Vol. 4 No. 2 – Tahun 2025
Indonesian Journal Computer Science
ISSN: 2829-3819

Analisa *Prompt Engineering* pada *Large Language Model* dengan *Retrieval-Augmented Generation* untuk Informasi Obat dan Vitamin

Imam Haromain[1], Sirojul Munir[2], Amalia Rahmah[3]
Sekolah Tinggi Teknologi Terpadu Nurul Fikri[1,2,3]
haromain@nurulfikri.ac.id[1], rojulman@nurulfikri.ac.id[2], amaliarahmah2@gmail.com[3]

Diterima	Direvisi	Disetujui
(19-10-2025)	(10-10-2025)	(24-10-2025)

Abstrak - Penelitian ini bertujuan melakukan analisa efektivitas dua gaya *prompt*, yaitu *prompt* bebas (*zero-shot*) dan *prompt* terbimbing (*few-shot*), pada model *Large Language Model* (LLM) berbasis *Retrieval-Augmented Generation* (RAG) dengan topik spesifik kesehatan, yaitu obat dan vitamin. Metode penelitian yang digunakan adalah eksperimen dengan menguji respons model terhadap sepuluh pertanyaan yang dirancang berdasarkan dokumen PDF dari sumber terpercaya, seperti Kementerian Kesehatan dan WHO. Proses ini bertujuan untuk mengevaluasi sejauh mana model mampu memberikan jawaban yang relevan, akurat, serta sesuai konteks ketika diberi perbedaan gaya *prompt*. Evaluasi kualitas jawaban dilakukan menggunakan dua metrik populer dalam *Natural Language Processing*, yaitu BERTScore untuk menilai kesesuaian semantik, dan ROUGE untuk mengukur kesesuaian tekstual. Hasil penelitian menunjukkan bahwa *prompt* bebas menghasilkan skor BERTScore yang cukup baik (Precision 69,74%, Recall 70,97%, F1 70,30%), namun cenderung rendah pada ROUGE. Sebaliknya, *prompt* terbimbing menunjukkan peningkatan kinerja, baik pada BERTScore (Precision 70,23%, Recall 73,32%, F1 71,64%) maupun ROUGE. Hasil penelitian menunjukan, penggunaan *prompt* terbimbing lebih efektif dalam menjaga keseimbangan antara kesesuaian semantik dan tekstual, sehingga berpotensi mendukung pengembangan sistem informasi kesehatan berbasis LLM secara lebih andal dan praktis.

Kata Kunci : LLM, Obat dan Vitamin, *Prompt Engineering*, RAG, Rouge dan BERTScore

Abstract *- This study aims to analyze the effectiveness of two prompt styles—free prompt (zero-shot) and guided prompt (few-shot)—in a Retrieval-Augmented Generation (RAG)-based Large Language Model (LLM) within the specific healthcare domain of medicines and vitamins. The research method employed is experimental, where the model's responses are tested against ten questions constructed from PDF documents sourced from reputable institutions, such as the Ministry of Health and the World Health Organization (WHO). The purpose of this process is to evaluate how well the model can provide relevant, accurate, and contextually appropriate answers under different prompting strategies. The quality of responses is assessed using two widely recognized metrics in Natural Language Processing: BERTScore, to evaluate semantic similarity, and ROUGE, to measure textual overlap. The findings indicate that free prompts achieve reasonably good performance on BERTScore (Precision 69.74%, Recall 70.97%, F1 70.30%) but tend to perform lower on ROUGE. In contrast, guided prompts demonstrate improved results in both BERTScore (Precision 70.23%, Recall 73.32%, F1 71.64%) and ROUGE. These results suggest that guided prompting is more effective in balancing semantic and textual consistency, thereby supporting the development of more reliable and practical LLM-based healthcare information systems.*

Keywords: LLM, Medicines and Vitamins, Prompt Engineering, RAG, ROUGE and BERTScore

I. PENDAHULUAN

Perkembangan terkini dalam teknologi kecerdasan buatan (AI) berdampak pada berbagai sektor, termasuk bidang kesehatan (Wang et al., 2023). Salah satu kemajuan yang paling besar dan signifikan adalah munculnya model bahasa besar (*Large Language Models*, LLM). Model LLM adalah model pemrosesan bahasa alami (*Natural Language Processing*) tingkat lanjut yang menganalisis masukan teks dan menghasilkan keluaran yang sesuai konteks (Shah et al., 2024). Model LLM juga dapat didefinisikan sebagai model komputasi yang memiliki kemampuan untuk memahami dan menghasilkan teks yang sangat mirip dengan teks buatan manusia (Chang et al., 2023). Model

ini memiliki potensi besar untuk mendukung sistem informasi kesehatan termasuk diantaranya diagnosis berbasis teks, informasi rekomendasi pengobatan, menjawab pertanyaan terkait kesehatan dan informasi kesehatan lainnya (Shah et al., 2024). Namun, salah satu tantangan terbesar dalam penerapan model LLM dalam bidang kesehatan adalah memanfaatkan potensi ini semaksimal mungkin. Salah satu teknik yang dapat digunakan untuk meningkatkan akurasi dan relevansi respons model LLM adalah teknik *prompt engineering*.

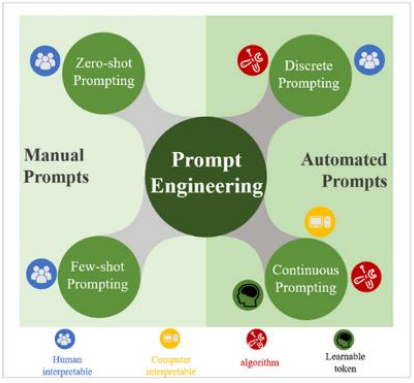
Prompt engineering merupakan bidang studi yang relatif baru yang merujuk pada praktik merancang, menyempurnakan, dan menerapkan perintah atau instruksi yang memandu keluaran dari model LLM dan membantu berbagai tugas dan memberikan informasi (Meskó, 2023). *Prompt* dapat dibagi menjadi dua jenis yaitu *manual-prompt* dan *automated-prompt*. *Manual-prompt* dibuat dengan bantuan manusia untuk memberikan instruksi eksplisit pada model LLM tentang jenis data apa yang harus difokuskan dan bagaimana melakukan pendekatan terhadap perintah atau tugas yang diberikan (Liu et al., 2021). Sedangkan *automated-prompt* adalah teknik prompt yang dihasilkan menggunakan berbagai algoritma dan teknik dengan tidak membutuhkan intervensi manusia (Lester et al., 2021). Beberapa metode diantaranya *Zero-shot Prompting*, *Few-shot Prompting*, *Discrete Prompting* dan *Continuous Prompting* (Wang et al., 2023) seperti terlihat pada Gambar 1. Penggunaan *prompt* memungkinkan model LLM berbasis AI untuk menjalankan berbagai perintah atau tugas AI salah satunya pada medis atau kesehatan.

Salah satu penerapan *prompt* dalam bidang kesehatan adalah untuk memberikan jawaban atas pertanyaan kesehatan (Singhal et al., 2023). *Prompt* bidang kesehatan akan mengajukan pertanyaan untuk dipahami dan ditanggapi secara akurat oleh model berdasarkan pengetahuan yang dimiliki pada bidang kesehatan atau medis (Wang et al., 2023). Pertanyaan kesehatan diantaranya mencakup pendukung keputusan medis seperti diagnosis pengobatan, meningkatkan komunikasi antara penyedia layanan kesehatan dan pasien seperti pengingat pengobatan, dan penjadwalan konsultasi, dan untuk kesehatan masyarakat dimana dalam skala yang lebih besar, dapat membantu dalam inisiatif kesehatan masyarakat dengan membantu menganalisis data kesehatan populasi, memprediksi tren penyakit, atau melakukan edukasi kesehatan kepada masyarakat (Meskó, 2023). Masalah utama yang dihadapi adalah bagaimana mengembangkan metode *prompt engineering* yang efektif dan efisien dalam konteks informasi kesehatan dengan beberapa metode yang ada. Selain itu, bagaimana metode ini dapat diterapkan dengan berbagai jenis sistem informasi kesehatan online, mulai dari konsultasi virtual hingga pertanyaan terkait topik kesehatan, medis maupun obat. Sebagai contoh, pengujian dapat dilakukan pada topik spesifik seperti informasi vitamin dan obat, misalnya terkait indikasi, dosis, interaksi, efek samping, maupun panduan penggunaan yang aman.

Namun model LLM terkadang memberikan hasil informasi jawaban pertanyaan tidak akurat, sehingga diperlukan penambahan informasi dari sumber eksternal yang dapat memberikan solusi untuk meningkatkan kinerja model (Rizky, 2025). Salah satu pendekatannya adalah dengan *Retrieval-Augmented Generation* (RAG). RAG adalah sebuah arsitektur yang memadukan mekanisme pencarian informasi dengan kemampuan menghasilkan teks (Albert & Voutama, 2025). RAG dapat juga didefinisikan sebagai teknik AI yang menggabungkan pencarian informasi dari sumber eksternal dengan kemampuan bahasa model untuk menghasilkan jawaban yang lebih akurat, relevan, dan terkini. Dengan implementasi RAG dapat membantu adanya kesenjangan informasi dengan mengkombinasikan pengetahuan dari informasi eksternal (Rizky, 2025). Meskipun RAG punya potensi besar, kualitas teks yang dihasilkan sangat dipengaruhi oleh *prompt* yang diberikan di awal. *Prompt* berfungsi memberi arahan pada model tentang bagaimana harus menjawab. Bahkan perubahan kecil pada *prompt* dapat mengubah struktur, relevansi, dan ketepatan jawaban . Secara sederhana, *prompt* membantu menetapkan konteks percakapan, menunjukkan informasi penting, serta menentukan bentuk dan isi keluaran yang diharapkan (Rizky, 2025).

Beberapa penelitian sebelumnya telah membahas penerapan LLM di berbagai bidang. Penerapan model LLM di bidang kesehatan telah mulai diterapkan dalam beberapa tahun terakhir. Penelitian sebelumnya yang dilakukan oleh Wang J, et al menyebutkan bahwa salah satu penerapan model LLM untuk kesehatan adalah untuk menjawab pertanyaan terkait kesehatan atau medis (Wang et al., 2023). Penelitian Holmes J, et al menggunakan model LLM untuk mendapatkan informasi mengenai topik radiasi (Holmes et al., 2023). Penelitian

Lee S, et al menggunakan model LLM untuk mendapatkan jawaban topik penyakit diabetes (Lee et al., 2023). Penelitian Dongyeop Jang dan Chang-Eop Kim menggunakan model LLM untuk mendapatkan jawaban mengenai obat tradisional korea (Dongyeop Jang & Chang-Eop Kim, 2023). Sebagai pembaruan dari penelitian sebelumnya, peneliti bermaksud untuk menggunakan dataset yang berbeda dan lebih spesifik yakni terkait obat dan vitamin. Penelitian ini akan menerapkan berbagai teknik *prompt* yang beragam untuk mengoptimalkan kinerja model. Hasil dari pengolahan model tersebut direncanakan untuk diaplikasikan dalam sistem informasi kesehatan, khususnya pada fitur chatbot kesehatan yang berfungsi dalam distribusi obat dan vitamin. Hal ini diharapkan dapat meningkatkan efisiensi dan akurasi sistem dalam memberikan informasi terkait obat dan vitamin kepada pengguna.



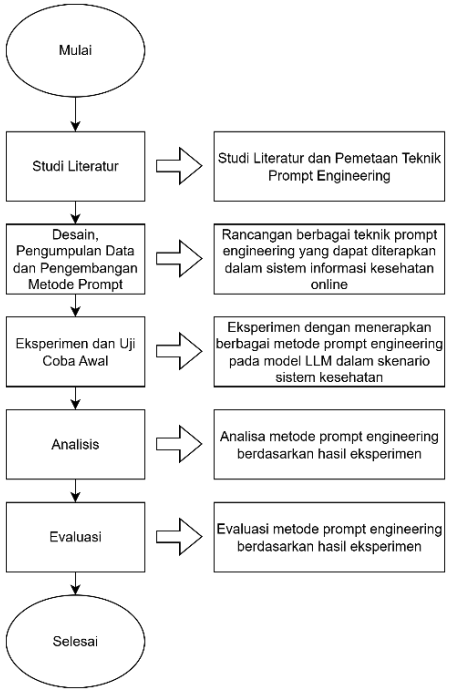
Sumber: (Wang et al., 2023)

Gambar 1. *Prompt engineering*

Penelitian ini bertujuan untuk melakukan evaluasi dan analisis gaya penulisan *prompt* yaitu dengan *zero shot prompt* dan *few shot prompt* dengan model *Large Language Model* (LLM) berbasis *Retrieval-Augmented Generation* (RAG) dengan topik spesifik kesehatan yaitu obat dan vitamin. *Zero-shot prompt* dapat dikategorikan sebagai *prompt* bebas, karena merupakan suatu teknik di mana model AI, seperti *Large Language Model* (LLM), diminta untuk menyelesaikan suatu tugas atau menjawab pertanyaan tanpa diberikan contoh spesifik maupun pelatihan khusus terkait tugas tersebut (Kuka, 2025). Sedangkan *few-shot prompt* dikategorikan sebagai *prompt* terbimbing karena model diberikan beberapa contoh yang berfungsi sebagai arahan atau panduan dalam menghasilkan jawaban yang sesuai. Eksperimen akan dilakukan dalam penelitian yaitu dengan menilai kualitas teks jawaban dengan gaya tersebut yang akan dievaluasi dengan metrik evaluasi ROUGE dan BERTScore. Diharapkan hasil penelitian ini dapat menjadi acuan dalam merancang prompt yang baik dan efektif guna meningkatkan kualitas informasi yang dihasilkan oleh LLM, serta dapat diintegrasikan ke dalam sistem informasi kesehatan, misalnya dalam bentuk chatbot.

II. METODOLOGI PENELITIAN

Metode penelitian yang digunakan adalah metode eksperimen. Tahapan penelitian yang dilakukan dapat dilihat pada Gambar 2.

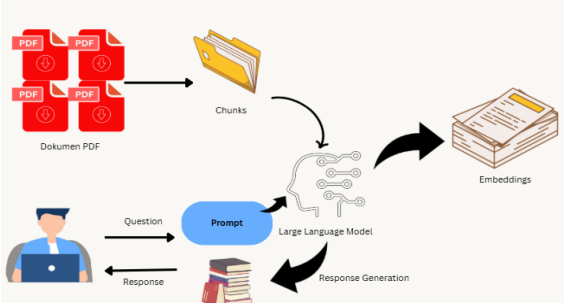


Sumber: Penelitian (2025)

Gambar 2. Tahapan Penelitian

1. Studi Literatur

Langkah pertama adalah melakukan studi literatur untuk memahami teori, konsep, serta penelitian sebelumnya yang relevan dengan *prompt engineering* dan penerapannya dalam konteks sistem informasi kesehatan. Kajian ini memberikan pemahaman mendalam tentang berbagai teknik *prompt engineering* serta penerapan *Large Language Model* (LLM), yang menjadi landasan penting bagi proses penelitian selanjutnya. Melalui studi literatur, peneliti dapat memetakan pendekatan yang sudah ada dan memilih metode yang paling sesuai untuk diujikan. Pada tahap ini peneliti memilih 2 jenis metode prompt yang akan digunakan dan dievaluasi yaitu *prompt* bebas (*zero shot prompt*) dan *prompt* terbimbing (*few shot prompt*).



Sumber: Penelitian (2025)

Gambar 3. Diagram alur kerja *Retrieval-Augmented Generation* (RAG).

2. Desain, Pengumpulan data dan pengembangan Metode *Prompt*

Setelah memperoleh dasar teori, tahap berikutnya adalah merancang dan mengembangkan metode *prompt engineering*. Untuk diagram alur kerja RAG dapat dilihat pada Gambar 3. Pada tahap ini, peneliti menyusun desain eksperimen, mengumpulkan data pendukung, serta mengimplementasikan kedua variasi prompt sebelumnya yang dapat diterapkan dalam topik kesehatan berbasis LLM. Dataset menggunakan beberapa dokumen PDF dengan bahasa indonesia yang peneliti ambil dari sumber ahli khususnya ahli kesehatan seperti dari kemenkes, ikatan dokter dan beberapa orang ahli dibidang kesehatan. Dokumen PDF yang dipilih juga spesifik yaitu mengenai topik obat, vitamin dan pencegahan penyakit dengan mencakup berbagai jenis penulisan, isi dan kompleksitas file. Beberapa metode *prompt* yang dirancang mengacu pada referensi penelitian sebelumnya, termasuk teknik berbasis pertanyaan langsung maupun instruksi khusus untuk menghasilkan *prompt* yang lebih akurat. Pertanyaan dibuat menyesuaikan dengan jenis *prompt* yang digunakan yaitu *prompt* bebas dan terbimbing dengan masing-masing 10 pertanyaan dengan topik spesifik kesehatan yaitu mengenai obat dan vitamin.

3. Eksperimen

Tahapan berikutnya adalah eksperimen, yaitu dengan menerapkan metode *prompt* engineering yang telah dirancang pada model LLM. Eksperimen dilakukan menggunakan AI *language models* populer dari Meta yaitu Llama dengan model versi llama-3.1-8b-instant. Eksperimen ini juga menggunakan beberapa alat bantuan yaitu dengan streamlit dan beberapa *library* file python dan dibuat skenario yang menyerupai sistem informasi kesehatan. Tujuannya adalah mengamati bagaimana model merespons variasi *prompt* yang diberikan serta mengevaluasi kelayakan penerapan metode tersebut dalam konteks kesehatan yaitu obat dan vitamin.

4. Analisis & Evaluasi

Hasil dari eksperimen kemudian dianalisis untuk menilai efektivitas tiap metode *prompt engineering*. Analisis difokuskan pada kriteria tertentu, seperti relevansi respons, akurasi dalam memberikan informasi terkait obat maupun vitamin, serta kemampuan memberikan rekomendasi jawaban yang dapat diterima. Melalui analisis ini, kelebihan dan kekurangan masing-masing metode dapat diidentifikasi, sehingga menjadi dasar dalam menentukan metode yang paling sesuai untuk diterapkan. Pada tahap evaluasi, metode yang telah dianalisis diuji kembali menggunakan metrik evaluasi otomatis, seperti ROUGE dan BERTScore, guna memberikan gambaran objektif mengenai kinerja tiap pendekatan.

a. BERTScore adalah metrik evaluasi otomatis yang digunakan untuk mengevaluasi hasil generasi teks. Metrik ini menghitung skor kesamaan untuk setiap token dalam kalimat yang dibuat dengan token dalam kalimat referensi. Cara kerjanya adalah dengan menghitung kesamaan kosinus antara representasi token (*embedding*) dari dua kalimat tersebut untuk menilai sejauh mana keduanya sama (Tri Utami Br. Lubis, 2024). BERTScore menghitung *precision*, *recall*, dan *F1-score* melalui *cosine similarity* antar embedding kontekstual yang dinormalisasi. Skor berada pada rentang 0–1, di mana nilai lebih tinggi menunjukkan kesesuaian semantik yang lebih baik antara teks kandidat dan referensi (Qaulan et al., 2025). Berikut ini persamaan untuk menghitung *Precision*, *Recall* dan *F1-Score* :

- *Precision*

  Precision dalam BERTScore dihitung dengan mencocokkan setiap token pada jawaban sistem (kandidat) dengan token pada teks referensi menggunakan *pairwise cosine similarity*. Semakin tinggi nilai *precision*, semakin banyak informasi dalam kandidat yang sesuai dengan konten referensi. Rumus perhitungan *precision* dapat dilihat pada Persamaan 1.

$$P_{BERT} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} Max\, x_i{}^T \hat{x}_j \quad (1)$$

- *Recall*
  Recall dihitung dengan mencocokkan token-token pada kalimat referensi terhadap token-token pada kalimat kandidat menggunakan metode *pairwise cosine similarity* yang dapat dilihat pada persamaan 2.

$$R_{BERT} = \frac{1}{|x|} \sum_{\hat{x}_i \in x} Max\, x_i{}^T \hat{x}_j \quad (2)$$

- *F1-Score*
  F1-score adalah metrik yang menggabungkan nilai *precision* dan *recall*. Rumus perhitungan *F1-score* ditampilkan pada persamaan 3.

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

b. ROUGE Score
ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) merupakan seperangkat metrik yang dirancang untuk mengevaluasi kualitas ringkasan otomatis maupun tugas generasi teks dengan membandingkan kesesuaian n-gram, urutan kata, atau pasangan kata antara teks hasil sistem dan teks referensi manusia (Rizky, 2025). ROUGE-N digunakan untuk menilai kecocokan n-gram dengan variasi nilai n, seperti unigram (n=1) dan bigram (n=2). Selain itu, ROUGE-L mengukur kesamaan berdasarkan *Longest Common Subsequence* (LCS), yaitu urutan kata terpanjang yang sama pada kedua teks. Perhitungan ROUGE umumnya mencakup *precision*, *recall*, dan *F1-score*, dengan formulasi matematis sebagai berikut (Rizky, 2025) :

- *Precision*
  Precision digunakan untuk mengevaluasi tingkat relevansi prediksi dengan menghitung kesesuaian kata (unigram, bigram, atau LCS) yang ditemukan, lalu membaginya dengan jumlah total kata pada ringkasan sistem yang dapat dilihat pada persamaan 4,5 dan 6.

$$ROUGE\ 1\ Precision = \frac{Jumlah\ unigram\ kata\ sama}{Keseluruhan\ kata\ ringkasan\ sistem} (4)$$

$$ROUGE\ 2\ Precision = \frac{Jumlah\ bigram\ kata\ sama}{Keseluruhan\ kata\ ringkasan\ sistem} (5)$$

$$ROUGE\ L\ Precision = \frac{Longest\ Common\ Subsequence\ (LCS)}{Keseluruhan\ kata\ ringkasan\ sistem} (6)$$

- *Recall*
  Recall dihitung sebagai rasio antara jumlah kata yang cocok, baik dalam bentuk unigram, bigram, maupun LCS, terhadap keseluruhan kata yang terdapat pada ringkasan manual yang dapat dilihat pada persamaan 7,8 dan 9.

$$ROUGE\ 1\ Recall = \frac{Jumlah\ unigram\ kata\ sama}{Keseluruhan\ kata\ ringkasan\ teks\ manusia} (7)$$

$$ROUGE\ 2\ Recall = \frac{Jumlah\ bigram\ kata\ sama}{Keseluruhan\ kata\ ringkasan\ teks\ manusia} (8)$$

$$ROUGE\ L\ Recall = \frac{Longest\ Common\ Subsequence\ (LCS)}{Keseluruhan\ kata\ ringkasan\ teks\ manusia} (9)$$

- *F1-Score*
  F1-score, atau *F-measure*, merupakan metrik yang menghitung rata-rata harmonik antara nilai recall dan precision yang dapat dilihat pada persamaan 10.

$$F_1 = 2\,.\frac{Precision\,.\,Recall}{Precision + Recall} (10)$$

III. HASIL DAN PEMBAHASAN

Eksperimen dilakukan dengan membandingkan dua pendekatan prompt engineering pada model *Retrieval-Augmented Generation* (RAG), yaitu *prompt* bebas dan *prompt* terbimbing. Data uji pada penelitian ini berasal dari dokumen PDF yang dikumpulkan dari berbagai sumber media dan berasal dari sumber terpercaya pada sektor kesehatan yaitu WHO, kemenkes maupun ahli di bidang kesehatan. Dokumen PDF berupa artikel maupun *e-book* dan digunakan sebagai acuan dalam menjawab pernyataan sekaliguse referensi dalam melakukan proses evaluasi menggunakan metrik BERTScore dan ROUGE.

Tabel 1. Pertanyaan *Prompt*

No	Pertanyaan	
	Prompt Bebas	*Prompt* Terbimbing
1	Apa definisi dari anemia?	Apa definisi medis anemia sesuai standar kesehatan?
2	Mengapa remaja putri lebih rentan dan berisiko terkena anemia?	Jelaskan faktor biologis yang memengaruhi risiko anemia pada remaja putri, termasuk kaitannya dengan kebutuhan zat besi, menstruasi, dan perubahan fisiologis pada masa pertumbuhan?

3	Bagaimana pencegahan dan penanggulangan anemia di sekolah?	Bagaimana upaya pencegahan dan penanggulangan anemia di sekolah, termasuk melalui program gizi, edukasi kesehatan, serta pemberian Tablet Tambah Darah (TTD)?
4	Apa fungsi vitamin D?	Apa fungsi vitamin D, khususnya dalam menjaga kesehatan tulang dan metabolisme kalsium serta perannya dalam mendukung sistem kekebalan tubuh?
5	Apa yang dimaksud dengan defisiensi vitamin D?	Apa yang dimaksud dengan defisiensi vitamin D, termasuk definisinya dalam konteks medis, penyebab umum terjadinya, serta dampaknya terhadap kesehatan tulang dan sistem tubuh lainnya?
6	Apa saja penggolongan jenis obat?	Apa saja penggolongan obat berdasarkan cara perolehannya?
7	Bagaimana cara menyimpan obat?	Bagaimana cara menyimpan obat dengan benar agar tetap aman dan gunakan bahasa sederhana yang mudah dipahami masyarakat umum?
8	Apa fungsi dari vitamin A?	Apa fungsi vitamin A, khususnya untuk kesehatan mata dan penglihatan, serta perannya dalam pertumbuhan, sistem kekebalan tubuh, dan kesehatan kulit?
9	Apa saja sumber untuk mendapatkan vitamin A?	Apa saja sumber vitamin A dari produk hewani dan nabati, serta apa perbedaan antara vitamin A aktif (retinol) dan provitamin A (beta-karoten)?
10	Apa dampak kekurangan vitamin A pada anak?	Apa dampak kekurangan vitamin A pada anak, khususnya gejala utama berupa gangguan penglihatan?

Sumber: Penelitian (2025)

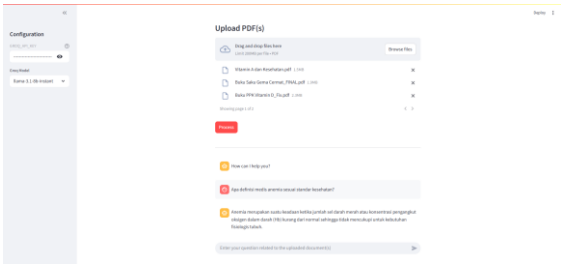
Pada tabel 1 adalah contoh pertanyaan yang dirancang untuk mengukur efektivitas kedua prompt tersebut, untuk topik yang digunakan khusus topik kesehatan yaitu topik obat dan vitamin.

Langkah selanjutnya adalah merancang antarmuka pengujian dengan menggunakan kedua jenis *prompt* yang telah disiapkan. Rancangan ini dibangun menggunakan teknologi Streamlit beserta beberapa library Python, serta memanfaatkan model bahasa besar dari Meta, yaitu Llama versi llama-3.1-8b-instant. Model ini tidak hanya digunakan untuk menghasilkan jawaban, tetapi juga diintegrasikan dalam kerangka *Retrieval-Augmented Generation* (RAG) sehingga mampu mengambil informasi relevan dari dokumen referensi sebelum menghasilkan respons. Dengan demikian, sistem dapat memberikan jawaban yang lebih akurat, kontekstual, dan sesuai dengan informasi yang terkandung dalam sumber data. Gambar 4, 5 dan 6 terlihat hasil rancangan yang dibuat.

1. Pengujian

Langkah pertama yang dilakukan dalam proses pengujian adalah menyiapkan antarmuka aplikasi yang digunakan sebagai media untuk melakukan evaluasi. Melalui antarmuka ini, peneliti dapat mengunggah dokumen uji serta mengatur jenis *prompt* yang akan digunakan. Pada tahap awal, peneliti mengunggah lima file PDF yang berisi materi mengenai vitamin dan obat. Dokumen-dokumen ini dipilih sebagai sumber pengetahuan yang akan digunakan sistem dalam menjawab pertanyaan yang diberikan. Setelah dokumen berhasil diunggah, langkah berikutnya adalah menyusun sepuluh pertanyaan yang akan diajukan kepada sistem.

Pertanyaan tersebut dirancang untuk diuji dengan dua pendekatan berbeda, yaitu *prompt* bebas dan *prompt* terbimbing. *Prompt* bebas memberikan kebebasan sistem dalam merumuskan jawaban sesuai pemahamannya terhadap dokumen, sedangkan *prompt* terbimbing diarahkan dengan instruksi yang lebih spesifik agar jawaban lebih terfokus.



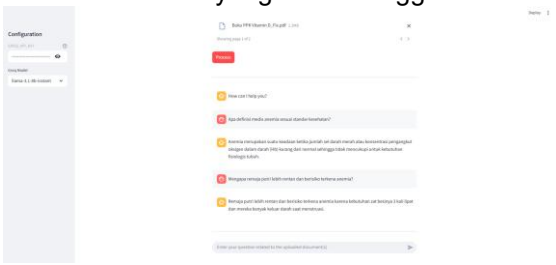
Sumber: Penelitian (2025)

Gambar 4. Tampilan PDF Upload

2. Evaluasi

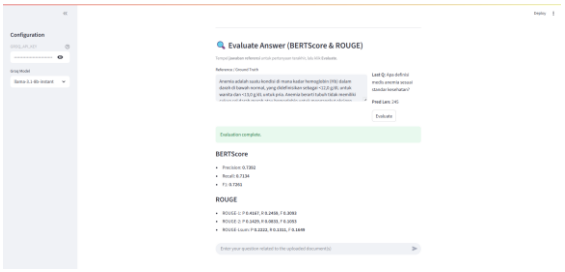
Setelah semua pertanyaan diberikan, sistem menghasilkan jawaban untuk masing-masing skenario. Hasil jawaban kemudian dievaluasi menggunakan dua metrik utama, yaitu BERTScore dan ROUGE. Kedua metrik ini dipilih karena mampu menilai kesesuaian antara jawaban sistem dengan referensi yang telah ditetapkan, baik dari sisi kemiripan semantik (BERTScore) maupun dari sisi kesesuaian kata dan frasa (ROUGE). Selanjutnya, skor yang diperoleh dari setiap gaya penulisan pada masing-masing metrik dibandingkan untuk

memperoleh gambaran menyeluruh mengenai kualitas respons. Hasil evaluasi ini disajikan dalam bentuk tabel, yaitu Tabel 2 dan Tabel 3, yang menunjukkan perbandingan kinerja sistem pada kedua jenis *prompt*. Melalui tahapan ini, peneliti dapat menganalisis secara mendalam sejauh mana sistem mampu memberikan jawaban yang relevan dan sesuai dengan konteks dokumen yang telah diunggah.



Sumber: Penelitian (2025)

Gambar 5. Tampilan Pertanyaan dan Respon Jawaban



Sumber: Penelitian (2025)

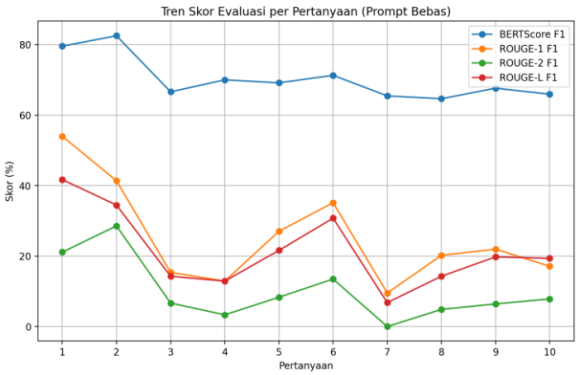
Gambar 6. Tampilan Evaluasi Respon Pertanyaan

Tabel 2. Evaluasi *Prompt* Bebas

No	Bert Score			ROUGE								
				ROUGE-1			ROUGE-2			ROUGE-L		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	77,64%	81,63%	79,58%	42,72%	73,33%	53,99%	16,67%	28,81%	21,12%	33,01%	56,67%	41,72%
2	85,31%	79,93%	82,53%	54,55%	33,33%	41,38%	38,10%	22,86%	28,57%	45,45%	27,78%	34,48%
3	63,94%	69,54%	66,62%	9,66%	37,84%	15,38%	4,17%	16,67%	6,67%	8,97%	35,14%	14,29%
4	67,38%	72,91%	70,03%	10,00%	18,18%	12,90%	2,56%	4,76%	3,33%	10,00%	18,18%	12,90%
5	71,66%	66,86%	69,18%	28,57%	25,64%	27,03%	8,82%	7,89%	8,33%	22,86%	20,51%	21,62%
6	71,04%	71,59%	71,31%	27,59%	48,48%	35,16%	10,53%	18,75%	13,48%	24,14%	42,42%	30,77%
7	65,06%	65,89%	65,47%	7,95%	11,86%	9,52%	0,00%	0,00%	0,00%	5,68%	8,47%	6,80%
8	64,35%	64,96%	64,66%	23,73%	21,21%	22,40%	5,17%	4,62%	4,88%	15,25%	13,64%	14,40%
9	66,63%	68,72%	67,66%	20,41%	23,81%	21,98%	6,25%	7,32%	6,74%	18,37%	21,43%	19,78%
10	64,35%	67,62%	65,94%	23,08%	32,81%	27,10%	6,67%	9,52%	7,84%	16,48%	23,44%	19,35%
Rerata	69,74%	70,97%	70,30%	24,83%	32,65%	26,68%	9,89%	12,12%	10,10%	20,02%	26,77%	21,61%

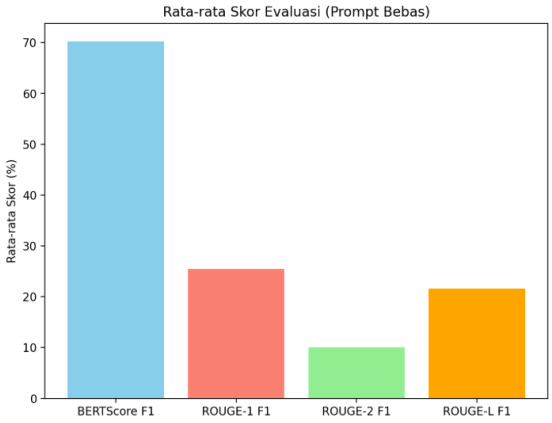
Sumber: Penelitian (2025)

Berdasarkan hasil evaluasi pengujian dengan *prompt* bebas yang ditampilkan pada Tabel 2 dan Gambar 7 serta 8, terlihat bahwa nilai BERTScore menunjukkan performa yang relatif baik dengan rata-rata *Precision* 69,74%, *Recall* 70,97%, dan F1 70,30%, yang mengindikasikan bahwa sistem mampu menghasilkan respons dengan kesesuaian semantik cukup tinggi terhadap referensi. Namun, pada metrik ROUGE, performa sistem cenderung lebih rendah. Rata-rata skor ROUGE-1 berada pada F1 sebesar 26,68%, sedangkan ROUGE-2 hanya mencapai 10,10%, dan ROUGE-L sebesar 21,61%. Hasil ini menunjukkan bahwa meskipun secara makna jawaban sistem cukup mendekati referensi (terlihat dari nilai BERTScore yang stabil di atas 70%), tetapi dari segi kesesuaian n-gram atau susunan kata yang diukur oleh ROUGE, kualitas jawaban masih rendah. Secara teoritis, hal ini dapat dipahami karena *prompt* bebas memberikan keleluasaan lebih besar kepada model dalam menyusun jawaban, sehingga sistem cenderung menghasilkan variasi kalimat yang berbeda dari referensi meskipun maknanya serupa. Dengan demikian, dapat disimpulkan bahwa pada penggunaan *prompt* bebas, sistem lebih kuat dalam mempertahankan kesesuaian semantik, tetapi kurang konsisten dalam reproduksi bentuk tekstual yang sesuai dengan jawaban acuan.



Sumber: Penelitian (2025)

Gambar 7. Evaluasi Pertanyaan (*Prompt* Bebas)

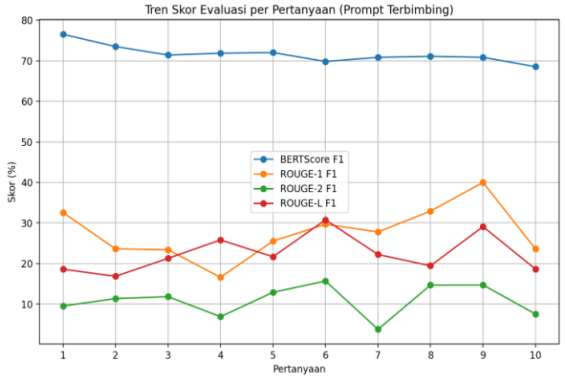


Sumber: Penelitian (2025)

Gambar 8. Nilai skor evaluasi (*Prompt* Bebas)

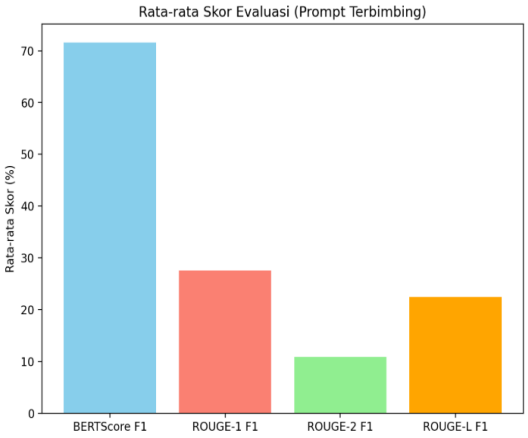
Kemudian berdasarkan hasil evaluasi pengujian dengan *prompt* terbimbing yang ditampilkan pada Tabel 3 dan Gambar 9 serta 10, nilai BERTScore menunjukkan rata-rata yang cukup stabil dengan *Precision* 70,23%, *Recall* 73,32%, dan F1 71,64%. Angka ini sedikit lebih tinggi dibandingkan dengan hasil pada *prompt* bebas, yang mengindikasikan bahwa arahan tambahan dalam bentuk *prompt* terbimbing membantu sistem menghasilkan jawaban dengan kesesuaian semantik yang lebih konsisten terhadap referensi. Hal ini sejalan dengan teori bahwa *prompt* yang lebih terarah dapat mengurangi variasi kalimat dan mendorong model untuk lebih fokus pada informasi yang relevan dengan pertanyaan, sehingga makna jawaban lebih dekat dengan jawaban acuan.

Kemudian jika ditinjau dari metrik ROUGE, hasil yang diperoleh juga menunjukkan tren peningkatan dibandingkan *prompt* bebas. Rata-rata skor ROUGE-1 F1 mencapai 28,47%, sedangkan ROUGE-2 F1 sebesar 11,72%, dan ROUGE-L F1 sebesar 22,57%. Nilai ini masih relatif rendah, tetapi lebih baik dibandingkan capaian pada *prompt* bebas. Hal ini menegaskan bahwa meskipun model dengan *prompt* terbimbing masih menghasilkan variasi kalimat, struktur kata dan kesesuaian n-gram dengan referensi menjadi lebih terjaga. Dengan demikian, dapat disimpulkan bahwa penggunaan *prompt* terbimbing memberikan keuntungan dalam menjaga keseimbangan antara kesesuaian semantik (BERTScore) dan kesesuaian tekstual (ROUGE), sehingga respons yang dihasilkan lebih relevan baik dari sisi makna maupun bentuk.



Sumber: Penelitian (2025)

Gambar 9. Evaluasi Pertanyaan (*Prompt* Terbimbing)



Sumber: Penelitian (2025)

Gambar 10. Nilai skor evaluasi (*Prompt* Terbimbing)

Tabel 3. Evaluasi *Prompt* Terbimbing

No	Bert Score			ROUGE								
				ROUGE-1			ROUGE-2			ROUGE-L		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	80,98%	72,49%	76,50%	53,85%	23,33%	32,56%	16,00%	6,78%	9,52%	30,77%	13,33%	18,60%
2	71,77%	75,32%	73,50%	14,79%	58,33%	23,60%	7,09%	28,57%	11,36%	10,56%	41,67%	16,85%
3	66,79%	76,69%	71,40%	14,57%	59,46%	23,40%	7,33%	30,56%	11,83%	13,25%	54,05%	21,28%
4	66,51%	78,12%	71,85%	9,29%	77,27%	16,59%	3,85%	33,33%	6,90%	8,74%	72,73%	15,61%
5	68,49%	75,90%	72,01%	16,11%	61,54%	25,53%	8,11%	31,58%	12,90%	13,42%	51,28%	21,28%
6	69,91%	69,72%	69,81%	23,94%	51,52%	32,69%	11,43%	25,00%	15,69%	22,54%	48,48%	30,77%
7	71,30%	70,37%	70,83%	30,61%	25,42%	27,78%	4,17%	3,45%	3,77%	24,49%	20,34%	22,22%
8	69,54%	72,69%	71,08%	30,25%	54,55%	38,92%	17,80%	32,31%	22,95%	24,37%	43,94%	31,35%
9	68,37%	73,50%	70,84%	26,83%	78,57%	40,00%	9,84%	29,27%	14,72%	19,51%	57,14%	29,09%
10	68,63%	68,42%	68,53%	19,59%	29,69%	23,60%	6,25%	9,52%	7,55%	15,46%	23,44%	18,63%
Rerata	70,23%	73,32%	71,64%	23,98%	51,97%	28,47%	9,19%	23,04%	11,72%	18,31%	42,64%	22,57%

Sumber: Penelitian (2025)

IV. KESIMPULAN

Hasil penelitian ini menegaskan bahwa penerapan teknik *prompt engineering* pada model *Large Language Model* (LLM) berbasis *Retrieval-Augmented Generation* (RAG) mampu meningkatkan kualitas jawaban dalam konteks informasi kesehatan, khususnya terkait obat dan vitamin. Melalui eksperimen dengan dua pendekatan, yaitu *prompt* bebas (*zero-shot*) dan *prompt* terbimbing (*few-shot*), diperoleh gambaran bahwa masing-masing memiliki keunggulan tersendiri. *Prompt* bebas cenderung menghasilkan respons dengan kesesuaian semantik yang tinggi, sebagaimana tercermin pada nilai BERTScore yang stabil, namun kurang konsisten dalam kesesuaian tekstual yang diukur dengan ROUGE. Sebaliknya, *prompt* terbimbing terbukti lebih mampu menjaga keseimbangan antara kesesuaian makna dan struktur kata, meskipun hasil ROUGE masih relatif rendah.

Hasil evaluasi ini menunjukkan bahwa desain *prompt* sangat berpengaruh terhadap kualitas keluaran model, di mana penambahan arahan yang lebih spesifik dapat membantu meningkatkan relevansi jawaban. Temuan ini mendukung teori bahwa *prompt* engineering tidak hanya menjadi instrumen teknis, tetapi juga bagian penting dari strategi penerapan LLM dalam domain seperti kesehatan. Dengan demikian, penelitian ini memberikan kontribusi praktis dalam pengembangan sistem informasi kesehatan berbasis AI, terutama untuk aplikasi chatbot yang berfungsi memberikan edukasi, rekomendasi, dan informasi terkait obat serta vitamin. Sebagai langkah pengembangan, penelitian ini membuka peluang untuk mengintegrasikan metode *prompt* terbimbing dengan strategi peningkatan kualitas lain, seperti penyaringan dataset, pemilihan model yang lebih adaptif, atau penggabungan teknik pencarian informasi yang lebih canggih.

UCAPAN TERIMA KASIH

Penulis menyampaikan apresiasi dan rasa terima kasih kepada Direktorat Riset, Teknologi, dan Pengabdian kepada Masyarakat, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi atas dukungan pendanaan hibah penelitian tahun 2025. Ucapan terima kasih juga ditujukan kepada Sekolah Tinggi Teknologi Terpadu Nurul Fikri yang telah memberikan dukungan berupa fasilitas, sarana, dan prasarana sehingga penelitian ini dapat terlaksana dengan baik. Selain itu, penulis juga menghargai kontribusi dari berbagai pihak yang telah membantu, baik secara langsung maupun tidak langsung, dalam mendukung kelancaran dan penyelesaian penelitian ini.

V. REFERENSI

Albert, G. D., & Voutama, A. (2025). Pengembangan Chatbot Berbasis PDF Menggunakan Local Retrieval-Augmented Generation (RAG) Dan Ollama. Jurnal Informatika Dan Teknik Elektro Terapan, 13(2). https://doi.org/10.23960/jitet.v13i2.6361

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2023). A Survey on Evaluation of Large Language Models. J. ACM, 37, 1–45. https://doi.org/https://doi.org/10.48550/arXiv.2307.03109

Dongyeop Jang, & Chang-Eop Kim. (2023). Exploring the Potential of Large Language models in Traditional Korean Medicine: A Foundation Model Approach to Culturally-Adapted Healthcare. ArXivLabs: Experimental Projects with Community Collaborators.

Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T. T., McGee, L. A., Ashman, J. B., Li, X., Liu,

T., Shen, J., & Liu, W. (2023). Evaluating Large Language Models on a Highly-specialized Topic, Radiation Oncology Physics. https://doi.org/10.3389/fonc.2023.1219326

Kuka, V. (2025, March 6). Technique #3: Examples in prompts: From zero-shot to few-shot. Learn Prompting. https://learnprompting.org/docs/basics/few_shot?utm_source=chatgpt.com

Lee, S., Lee, D. Y., Im, S., Kim, N. H., & Park, S.-M. (2023). Clinical Decision Transformer: Intended Treatment Recommendation through Goal Prompting. http://arxiv.org/abs/2302.00612

Lester, B., Al-Rfou, R., & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. Empirical Methods in Natural Language Processing. https://doi.org/https://doi.org/10.48550/arXiv.2104.08691

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. 1–46. https://doi.org/https://doi.org/10.48550/arXiv.2107.13586

Meskó, B. (2023). Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. Journal of Medical Internet Research, 25, e50638. https://doi.org/10.2196/50638

Qaulan, M. A., Wahyuni, & Adytia, P. (2025). Pengembangan Chatbot Berbasis AI untuk Mendukung Pelayanan Perpustakaan . Tematik : Jurnal Teknologi Informasi Komunikasi (e-Journal), 12(1), 23–30.

Rizky, M. A. (2025). Analisis Efektivitas Dua Jenis Gaya Prompt dalam Model LLM Berbasis RAG. Jurnal Komtika (Komputasi dan Informatika), 9(1), 76–86. https://doi.org/10.31603/komtika.v9i1.13488

Shah, K., Xu, A. Y., Sharma, Y., Daher, M., McDonald, C., Diebo, B. G., & Daniels, A. H. (2024). Large Language Model Prompting Techniques for Advancement in Clinical Medicine. Journal of Clinical Medicine, 13(17), 5101. https://doi.org/10.3390/jcm13175101

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Lubis, T. U. B. (2024). Question answering system menggunakan large language models (LLM) dan LangChain (Studi kasus: UU Kesehatan). [Skripsi, Universitas Islam Negeri Sultan Syarif Kasim Riau].

Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., Yang, Q., Kang, Y., Wu, J., Hu, H., Yue, C., Zhang, H., Liu, Y., Pan, Y., Liu, Z., Sun, L., Li, X., Ge, B., Jiang, X., … Zhang, S. (2023). Prompt Engineering for Healthcare: Methodologies and Applications. JOURNAL OF LATEX CLASS FILES, 14, 1–18. https://doi.org/https://doi.org/10.48550/arXiv.2304.146