BAB 2
LANDASAN TEORI

Bab ini memaparkan teori-teori yang menjadi landasan ilmiah penelitian, meliputi konsep *Large Language Model* (LLM), *Retrieval-Augmented Generation* (RAG), model Mistral-7B, platform Ollama, basis data vektor ChromaDB, *framework* evaluasi RAGAS, dataset yang digunakan, serta penelitian terkait.

2.1 Large Language Model

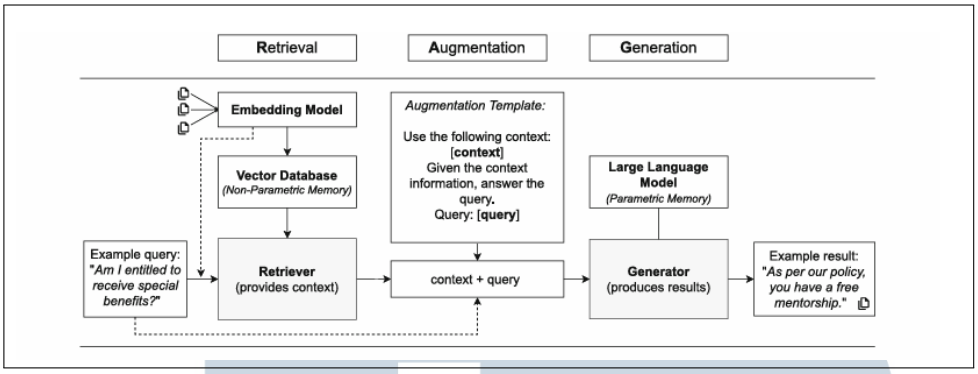
Large Language Model (LLM) adalah model kecerdasan buatan berbasis arsitektur *transformer* yang dilatih menggunakan data teks berskala sangat besar, sehingga mampu memahami dan menghasilkan teks dalam bahasa manusia dengan tingkat koherensi yang tinggi [2]. LLM memiliki miliaran parameter yang memungkinkannya menangkap pola bahasa yang kompleks, melakukan penalaran kontekstual, dan menghasilkan respons yang relevan terhadap berbagai jenis pertanyaan [14].

Dalam penerapannya pada domain kesehatan dan kebugaran, LLM mampu memberikan panduan latihan fisik, rekomendasi nutrisi, serta dukungan motivasi secara percakapan (*conversational*) [4, 26]. Namun, LLM generik memiliki keterbatasan mendasar yaitu fenomena halusinasi, yakni kondisi di mana model menghasilkan informasi yang terdengar meyakinkan namun tidak akurat secara faktual [11, 12, 27]. Keterbatasan ini menjadi perhatian serius dalam domain yang membutuhkan akurasi tinggi seperti kesehatan dan kebugaran [10].

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) adalah paradigma yang menggabungkan mekanisme pengambilan informasi (*retrieval*) dari basis pengetahuan eksternal dengan kemampuan generatif LLM [28, 14]. Arsitektur RAG terdiri dari tiga komponen utama yaitu *retriever* yang bertugas mencari potongan dokumen yang relevan, *generator* berupa LLM yang menghasilkan jawaban, dan basis data vektor sebagai penyimpan representasi numerik dokumen [16].

Alur kerja RAG sebagaimana ditunjukkan pada Gambar 2.1 dimulai dari proses *indexing*, di mana dokumen sumber dipotong menjadi potongan-potongan



Gambar 2.1. Arsitektur sistem RAG *chatbot* gym dan nutrisi berbasis Mistral-7B

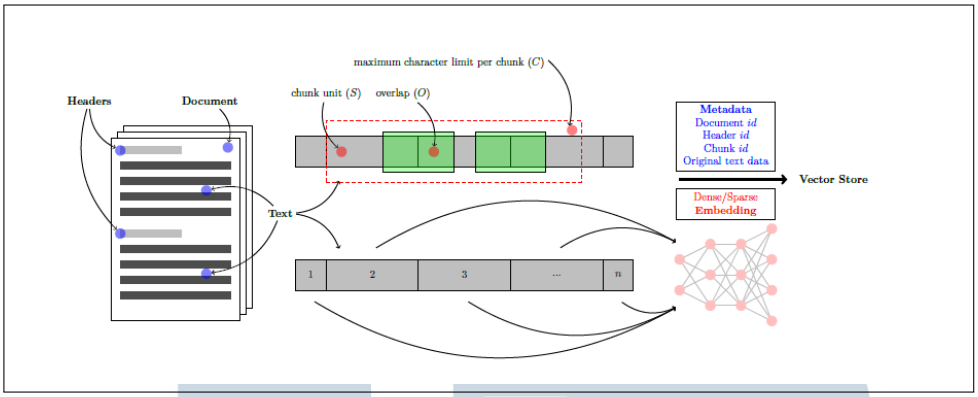
kecil (*chunks*), dikonversi menjadi vektor numerik melalui model *embedding*, dan disimpan dalam basis data vektor. Saat pengguna mengajukan pertanyaan, sistem mengubah pertanyaan tersebut menjadi vektor, kemudian mencari potongan dokumen yang paling relevan berdasarkan kemiripan vektor. Potongan-potongan relevan tersebut kemudian digabungkan dengan pertanyaan asli dan dikirimkan ke LLM sebagai konteks untuk menghasilkan jawaban [11].

Penerapan RAG pada *chatbot* domain spesifik seperti *gym* dan nutrisi terbukti secara signifikan mengurangi halusinasi karena model dikondisikan untuk hanya menjawab berdasarkan dokumen yang telah terverifikasi [14]. Pendekatan ini juga memberikan transparansi sumber informasi yang dapat ditelusuri oleh pengguna [11].

2.2.1 Chunking

Chunking adalah proses memecah dokumen teks menjadi potongan-potongan yang lebih kecil sebelum disimpan ke dalam basis data vektor [18]. Proses ini merupakan tahap krusial dalam *pipeline* RAG karena ukuran dan konfigurasi potongan secara langsung mempengaruhi kualitas pengambilan informasi dan relevansi konteks yang diberikan kepada LLM [20, 22].

Sebagaimana ditunjukkan pada Gambar 2.2, parameter utama dalam proses *chunking* meliputi *chunk size* yang menentukan jumlah karakter atau token dalam setiap potongan, dan *chunk overlap* yang menentukan jumlah karakter yang tumpang tindih antara dua potongan yang berurutan [18]. Potongan yang terlalu kecil dapat menyebabkan hilangnya konteks penting, sementara potongan yang terlalu besar dapat menurunkan presisi pencarian [19]. Penggunaan *chunk overlap* bertujuan untuk menjaga kontinuitas informasi di batas antar potongan sehingga



Gambar 2.2. Ilustrasi proses *chunking* dokumen dengan *chunk size* dan *chunk overlap*

tidak ada informasi penting yang terpotong [20].

2.2.2 Embedding

Model *embedding* berfungsi mengubah teks menjadi representasi numerik berupa vektor berdimensi tinggi yang menangkap makna semantik dari teks tersebut [29]. Dalam penelitian ini digunakan model *Language-Agnostic BERT Sentence Embedding* (LaBSE) dari *sentence-transformers* yang dikembangkan oleh Feng et al. [29] dan dipublikasikan di *Annual Meeting of the Association for Computational Linguistics* (ACL) 2022. Model ini menghasilkan vektor berukuran 768 dimensi dan dilatih menggunakan data dari 109 bahasa termasuk bahasa Indonesia, sehingga mampu menangkap kemiripan semantik lintas bahasa dengan akurasi tinggi [29].

Model LaBSE dipilih dalam penelitian ini karena kemampuannya memproses teks berbahasa Indonesia secara lebih akurat dibandingkan model *monolingual* yang dioptimalkan khusus untuk bahasa Inggris. Mengingat pertanyaan evaluasi dan jawaban sistem dalam penelitian ini menggunakan bahasa Indonesia, penggunaan model *multilingual* seperti LaBSE memberikan representasi vektor yang lebih tepat sehingga meningkatkan akurasi penghitungan kemiripan semantik pada metrik evaluasi RAGAS, khususnya *response relevancy* [29].

2.2.3 Top-K Retrieval

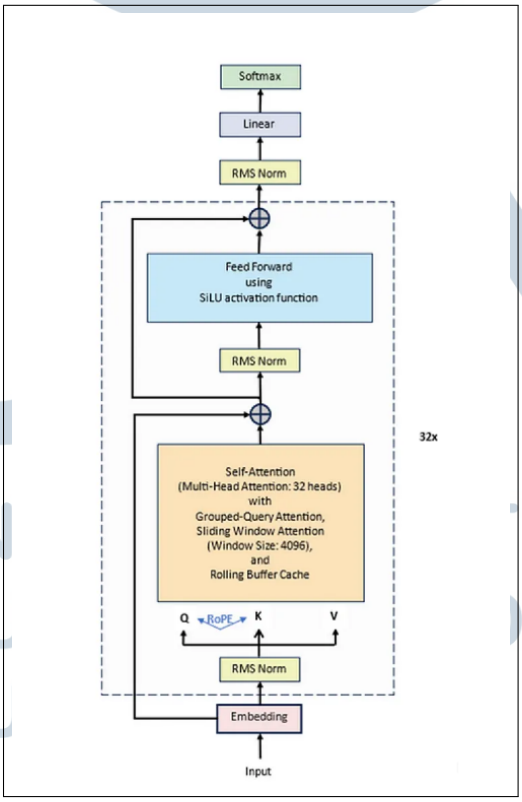
Top-k retrieval adalah parameter yang menentukan jumlah potongan dokumen yang diambil dari basis data vektor untuk dijadikan konteks bagi LLM [20]. Nilai *top-k* yang terlalu kecil dapat menyebabkan konteks yang tidak lengkap, sehingga LLM kekurangan informasi untuk menjawab pertanyaan secara

komprehensif [19]. Sebaliknya, nilai *top-k* yang terlalu besar dapat memasukkan potongan yang tidak relevan ke dalam konteks yang justru dapat menurunkan kualitas jawaban [20, 23].

Lyu et al. [20] menunjukkan bahwa nilai *top-k* mempengaruhi kekayaan dan keragaman konteks yang diberikan kepada LLM — nilai yang terlalu kecil menyebabkan konteks tidak lengkap, sementara nilai yang terlalu besar memasukkan potongan tidak relevan.Penelitian ini menguji tiga nilai *top-k* yaitu 3, 5, dan 10 untuk mengidentifikasi nilai yang menghasilkan performa terbaik pada domain *gym* dan nutrisi.

2.3 Mistral-7B

Mistral-7B adalah model LLM *open-source* dengan 7 miliar parameter yang dikembangkan oleh Mistral AI [14]. Model ini menggunakan arsitektur *transformer* dengan sejumlah inovasi teknis seperti *grouped-query attention* (GQA) yang mempercepat proses inferensi, dan *sliding window attention* (SWA) yang memungkinkan pemrosesan konteks yang lebih panjang secara efisien [24].



Gambar 2.3. Arsitektur model Mistral-7B

Sumber: Diadaptasi dari Jiang et al. [30]

Sebagaimana ditunjukkan pada Gambar 2.3, arsitektur Mistral-7B terdiri dari tiga inovasi utama dibandingkan *transformer* standar. Pertama, *Sliding Window Attention* (SWA) membatasi setiap token agar hanya memperhatikan sejumlah token terdekat dalam jendela tertentu, sehingga mengurangi kompleksitas komputasi secara signifikan [30]. Kedua, *Grouped Query Attention* (GQA) mengelompokkan beberapa *query head* untuk berbagi satu *key* dan *value head*, sehingga mempercepat proses inferensi dan mengurangi kebutuhan memori [30]. Ketiga, *Rolling Buffer KV Cache* menjaga ukuran cache tetap konstan selama inferensi dengan mengganti nilai lama secara bergulir, sehingga memungkinkan pemrosesan teks panjang tanpa peningkatan memori yang signifikan [30].

Meskipun berukuran relatif kecil dibandingkan model komersial sepertiGPT-4, Mistral-7B terbukti memiliki performa yang kompetitif dalam berbagai *benchmark* NLP dan bahkan mengungguli model yang lebih besar seperti Llama-2 13B pada beberapa tugas [14]. Karakteristik ini menjadikan Mistral-7B pilihan yang tepat untuk penelitian yang mengutamakan efisiensi komputasi dan kemampuan dijalankan secara lokal tanpa memerlukan infrastruktur *cloud* yang mahal [24].

Dalam konteks sistem RAG, Mistral-7B berperan sebagai komponen *generator* yang menerima konteks dari *retriever* dan menghasilkan jawaban berbasis bahasa alami. Kemampuan model dalam mengikuti instruksi (*instruction following*) melalui versi *instruct*-nya membuatnya sangat sesuai untuk tugas *chatbot* yang membutuhkan respons yang terfokus dan sesuai domain [14].

2.4 Ollama

Ollama adalah platform *open-source* yang memungkinkan pengguna untuk menjalankan berbagai model LLM secara lokal di perangkat pribadi [14]. Platform ini menyederhanakan proses instalasi, manajemen, dan inferensi model LLM melalui antarmuka baris perintah yang intuitif serta API berbasis HTTP yang dapat diintegrasikan dengan berbagai bahasa pemrograman termasuk Python [14].

Ollama menyediakan lapisan abstraksi yang memungkinkan model LLM seperti Mistral-7B untuk berjalan secara lokal tanpa memerlukan koneksi internet atau biaya penggunaan API *cloud*. Platform ini menjalankan *server* LLM lokal pada port 11434 yang dapat diakses melalui protokol HTTP standar, sehingga memudahkan integrasi dengan *framework* seperti LangChain [18]. Penggunaan Ollama dalam penelitian ini mendukung aspek privasi data pengguna karena seluruh

pemrosesan dilakukan secara lokal tanpa pengiriman data ke server eksternal [24].

2.5 ChromaDB

ChromaDB adalah basis data vektor *open-source* yang dirancang khusus untuk menyimpan dan melakukan pencarian efisien pada vektor *embedding* [16]. Basis data vektor berbeda dari basis data relasional konvensional karena mengoptimalkan operasi pencarian kemiripan (*similarity search*) antar vektor berdimensi tinggi menggunakan algoritma *Approximate Nearest Neighbor* (ANN) [14].

Dalam *pipeline* RAG, ChromaDB berperan sebagai komponen *vector store* yang menyimpan representasi vektor dari seluruh potongan dokumen *knowledge base*. Ketika sistem menerima pertanyaan, ChromaDB melakukan pencarian *cosine similarity* antara vektor pertanyaan dengan seluruh vektor dokumen yang tersimpan, kemudian mengembalikan sejumlah *top-k* potongan dengan kemiripan tertinggi [16]. ChromaDB dipilih dalam penelitian ini karena kemudahannya dalam integrasi dengan *framework* LangChain serta kemampuannya untuk menyimpan data secara persisten di penyimpanan lokal [24].

2.6 LangChain

LangChain adalah *framework* pengembangan aplikasi berbasis LLM yang menyediakan abstraksi dan komponen modular untuk membangun sistem seperti *chatbot*, agen, dan *pipeline* RAG [16]. *Framework* ini menyederhanakan proses penghubungan berbagai komponen sistem RAG seperti *document loader*, *text splitter*, model *embedding*, basis data vektor, dan LLM menjadi sebuah *pipeline* yang terintegrasi [14].

LangChain menyediakan antarmuka yang konsisten untuk berbagai penyedia LLM dan basis data vektor, sehingga memudahkan penggantian komponen tanpa perubahan kode yang signifikan. Dalam penelitian ini, LangChain digunakan untuk mengintegrasikan Mistral-7B via Ollama, ChromaDB sebagai *vector store*, dan model *embedding* dari HuggingFace ke dalam satu *pipeline* RAG yang kohesif [16].

2.7 Framework RAGAS

RAGAS (*Retrieval-Augmented Generation Assessment*) adalah *framework* evaluasi *open-source* yang dirancang khusus untuk mengukur performa sistem RAG secara otomatis [21]. RAGAS menggunakan LLM sebagai *evaluator* untuk menilai kualitas jawaban yang dihasilkan sistem RAG berdasarkan metrik yang telah terdefinisi dengan baik [14].

Keunggulan utama RAGAS dibandingkan metrik evaluasi tradisional seperti ROUGE atau BLEU adalah kemampuannya dalam menilai kualitas semantik jawaban, bukan sekadar kemiripan leksikal dengan jawaban referensi [16]. RAGAS versi terbaru memiliki lebih dari dua puluh metrik evaluasi, namun tidak semua metrik dapat diterapkan pada setiap jenis sistem RAG. Es et al. [21] dalam paper resmi RAGAS yang dipublikasikan di *European Chapter of the Association for Computational Linguistics* (EACL) 2024 menyatakan bahwa untuk evaluasi *reference-free* yaitu evaluasi tanpa memerlukan jawaban referensi yang telah dianotasi secara manual tiga metrik utama yang direkomendasikan adalah *faithfulness*, *answer relevancy*, dan *context precision*.

Penelitian ini mengadopsi ketiga metrik tersebut karena dataset evaluasi yang digunakan tidak memiliki jawaban referensi yang telah dianotasi (*ground truth*), mengingat belum tersedia *benchmark* standar untuk domain *gym* dan nutrisi makanan. Metrik lain seperti *context recall* dan *answer correctness* membutuhkan *ground truth* jawaban yang benar sehingga tidak dapat diterapkan. Selain itu, metrik *multimodal faithfulness* dan *multimodal relevance* dirancang untuk sistem RAG yang memproses gambar atau video, sehingga tidak relevan untuk penelitian ini yang hanya menggunakan dokumen teks. Pendekatan evaluasi dengan tiga metrik *reference-free* ini juga digunakan oleh penelitian RAG *chatbot* di bidang kesehatan yang diterbitkan di JMIR AI [17].

2.7.1 Faithfulness

Faithfulness mengukur seberapa konsisten jawaban yang dihasilkan sistem terhadap konteks dokumen yang diambil oleh *retriever* [21]. Metrik ini menilai apakah setiap klaim dalam jawaban dapat didukung secara faktual oleh informasi yang terdapat dalam potongan dokumen yang diberikan sebagai konteks [14].

Proses penghitungan *faithfulness* dilakukan dengan memecah jawaban menjadi pernyataan-pernyataan atomik (*atomic statements*), kemudian setiap

pernyataan dinilai apakah dapat dibuktikan dari konteks yang tersedia. Penilaian ini menghasilkan sebuah *verdict* untuk setiap pernyataan, yaitu *Yes* (bernilai 1) apabila pernyataan didukung oleh konteks, atau *No* (bernilai 0) apabila pernyataan tidak dapat dibuktikan dari konteks yang tersedia, disertai alasan (*reason*) yang menjelaskan dasar penilaian tersebut.

Sebagai ilustrasi, misalkan sistem menghasilkan jawaban terhadap pertanyaan "Berapa menit aktivitas fisik intensitas sedang yang direkomendasikan per minggu?" berupa: *"Minimal 150 menit intensitas sedang per minggu. Bisa berupa jalan cepat atau bersepeda. 75 menit intensitas tinggi sebagai alternatif."* Jawaban tersebut dipecah menjadi tiga pernyataan atomik dan dinilai terhadap konteks dokumen yang di-*retrieve*, sebagaimana ditunjukkan pada Tabel 2.1.

Tabel 2.1. Contoh proses penilaian *faithfulness*

#	**Pernyataan Atomik**	**Verdict**	**Alasan**
S1	Minimal 150 menit intensitas sedang per minggu	Yes (1)	Terdapat dalam konteks dokumen
S2	Bisa berupa jalan cepat atau bersepeda	No (0)	Tidak terdapat dalam konteks dokumen
S3	75 menit intensitas tinggi sebagai alternatif	Yes (1)	Terdapat dalam konteks dokumen

Berdasarkan Tabel 2.1, dari tiga pernyataan atomik yang dihasilkan, dua pernyataan (S1 dan S3) memperoleh *verdict Yes* karena dapat dibuktikan langsung dari konteks dokumen, sedangkan satu pernyataan (S2) memperoleh *verdict No* karena informasi tersebut tidak disebutkan secara eksplisit dalam konteks yang diberikan, meskipun secara umum benar dan relevan dengan topik. Skor *faithfulness* kemudian dihitung menggunakan rumus sebagai berikut [21]:

$$Faithfulness = \frac{|S_{supported}|}{|S_{total}|} \tag{2.1}$$

di mana $|S_{supported}|$ adalah jumlah pernyataan atomik dalam jawaban yang dapat dibuktikan dari konteks dokumen, dan $|S_{total}|$ adalah total jumlah pernyataan atomik dalam jawaban. Berdasarkan contoh pada Tabel 2.1, skor *faithfulness* yang dihasilkan adalah $2/3 = 0.6667$. Nilai *faithfulness* berada pada rentang 0 hingga

1, di mana nilai lebih tinggi menunjukkan konsistensi yang lebih baik dan tingkat halusinasi yang lebih rendah [21].

2.7.2 Response Relevancy

Response relevancy mengukur seberapa relevan jawaban yang dihasilkan terhadap pertanyaan yang diajukan pengguna [21]. Metrik ini menilai apakah jawaban benar-benar menjawab inti pertanyaan atau justru mengandung informasi yang tidak diminta, bertele-tele, atau menyimpang dari topik [16].

Penghitungan *response relevancy* dilakukan dengan menggunakan LLM untuk menghasilkan sejumlah pertanyaan hipotetis (*generated questions*) dari jawaban yang diberikan, kemudian menghitung rata-rata *cosine similarity* antara pertanyaan hipotetis tersebut dengan pertanyaan asli. Prinsipnya, apabila jawaban benar-benar relevan dan tepat sasaran, pertanyaan hipotetis yang dihasilkan dari jawaban tersebut akan sangat mirip dengan pertanyaan asli yang diajukan pengguna.

Sebagai ilustrasi, misalkan pengguna mengajukan pertanyaan asli "Berapa gram protein per kg berat badan untuk atlet kekuatan?", dan sistem menjawab "1,6–1,7 gram protein per kg berat badan untuk atlet kekuatan." LLM kemudian menghasilkan tiga pertanyaan hipotetis dari jawaban tersebut beserta nilai *cosine similarity*-nya terhadap pertanyaan asli, sebagaimana ditunjukkan pada Tabel 2.2.

Tabel 2.2. Contoh proses penilaian *response relevancy*

#	**Pertanyaan Hipotetis (Dihasilkan LLM)**	**Cosine Similarity**
Q1	Berapa kebutuhan protein untuk atlet kekuatan?	0.95
Q2	Berapa gram protein per kg untuk atlet?	0.90
Q3	Apa rekomendasi asupan protein harian?	0.78

Berdasarkan Tabel 2.2, ketiga pertanyaan hipotetis yang dihasilkan LLM memiliki kemiripan tinggi dengan pertanyaan asli, mengindikasikan bahwa jawaban sistem cukup fokus dan relevan terhadap inti pertanyaan pengguna. Skor *response relevancy* dihitung menggunakan rumus berikut [21]:

$$ResponseRelevancy = \frac{1}{N}\sum_{i=1}^{N}\cos(E_q, E_{q_i}) \tag{2.2}$$

di mana E_q adalah vektor *embedding* pertanyaan asli, E_{q_i} adalah vektor *embedding* pertanyaan hipotetis ke-i yang di-*generate* dari jawaban, dan N adalah jumlah pertanyaan hipotetis yang dihasilkan. Berdasarkan contoh pada Tabel 2.2, skor *response relevancy* yang dihasilkan adalah rata-rata dari ketiga nilai *cosine similarity* tersebut, yaitu $(0.95 + 0.90 + 0.78)/3 = 0.8767$. Nilai *response relevancy* yang tinggi menunjukkan bahwa jawaban secara tepat menjawab apa yang ditanyakan tanpa informasi yang tidak perlu [21].

2.7.3 Context Precision

Context precision mengukur seberapa tepat peringkat potongan dokumen yang diambil oleh *retriever* [21]. Metrik ini berbeda dari *precision* konvensional dalam *information retrieval*. *Precision* konvensional hanya mengukur proporsi dokumen relevan dari seluruh dokumen yang diambil, sedangkan *context precision* dalam RAGAS juga mempertimbangkan *urutan peringkat* dari potongan yang relevan — potongan yang relevan harus berada di posisi teratas agar LLM mendapatkan konteks terbaik [21].

Sebagai ilustrasi, misalkan sistem menggunakan *top-k* = 3 untuk pertanyaan "Berapa menit aktivitas fisik intensitas sedang yang direkomendasikan per minggu?", dan *retriever* mengembalikan tiga potongan dokumen dengan urutan sebagaimana ditunjukkan pada Tabel 2.3.

Tabel 2.3. Contoh proses penilaian *context precision*

Posisi	Ringkasan Potongan Dokumen	Relevan?
1	Rekomendasi WHO tentang durasi aktivitas fisik mingguan	Ya (1)
2	Klasifikasi intensitas latihan menurut ACSM	Ya (1)
3	Panduan porsi makan berdasarkan Pedoman Gizi Seimbang	Tidak (0)

Berdasarkan Tabel 2.3, potongan dokumen pada posisi 1 dan 2 relevan dengan pertanyaan, sedangkan potongan pada posisi 3 tidak relevan. Skor *context precision* dihitung menggunakan rumus berikut [21]:

$$ContextPrecision@K = \frac{\sum_{k=1}^{K} (Precision@k \times rel_k)}{\sum_{k=1}^{K} rel_k} \tag{2.3}$$

di mana K adalah nilai *top-k* yang digunakan, $Precision@k$ adalah presisi pada posisi ke-k dalam daftar hasil, dan rel_k adalah indikator relevansi potongan pada posisi ke-k (bernilai 1 jika relevan dan 0 jika tidak relevan). Berdasarkan contoh pada Tabel 2.3, perhitungan dilakukan sebagai berikut: pada posisi 1, $Precision@1 = 1/1 = 1.0$ (satu dari satu potongan relevan); pada posisi 2, $Precision@2 = 2/2 = 1.0$ (dua dari dua potongan relevan); posisi 3 tidak dihitung karena $rel_3 = 0$. Skor *context precision* yang dihasilkan adalah $(1.0 \times 1 + 1.0 \times 1)/(1+1) = 1.0$, menunjukkan bahwa kedua potongan relevan berhasil ditempatkan pada posisi teratas oleh *retriever*. Nilai *context precision* yang tinggi menunjukkan bahwa *retriever* berhasil menempatkan potongan-potongan paling berguna di urutan teratas sehingga LLM mendapatkan konteks berkualitas tinggi [21].

2.7.4 Rata-Rata Skor RAGAS

Untuk mendapatkan satu nilai representatif yang mencerminkan performa keseluruhan sistem pada setiap konfigurasi eksperimen, dihitung rata-rata skor RAGAS yang merupakan nilai rata-rata aritmatik dari ketiga metrik yang digunakan, yaitu *faithfulness*, *response relevancy*, dan *context precision*. Skor rata-rata RAGAS dihitung menggunakan rumus berikut:

$$RAGAS_{avg} = \frac{Faithfulness + ResponseRelevancy + ContextPrecision}{3} \tag{2.4}$$

Sebagai ilustrasi, konfigurasi eksperimen dengan *chunk size* 512 token, *chunk overlap* 100 token, dan *top-k* 10 (Eksperimen 11) memperoleh skor *faithfulness* sebesar 0.8422, *response relevancy* sebesar 0.6924, dan *context precision* sebesar 1.0000. Berdasarkan rumus di atas, rata-rata skor RAGAS untuk konfigurasi tersebut adalah:

$$RAGAS_{avg} = \frac{0.8422 + 0.6924 + 1.0000}{3} = 0.8449$$

Nilai rata-rata RAGAS ini digunakan sebagai dasar utama untuk membandingkan performa antar konfigurasi eksperimen dan menentukan konfigurasi optimal, sebagaimana dibahas lebih lanjut pada Bab 4.

2.8 Dataset Penelitian

Seluruh dokumen diunduh dari situs resmi masing-masing organisasi penerbit dalam format PDF dan tersedia secara terbuka untuk keperluan penelitian akademik, kemudian dikonversi ke format teks untuk memudahkan proses *chunking* dan *embedding*.

Penelitian ini menggunakan sepuluh dokumen sebagai *knowledge base* sistem RAG, terdiri dari tiga dokumen domain latihan fisik, lima dokumen domain nutrisi makanan, dan dua dokumen domain gabungan latihan fisik dan nutrisi. Tabel 2.4 merangkum informasi keseluruhan dataset yang digunakan.

Tabel 2.4. Dataset *knowledge base* sistem RAG

No	Nama Dataset	Penerbit	Tahun	Domain
1	WHO Global Action Plan on Physical Activity (GAPPA)	WHO	2018	Latihan fisik
2	ACSM Guidelines for Exercise Testing and Prescription (Ed. 11)	ACSM	2021	Latihan fisik
3	Physical Activity Guidelines for Americans (2nd Ed.)	HHS/ODPHP	2018	Latihan fisik
4	ACSM Position Stand: Nutrition and Athletic Performance	ACSM, AND, DC	2016	Latihan fisik dan nutrisi
5	Dietary Guidelines for Americans 2020–2025	USDA & HHS	2020	Nutrisi makanan
6	USDA FoodData	USDA	2020	Nutrisi makanan
7	Pedoman Gizi Seimbang	Kemenkes RI	2014	Nutrisi makanan
8	Dietary Guidance to Improve Cardiovascular Health	AHA	2021	Nutrisi makanan
9	WHO Healthy Diet Guidelines	WHO	2020	Nutrisi makanan
10	Panduan Umum Gym dan Nutrisi	Diolah oleh peneliti	2025	Kebugaran dan nutrisi

Untuk memperjelas cakupan domain masing-masing dokumen, berikut rincian pengelompokan sepuluh dokumen berdasarkan domain:

1. **Domain latihan fisik (3 dokumen)**: WHO Global Action Plan on Physical Activity (GAPPA), ACSM Guidelines for Exercise Testing and Prescription, dan Physical Activity Guidelines for Americans.

2. **Domain nutrisi makanan (5 dokumen)**: Dietary Guidelines for Americans 2020–2025, USDA FoodData, Pedoman Gizi Seimbang, Dietary Guidance to Improve Cardiovascular Health, dan WHO Healthy Diet Guidelines.

3. **Domain gabungan latihan fisik dan nutrisi (2 dokumen)**: ACSM Position Stand: Nutrition and Athletic Performance, yang secara spesifik membahas kebutuhan nutrisi untuk mendukung performa atletik (menjembatani domain latihan fisik dan nutrisi); serta Panduan Umum Gym dan Nutrisi yang disusun oleh peneliti berdasarkan sintesis dari kesembilan dokumen resmi lainnya, untuk melengkapi cakupan pertanyaan yang membutuhkan kombinasi informasi latihan fisik dan nutrisi secara bersamaan.

2.8.1 WHO Global Action Plan on Physical Activity 2018–2030

WHO Global Action Plan on Physical Activity (GAPPA) 2018–2030 adalah panduan aktivitas fisik global yang diterbitkan oleh *World Health Organization* (WHO) [3]. Dokumen ini menetapkan rekomendasi aktivitas fisik berbasis bukti untuk berbagai kelompok usia, mendefinisikan klasifikasi intensitas latihan, serta menjabarkan manfaat aktivitas fisik bagi kesehatan kardiovaskular, metabolisme, dan kesehatan mental.

2.8.2 ACSM Guidelines for Exercise Testing and Prescription

ACSM's Guidelines for Exercise Testing and Prescription edisi ke-11 adalah panduan latihan fisik komprehensif yang diterbitkan oleh *American College of Sports Medicine* (ACSM) [10]. Dokumen ini mencakup prinsip-prinsip latihan berbasis FITT-VP (*Frequency, Intensity, Time, Type, Volume, Progression*), program latihan untuk pemula hingga lanjutan, serta gerakan latihan beban utama.

2.8.3 Physical Activity Guidelines for Americans 2018

Physical Activity Guidelines for Americans edisi ke-2 adalah panduan aktivitas fisik resmi yang diterbitkan oleh *U.S. Department of Health and Human Services* (HHS) [13]. Dokumen ini memberikan rekomendasi berbasis bukti

ilmiah tentang jumlah dan jenis aktivitas fisik yang diperlukan untuk meningkatkan kesehatan dan mengurangi risiko penyakit kronis bagi semua kelompok usia.

2.8.4 ACSM Position Stand: Nutrition and Athletic Performance

ACSM Position Stand: Nutrition and Athletic Performance adalah panduan nutrisi olahraga resmi yang diterbitkan secara bersama oleh ACSM, *Academy of Nutrition and Dietetics* (AND), dan *Dietitians of Canada* (DC) [10]. Dokumen ini memuat rekomendasi kebutuhan energi, karbohidrat, protein, lemak, hidrasi, dan suplemen olahraga berbasis bukti ilmiah untuk atlet dan individu aktif secara fisik, sehingga dikategorikan sebagai dokumen domain gabungan karena membahas aspek nutrisi yang secara spesifik ditujukan untuk mendukung aktivitas dan performa latihan fisik.

2.8.5 USDA Dietary Guidelines for Americans 2020–2025

Dietary Guidelines for Americans 2020–2025 adalah panduan nutrisi resmi yang diterbitkan oleh *U.S. Department of Agriculture* (USDA) dan HHS [13]. Dokumen ini memuat rekomendasi distribusi makronutrien, kebutuhan kalori harian berdasarkan usia dan tingkat aktivitas, serta batasan konsumsi gula, lemak jenuh, dan natrium.

2.8.6 Pedoman Gizi Seimbang Kementerian Kesehatan RI

Pedoman Gizi Seimbang adalah panduan nutrisi resmi yang diterbitkan oleh Kementerian Kesehatan Republik Indonesia berdasarkan Peraturan Menteri Kesehatan Nomor 41 Tahun 2014 [31]. Dokumen ini memuat sepuluh pesan gizi seimbang, konsep Tumpeng Gizi Seimbang, Angka Kecukupan Gizi (AKG) 2019, serta panduan porsi makan untuk berbagai kelompok usia di Indonesia.

2.8.7 AHA Dietary Guidance to Improve Cardiovascular Health 2021

Dietary Guidance to Improve Cardiovascular Health adalah *scientific statement* yang diterbitkan oleh *American Heart Association* (AHA) pada tahun 2021 [1]. Dokumen ini memuat sepuluh fitur pola makan sehat jantung berbasis bukti ilmiah, meliputi rekomendasi konsumsi buah dan sayuran, biji-bijian utuh,

protein sehat, lemak tak jenuh, serta batasan konsumsi gula, garam, dan makanan ultra-olahan.

2.8.8 WHO Healthy Diet Guidelines

WHO Healthy Diet Guidelines adalah panduan diet sehat global yang diterbitkan oleh *World Health Organization* (WHO) [3]. Dokumen ini memuat rekomendasi konsumsi buah dan sayuran minimal 400 gram per hari, batasan konsumsi gula bebas, lemak jenuh, lemak trans, dan garam, serta panduan diet sehat untuk berbagai kelompok usia berdasarkan bukti ilmiah internasional.

2.9 Penelitian Terkait

Beberapa penelitian terdahulu yang relevan dengan penelitian ini dirangkum pada Tabel 2.5.

Tabel 2.5. Ringkasan penelitian terkait sistem RAG dan *chatbot* berbasis LLM

Peneliti	**Metode**	**Domain**	**Hasil**
Dhaman et al. (2025)[14]	RAG + *nomic-embed-text* + Llama3.1/3.2	Layanan pelanggan	Evaluasi ROUGE + *human evaluation* Likert 1–5
Saman et al. (2025) [16]	RAG + LLaMA 3B + Indo-SBERT	Informasi akademik	Nilai RAGAS 0.9, latensi $<$ 1 detik
GastroBot (2024) [17]	RAG berbasis GPT+ *gte-base-zh* (*fine-tuned*)	Gastrointestinal	*Faithfulness* 93.73%, *relevancy* 92.28%
Lyu et al. (2025) [20]	CRUD-RAG *benchmark*	*Question answering*	*Chunk size* kecil lebih baik untuk QA
Smole et al. (2025) [19]	*Benchmark* 90 konfigurasi	*Information retrieval*	*Window* 512 + *overlap* 200 token, IoU tertinggi

Tabel 2.5 menunjukkan beberapa hal penting terkait posisi penelitian ini, khususnya pada aspek pemilihan model *embedding*. Saman et al. mengukur performa RAG menggunakan RAGAS dengan nilai komposit sebesar 0.9, namun juga menggunakan *Mean Reciprocal Rank* (MRR) 0.5 dan latensi di bawah 1 detik sebagai metrik tambahan. Penelitian ini tidak mengadopsi MRR karena MRR membutuhkan data *ground truth* urutan relevansi dokumen yang tidak tersedia.

Evaluasi performa sistem dalam penelitian ini difokuskan pada tiga metrik RAGAS *reference-free* yang lebih sesuai dengan kondisi dataset yang tersedia.

Dhaman et al. [14] menggunakan model *embedding nomic-embed-text* yang dijalankan secara lokal melalui Ollama untuk *chatbot* layanan pelanggan berbahasa Indonesia, dengan *chunk size* 1000 kata, *overlap* 200 kata, dan *top-k* 4 menggunakan *cosine similarity*. Penelitian tersebut memfokuskan eksperimen pada perbandingan ukuran parameter LLM (Llama3.1:8B dibandingkan Llama3.2:1B/3B), bukan pada perbandingan atau justifikasi pemilihan model *embedding* itu sendiri, dan tidak membahas kesesuaian *nomic-embed-text* untuk konten multibahasa atau *code-mixing*. Selain itu, evaluasi yang digunakan adalah ROUGE yang membutuhkan jawaban referensi (*ground truth*) serta *human evaluation* berskala Likert 1–5, berbeda dengan pendekatan *reference-free* pada *framework* RAGAS yang digunakan dalam penelitian ini. Dhaman et al. [14] turut mengidentifikasi bahwa penelitian RAG untuk dokumen berbahasa Indonesia dan model LLM ringan (Llama3.2 1B/3B) masih terbatas.

Sementara itu, GastroBot [17] melakukan *fine-tuning* model *embedding gte-base-zh* secara khusus untuk domain penyakit gastrointestinal berbahasa Mandarin, menunjukkan pendekatan *domain-specific fine-tuning* sebagai salah satu strategi peningkatan kualitas *embedding*. Pendekatan ini efektif namun memerlukan proses *fine-tuning* tambahan yang membutuhkan data dan sumber daya komputasi lebih besar dibandingkan penggunaan model *embedding* multibahasa yang sudah terlatih (*pre-trained*).

Dibandingkan dengan kedua penelitian tersebut, penelitian ini memilih LaBSE sebagai model *embedding* multilingual siap pakai (*off-the-shelf*, tanpa proses *fine-tuning* tambahan) yang telah dilatih untuk 109 bahasa termasuk bahasa Indonesia [29], sebagai solusi atas kesenjangan kesesuaian model *embedding* untuk domain kebugaran dan nutrisi berbahasa Indonesia. Penelitian ini juga menggunakan pendekatan evaluasi *reference-free* melalui *framework* RAGAS, yang lebih sesuai untuk sistem RAG tanpa ketersediaan jawaban referensi berlabel, dibandingkan pendekatan ROUGE yang memerlukan *ground truth* seperti yang digunakan oleh Dhaman et al. [14].

Berdasarkan tinjauan penelitian terdahulu, dapat diidentifikasi bahwa implementasi sistem RAG untuk domain *gym* dan nutrisi makanan secara bersamaan dengan evaluasi menggunakan parameter *chunk size*, *chunk overlap*, dan *top-k* belum pernah dilakukan sebelumnya. Penelitian ini hadir untuk mengisi celah tersebut.