi-hop retrieval*, serta metode *sequence-to-sequence* (*seq2seq*) yang dapat digunakan untuk meningkatkan efisiensi pengambilan dan penyajian informasi pada sistem berbasis *RAG*. Teknik-teknik ini memungkinkan integrasi data yang lebih kompleks dan mendukung kemampuan chatbot untuk menangani skenario percakapan yang lebih mendalam dan spesifik [11], [12].

Penelitian-penelitian sebelumnya menunjukkan bahwa *RAG* memiliki keunggulan dalam hal akurasi dan relevansi respons yang diberikan. Pendekatan ini sangat mendukung penyelesaian tugas-tugas kompleks seperti chatbot kesehatan pencernaan, yang membutuhkan data spesifik untuk menghasilkan respons yang tepat dan relevan [13], [14]. Dengan integrasi metode tersebut, diharapkan penelitian ini dapat memberikan kontribusi dalam pengembangan chatbot berbasis kesehatan pencernaan dengan tingkat akurasi dan efisiensi yang tinggi.

Selain sebagai alat pendukung kesehatan, chatbot berbasis *RAG* juga memiliki potensi besar untuk digunakan dalam sektor komersial. Teknologi ini dapat diterapkan dalam berbagai aplikasi, seperti memberikan rekomendasi produk kesehatan, menjelaskan manfaat alat atau obat medis, hingga membantu konsumen dalam mengambil keputusan pembelian secara lebih terinformasi. Dengan kemampuan menghasilkan respons yang relevan dan berbasis data, chatbot ini tidak hanya meningkatkan pengalaman pengguna tetapi juga membuka peluang monetisasi yang signifikan bagi perusahaan yang bergerak di bidang kesehatan [15].

2. METODE PENELITIAN

Metode yang digunakan dalam penelitian ini adalah metode eksperimen. Penelitian ini dirancang mengikuti tahapan-tahapan yang sesuai dengan perancangan *Retrieval Augmented Generation (RAG)*. Berikut diagram alur gambaran besar proses penelitian:



Gambar 2.1 *Flowchart*

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 181~188
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Adapun detail tahapan eksperimen yang dilakukan adalah sebagai berikut:

Pengambilan Data *(Data Scraping)*

Data diambil dari website Alodokter, yang menyediakan halaman khusus bagi komunitas untuk bertanya dan dijawab langsung oleh dokter ahli. Teknik *scraping* dilakukan menggunakan pustaka Python seperti *BeautifulSoup* untuk mengumpulkan data dari berbagai kategori penyakit yang tersedia di situs tersebut. Dalam penelitian ini, data difokuskan secara spesifik pada kategori kesehatan pencernaan, seperti GERD, maag, diare, konstipasi, dan gangguan usus lainnya.

- **Sumber data**: Website Alodokter (kategori kesehatan terkait pencernaan).
- **Rasio data**: Setelah dilakukan *scraping*, terkumpul sebanyak **7.860 data**.

Pembersihan Data *(Data Cleaning)*

Setelah data diperoleh, dilakukan proses pembersihan data untuk memastikan konsistensi dan kesesuaian dengan kebutuhan penelitian. Proses pembersihan meliputi:

- Mengubah seluruh teks menjadi huruf kecil (lowercase).
- Membatasi panjang kalimat maksimal 1.000 karakter pada kolom jawaban.
- Mengambil bagian tengah kalimat sepanjang 800 karakter untuk jawaban yang terlalu panjang.
- Koreksi dan standarisasi kolom pertanyaan.
- Menggabungkan beberapa kolom menjadi satu kolom tunggal (pertanyaan dan jawaban).
- Menghapus kolom yang tidak relevan.

Pembuatan Representasi Kalimat *(Sentence Embedding)*

Representasi kalimat dilakukan untuk mengubah teks menjadi vektor angka menggunakan model *embedding*. Model yang digunakan dalam penelitian ini adalah *SBERT (Sentence-BERT)*, yang dioptimalkan untuk tugas *retrieval.* Setelah proses *embedding* selesai, hasil *embedding* akan disimpan kedalam database vektor untuk melakukan *searching* vektor.

Evaluasi Proses Pencarian Informasi *(Retrieval Evaluation)*

Proses evaluasi metode *Retrieval Augmented Generation* dilakukan secara menual melihat dan mengoreksi datanya, disini penulis menggunakan 10 sample data dalam melakukan evaluasi. Evaluasi matrik yang digunakan adalah *MRR (Mean Reciprocal Rank)* Berikut merupakan hasilnya,

Tabel 2.1 Evaluasi Retrieval

Dokumen	Rank in K	Result
Doc 1	1	1/1
Doc 2	1	1/1
Doc 3	1	1/1
Doc 4	1	1/1
Doc 5	3	1/3
Doc 6	1	1/1
Doc 7	1	1/1
Doc 8	1	1/1
Doc 9	1	1/1
Doc 10	1	1/1

Evaluasi Proses Generasi Jawaban *(Generative Evaluation)*

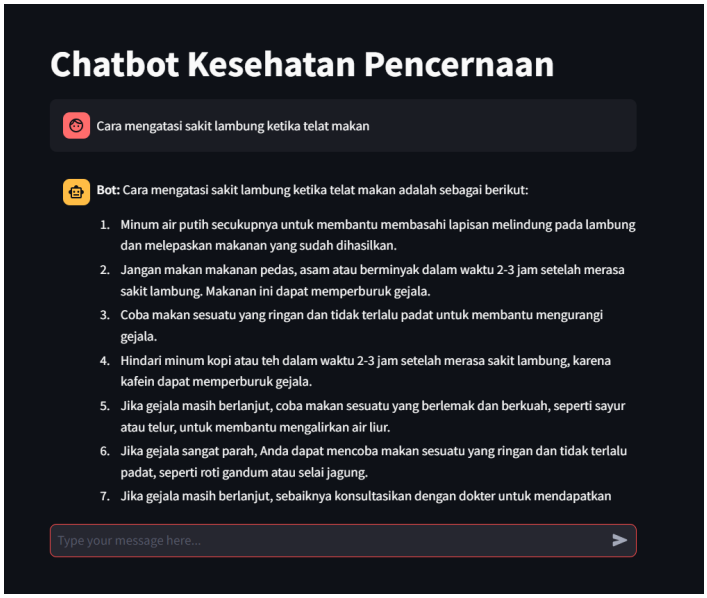
Evaluasi pada proses *generation* memiliki tiga matrik yaitu, *faithfullness*, *answer relevancy*, *semantic similarity*. Dengan tujuan untuk menghitung seberapa baik hasil dari generation *RAG*. Berikut merupakan hasilnya,

Tabel 2.2 Evaluasi Generation

Dokumen	Faithfullness	Answer Relevancy	Semantic Similarity
Doc 1	0.6	0.566872	0.886848
Doc 2	0.6	0.626669	0.797220
Doc 3	0	0.480409	0.599470
Doc 4	0.8	0.633684	0.912981
Doc 5	0.6	0.611504	0.888766
Doc 6	0.8	0.578719	0.740155
Doc 7	0.8	0.422533	0.757962
Doc 8	0.8	0.680173	0.890405
Doc 9	0.6	0.667278	0.867403
Doc 10	0.6	0.501049	0.857279

Perancangan Aplikasi berbasis Web

Dalam kemudahan akses maka di sediakan aplikasi web sederhana untuk mengakses *chatbot* kesehatan pencernaan, dengan desain yang sederhana agar mudah dalam penggunaannya. Berikut merupakan desain aplikasi *chatbot* kesehatan pencernaan menggunakan streamlit,

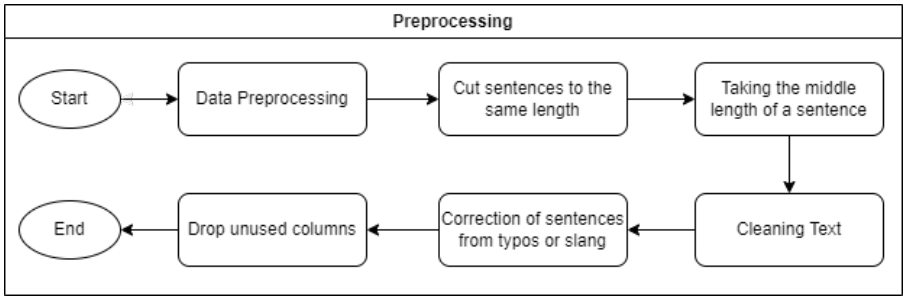


Gambar 2.2 Tampilan Aplikasi

3. HASIL DAN ANALISIS

Persiapan Data

Pengumpulan data dilakukan dengan cara scraping data pada website alodokter, sebuah website kesehatan yang ada di indonesia. Data umum berhasil di kumpulkan sebanyak lebih dari dua ratus ribu data kemudian dilakukan filtering untuk masalah pencernaan dan didapatkan jumlah akhir data sebanyak **7.860 data**. Kemudian tahap selanjutnya dilakukan *cleaning* data seperti, *lowercase*, mengatur panjang kalimat hanya seribu pada kolom jawaban, mengambil kalimat tengah sebanyak 800, koreksi kolom pertanyaan, menggabungkan kolom menjadi satu kolom, menghapus kolom yang tidak digunakan.



Gambar 3.1 *Preprocessing*

Embedding

Setelah data bersih, proses *embedding* data menggunakan *LLM* khusus *embedding* dari huggingface dengan nama firqaaa/indo-sentence-bert-base yang merupakan model *embedding* khusus bahasa indonesia. Huggingface merupakan sebuah platform untuk mendapat model-model *open-source* siap pakai, yang merupakan bagian dari *BERT* model [16], [17], [18]. Ketika data sudah semua di *embedding* maka akan di simpan kedalam database vektor [19], dan dilakukan *index* menggunakan *index* dengan nama *HNSW (Hierarchical Navigable Small World)* [20].

Llama Model

Llama adalah model *LLM* seperti GPT, Gemini, dan lain-lain yang bisa memberikan respon *generative*. Dalam penelitian ini model *LLM* yang digunakan adalah Llama3.1:8b yang merupakan model *LLM open source* yang di sediakan oleh Facebook. Respon yang dihasilkan sepenuhnya menggunakan model tersebut untuk menjawab [21].

Mean Reciprocal Rank (MRR)

Secara sederhana *MRR* mengambil potongan *chunk* relevan pertama dari daftar top-k, *MRR* yang tinggi menunjukan bahwa informasi yang relevan muncul lebih dekat ke bagian atas daftar. Rumus *MRR* sebagai berikut,

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank\ i}$$

Keterangan simbol:

- Q: Jumlah pertanyaan atau *query* yang diuji.
- $rank_i$: Peringkat dari dokumen relevan pertama untuk *query* ke-i.
- $|Q|$: Total jumlah *query* yang diuji.

Pengujian dilakukan menggunakan 10 sample data diluar dari data yang dimiliki, berikut merupakan hasil dari perhitungan evaluasi 10 sample datanya,

$$MRR = \frac{1}{10}\ (1 +\ 1 +\ 1 +\ 1 +\ 0.33 + 1 + 1 + 1 + 1 + 1) = \frac{1}{10} \times 9.33 =\ 0.933$$

Faithfullness

Faithfullness adalah menghitung seberapa akurat jawaban yang dihasilkan dengan informasi (*chunks*) yang disediakan, top-k *chunks* yang digunakan merupakan 10. Melihat keseluruhan potongan tersebut dan indentifikasi jika potongan relevan dengan jawaban akan di nilai 1 jika tidak maka 0, dan dihitung masing-masing dokumennya. Berikut rumus dari *Faithfullness*,

$$Faithfullness = \frac{|Number\ of\ claim\ generate\ answer\ that\ can\ be\ inferred\ from\ the\ given\ context|}{|Total\ number\ of\ claim\ in\ the\ generated\ answer|}$$

Kita ambil seberapa banyak dari top-k per dokumen potongan yang relevan kemudian menggunakan formula *faithfullness* untuk melihat akurasi dari dokumen tersebut.

$$Faithfullness\ (doc\ 1) = \frac{6}{10} = 0.6$$

Untuk melihat total dari keseluruhan dokumen bisa dilihat dengan rumus rata-rata (*mean*), maka didapatkan total dari keseluruan dokumen adalah,

$$Mean = \frac{(0.6 + 0.6 + 0 + 0.8 + 0.6 + 0.8 + 0.8 + 0.8 + 0.6 + 0.6)}{10} = 0.62$$

Answer Relevancy

Metrik ini menghitung relevansi jawaban yang dihasilkan dengan pertanyaan yang diajukan. Relevansi dihitung menggunakan *cosine similarity* antara vektor *embedding* dari jawaban yang dihasilkan dan pertanyaan asli.

$$Answer\ Relevancy = \frac{1}{N}\sum_{i=1}^{N}\cos(Eg, Eo)$$

Keterangan simbol:

- N: Jumlah total dokumen yang diuji.
- E_g: *Embedding* jawaban yang dihasilkan (*Generated Answer*).
- E_o: *Embedding* pertanyaan asli (*Original Question*).

Dihitung untuk masing-masing dokumen dari *cosine similarity* lalu untuk mendapat nilai akurasi akhirnya menggunakan rumus *mean*. Total akurasi keseluruhan nya menjadi,

$$Mean = \frac{(0.56 + 0.62 + 0.48 + 0.63 + 0.61 + 0.57 + 0.42 + 0.68 + 0.66 + 0.50)}{10} = 0.573$$

Semantic Similarity

metrik yang menghitung relevansi jawaban yang dihasilkan dengan jawaban yang sebenarnya. Matrik ini membandingkan kesamaan cosinus antara jawaban yang dihasilkan dengan jawaban sebenarnya, berikut rumusnya,

$$Semantic\ Similarity = cosine\ similarity\ (V_{generate}, V_{reference})$$

Keterangan Simbol:

- $V_{generate}$: Vektor *embedding* dari jawaban yang dihasilkan.
- $V_{reference}$: Vektor *embedding* dari jawaban sebenarnya.

Dihitung untuk masing-masing dokumen dari *cosine similarity* lalu untuk mendapat nilai akurasi akhirnya menggunakan rumus *mean*. Total akurasi keseluruhan nya menjadi,

$$Mean = \frac{(0.88 + 0.79 + 0.59 + 0.91 + 0.88 + 0.74 + 0.75 + 0.89 + 0.86 + 0.85)}{10} = 0.814$$

Hasil Pengujian

Hasil dari pengujian evaluasi mulai dari *Mean Reciprocal Rank (MRR), Faithfullness, Answer Relvancy, dan Semantic Similarity*. Dengan data uji berjumlah 10 dan jumlah k=10 maka hasil dari tabel evaluasinya menjadi seperti berikut,

Tabel 3.1 Evaluasi Matrik RAG

Metrik Evaluasi	Result
Mean Reciprocal Rank (MRR)	0.933
Faithfullness	0.620
Answer Relvancy	0.573
Semantic Similarity	0.814

- ***Mean Reciprocal Rank (MRR)***: Nilai *MRR* sebesar **0.933** menunjukkan bahwa dokumen yang relevan rata-rata muncul sangat dekat dengan peringkat atas dalam daftar hasil pencarian (*top-k*). Semakin tinggi nilai *MRR*, semakin baik kinerja sistem dalam mengembalikan hasil yang relevan pada posisi awal. Sistem berhasil dengan sangat baik menempatkan informasi yang relevan pada peringkat teratas, yaitu rata-rata peringkat pertama.
- ***Faithfullness***: Nilai *Faithfulness* sebesar **0.620** menunjukkan bahwa sekitar **62%** dari klaim yang dihasilkan oleh sistem dapat ditelusuri kembali ke konteks (*chunks*) yang diberikan. Sistem ini cukup andal dalam menjaga akurasi dengan data yang tersedia, tetapi memerlukan perbaikan untuk lebih meningkatkan kepercayaan pada jawaban yang dihasilkan.
- ***Answer Relevancy***: Nilai *Answer Relevancy* sebesar **0.573** menunjukkan bahwa relevansi antara jawaban yang dihasilkan dengan pertanyaan yang diajukan memiliki rata-rata skor kesamaan sekitar **57.3%** berdasarkan perhitungan *cosine similarity*. Skor ini dapat ditingkatkan dengan pengoptimalan *embedding* atau strategi pengambilan data (*retrieval*), sehingga relevansi antara pertanyaan dan jawaban semakin tinggi.
- ***Semantic Similarity***: Nilai *Semantic Similarity* sebesar **0.814** menunjukkan bahwa rata-rata kesamaan antara jawaban yang dihasilkan oleh sistem dengan jawaban sebenarnya mencapai **81.4%**. Nilai ini menunjukkan bahwa

model generatif dalam *RAG* telah bekerja secara optimal dalam memahami konteks data referensi dan menghasilkan jawaban yang hampir mirip dengan jawaban manual.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa pendekatan *Retrieval-Augmented Generation* (RAG) efektif dalam mengembangkan chatbot kesehatan pencernaan. Hasil evaluasi menunjukkan kinerja yang memuaskan, dengan nilai MRR 93%, Faithfullness 62%, Answer Relevancy 57%, dan Semantic Similarity 81%. Meskipun ada ruang untuk perbaikan dalam relevansi jawaban, dan ini merupakan hasil yang cukup bagus dalam metrik *Retrieval Augmented Generation.* Hasil ini menunjukkan potensi besar untuk pengembangan lebih lanjut, seperti integrasi teknik lanjutan dan pengoptimalan model, yang dapat memperluas penerapan *chatbot* ini ke bidang kesehatan lainnya.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada semua pihak yang telah memberikan kontribusi dalam penyelesaian penelitian ini. Terima kasih atas dukungan yang diberikan selama proses publikasi artikel ini. Apresiasi juga ditujukan kepada keluarga yang selalu memberikan semangat.

REFERENSI

[1] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," Dec. 2024, [Online]. Available: http://arxiv.org/abs/2312.10997
[2] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of Retrieval-Augmented Generation: A Survey," May 2024, [Online]. Available: http://arxiv.org/abs/2405.07437
[3] M. R. J, K. VM, H. Warrier, and Y. Gupta, "Fine Tuning LLM for Enterprise: Practical Guidelines and Recommendations," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2404.10779
[4] T. Yu, A. Xu, and R. Akkiraju, "In Defense of RAG in the Era of Long-Context Language Models," Sep. 2024, [Online]. Available: http://arxiv.org/abs/2409.01666
[5] F. Wang, X. Wan, R. Sun, J. Chen, and S. Ö. Arık, "Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.07176
[6] R. Qin *et al.*, "Robust Implementation of Retrieval-Augmented Generation on Edge-based Computing-in-Memory Architectures," May 2024, [Online]. Available: http://arxiv.org/abs/2405.04700
[7] B. Jin, J. Yoon, J. Han, and S. O. Arik, "Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.05983
[8] Z. Wang *et al.*, "Speculative RAG: Enhancing Retrieval Augmented Generation through Drafting," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.08223
[9] S. Wu *et al.*, "Retrieval-Augmented Generation for Natural Language Processing: A Survey," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.13193
[10] S. Zhao, Y. Yang, Z. Wang, Z. He, L. K. Qiu, and L. Qiu, "Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely," Sep. 2024, [Online]. Available: http://arxiv.org/abs/2409.14924
[11] S. Bin Islam, M. A. Rahman, K. S. M. T. Hossain, E. Hoque, S. Joty, and M. R. Parvez, "Open-RAG: Enhanced Retrieval-Augmented Reasoning with Open-Source Large Language Models," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.01782
[12] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," May 2021, [Online]. Available: http://arxiv.org/abs/2005.11401
[13] Z. Guo, L. Xia, Y. Yu, T. Ao, and C. Huang, "LightRAG: Simple and Fast Retrieval-Augmented Generation," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.05779
[14] Z. Feng, X. Feng, D. Zhao, M. Yang, and B. Qin, "Retrieval-Generation Synergy Augmented Large Language Models," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.05149
[15] A. Azis, A. T. Zy, and A. S. Sunge, "Prediksi Penjualan Obat Dan Alat Kesehatan Terlaris Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 6, no. 1, pp. 117–124, Jan. 2024, doi: 10.47233/jteksis.v6i1.1078.
[16] M. R. Douglas, "Large Language Models," Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.05782
[17] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," May 2020, [Online]. Available: http://arxiv.org/abs/2005.14165
[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805
[19] Y. Han, C. Liu, and P. Wang, "A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.11703

[20] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs," 2016.
[21] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.13971