BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Akselerasi transformasi digital dalam sektor kesehatan dan kebugaran telah memicu integrasi kecerdasan buatan (*Artificial Intelligence*/AI) secara masif [1]. Di era modern, industri kebugaran mulai beralih pada teknologi *chatbot* cerdas sebagai solusi penyediaan panduan kesehatan, perancangan rencana latihan (*workout plans*), hingga dukungan motivasi yang dipersonalisasi [2, 3]. Pemanfaatan *Large Language Model* (LLM) pada *chatbot* menjadi pilihan utama karena kemampuannya dalam berinteraksi melalui bahasa alami yang luwes, ketersediaan layanan 24/7, serta kemudahan akses bagi pengguna dalam menunjang aktivitas fisik [2, 4]. Jrke et al. [5] melalui pengembangan GPTCoach menunjukkan bahwa *chatbot* berbasis LLM mampu memberikan panduan aktivitas fisik yang terpersonalisasi, sementara Khamesian et al. [6] mengembangkan NutriGen berbasis LLM yang mampu menghasilkan rencana makan personal sesuai kebutuhan gizi pengguna.

Kebugaran fisik (*physical fitness*) dan nutrisi merupakan dua aspek yang tidak dapat dipisahkan dalam konteks kesehatan holistik. Secara akademik, kebugaran fisik didefinisikan sebagai kondisi yang umumnya dicapai melalui nutrisi yang tepat, latihan fisik dengan intensitas sedang hingga tinggi, serta istirahat yang memadai [7]. Lebih lanjut, *body composition* salah satu dari lima komponen *health-related fitness* versi ACSM secara langsung dipengaruhi oleh kebiasaan makan dan metabolisme individu [8]. Nazni dan Vimala [9] dalam penelitiannya yang dipublikasikan di *Asian Journal of Sports Medicine* secara eksplisit menyatakan bahwa nutrisi merupakan komponen penting dari setiap program kebugaran fisik, dengan tujuan utama memperoleh asupan gizi yang memadai untuk mengoptimalkan kesehatan dan meningkatkan performa fisik. Oleh karena itu, penggabungan domain *gym* dan nutrisi dalam satu sistem *chatbot* memiliki landasan ilmiah yang kuat.

Namun, LLM generik sering kali menemui kendala dalam menjalankan peran sebagai asisten *gym* yang spesifik. Model bahasa umum yang tidak terikat pada basis pengetahuan domain tertentu kerap gagal mengintegrasikan prinsip latihan standar, seperti pedoman dari *American College of Sports Medicine* (ACSM)

[10]. Kendala paling krusial adalah fenomena halusinasi, yakni kondisi di mana LLM menghasilkan informasi yang terdengar meyakinkan namun secara faktual keliru [11, 12]. Zaleski et al. [10] dalam penelitiannya yang dipublikasikan di *JMIR Medical Education* menemukan bahwa rekomendasi latihan fisik yang dihasilkan oleh ChatGPT memiliki tingkat kelengkapan dan akurasi yang bervariasi, serta belum sepenuhnya sesuai dengan pedoman ACSM. Senada dengan itu, Solomon et al. [13] yang mengevaluasi berbagai model LLM yaitu ChatGPT, Gemini, dan Claude dalam menjawab pertanyaan nutrisi olahraga menemukan bahwa pada Experiment 1 akurasi bervariasi dari 31% (ClaudePro) hingga 74% (Gemini1.5pro), sedangkan pada Experiment 2 akurasi berkisar antara 61% hingga 89%, dengan banyak jawaban yang tidak konsisten dengan pedoman nutrisi ilmiah yang terverifikasi.

Sebagai ilustrasi konkret, dilakukan perbandingan terhadap tiga pertanyaan seputar latihan fisik dan nutrisi antara ChatGPT tanpa RAG dan sistem *chatbot* RAG yang dibangun dalam penelitian ini. Pada pertanyaan mengenai kebutuhan protein untuk atlet kekuatan, ChatGPT memberikan rentang 1,6–2,2 gram per kg berat badan disertai rincian perhitungan dan berbagai sub-kondisi tanpa menyebutkan sumber referensi, sedangkan sistem RAG memberikan jawaban 1,6–1,7 gram per kg berat badan yang secara langsung dapat ditelusuri ke dokumen *ACSM Position Stand: Nutrition and Athletic Performance* [8]. Pada pertanyaan mengenai durasi aktivitas fisik mingguan dan manfaat latihan *squat*, kedua sistem memberikan jawaban yang secara substansi konsisten dan tidak saling bertentangan. Temuan ini menunjukkan bahwa perbedaan mendasar antara RAG dan LLM generik bukan selalu terletak pada ada tidaknya kesalahan faktual, melainkan pada aspek *traceability*: jawaban sistem RAG dapat ditelusuri secara langsung ke dokumen pedoman resmi yang telah terverifikasi, sedangkan jawaban ChatGPT tidak menyertakan sumber yang dapat diverifikasi meskipun disampaikan dengan tingkat kepercayaan diri yang tinggi — suatu karakteristik yang menjadi akar dari risiko halusinasi pada LLM [12]. Karakteristik *traceability* inilah yang menjadi nilai tambah utama pendekatan RAG, terutama untuk domain kesehatan yang membutuhkan akuntabilitas sumber informasi.

Untuk memitigasi risiko tersebut, metode *Retrieval-Augmented Generation* (RAG) muncul sebagai solusi strategis [14, 15]. RAG mengombinasikan mekanisme pencarian informasi (*retrieval*) dari basis data eksternal yang valid dengan kapabilitas generatif teks LLM [14, 16]. Melalui arsitektur ini, *chatbot* tidak hanya bersandar pada data pelatihan internal model, melainkan melakukan

ekstraksi informasi dari dokumen pedoman yang terverifikasi sebelum menyusun respons [11]. Pendekatan ini secara signifikan mampu mereduksi tingkat halusinasi, meningkatkan akurasi faktual, dan memberikan transparansi berbasis sumber bukti yang relevan [11, 14].

RAG telah terbukti efektif dalam meningkatkan akurasi jawaban sistem *chatbot* di berbagai domain. Dhaman et al. [14] mengimplementasikan RAG untuk sistem *chatbot* layanan pelanggan menggunakan *framework* LangChain dan ChromaDB, dan melaporkan peningkatan akurasi jawaban yang signifikan dibandingkan sistem tanpa RAG. Saman et al. [16] membangun *chatbot* informasi akademik berbasis RAG menggunakan model LLaMA 3B dan Indo-SBERT, dan melaporkan nilai RAGAS sebesar 0.9 dengan latensi respons di bawah 1 detik. Pada domain kesehatan yang lebih spesifik, GastroBot yang dikembangkan sebagai *chatbot* penyakit gastrointestinal berbasis RAG melaporkan nilai *faithfulness* sebesar 93.73%, *answer relevancy* sebesar 92.28%, dan *context recall* sebesar 95% [17].

Meskipun penelitian-penelitian tersebut telah membuktikan efektivitas RAG, terdapat celah penelitian yang belum terjawab secara spesifik. Dhaman et al. [14] dan Saman et al. [16] mengimplementasikan RAG pada domain layanan pelanggan dan informasi akademik, namun tidak menganalisis pengaruh variasi parameter *chunk size*, *chunk overlap*, maupun *top-k* terhadap performa sistem — kedua penelitian tersebut hanya melaporkan hasil dari satu konfigurasi parameter tanpa perbandingan sistematis. Demikian pula GastroBot [17] pada domain gastrointestinal, yang melaporkan skor *faithfulness* dan *relevancy* tinggi namun tanpa mengungkap bagaimana nilai tersebut dipengaruhi oleh konfigurasi *retrieval* yang digunakan. Di sisi lain, penelitian yang secara khusus menganalisis parameter *chunking* seperti Hladěna et al. [18], Smole et al. [19], dan Lyu et al. [20] berfokus pada domain *general-purpose information retrieval* dan *question answering*, bukan pada domain kesehatan atau kebugaran yang memiliki karakteristik dokumen dan pola pertanyaan yang berbeda. Belum ditemukan penelitian yang secara sistematis mengombinasikan kedua aspek tersebut, yaitu evaluasi pengaruh parameter *chunk size* (64, 128, 256, 512, 1024 token), *chunk overlap* (0%, 10%, 20%), dan *top-k retrieval* (3, 5, 10) secara bersamaan pada domain *gym* dan nutrisi makanan. Selain itu, sebagian besar penelitian RAG yang telah disebutkan menggunakan model LLM komersial berbasis *cloud* (seperti GPT), yang memerlukan biaya API dan tidak menjamin privasi data pengguna, sehingga kurang sesuai untuk diterapkan pada konteks aplikasi kesehatan yang bersifat sensitif. Penelitian ini

hadir untuk mengisi celah tersebut dengan menggunakan model LLM *open-source* yang dijalankan secara lokal.

Untuk memperjelas posisi penelitian ini terhadap penelitian terdahulu, Tabel 1.1 merangkum perbandingan cakupan penelitian RAG yang relevan.

Tabel 1.1. Perbandingan cakupan penelitian RAG terdahulu dengan penelitian ini

Penelitian	**Domain**	**Uji Variasi Parameter *Chunking*?**	**Model LLM**	**Lokal/Privasi Data?**
Dhaman et al. [14]	Layanan pelanggan	Tidak	LangChain + LLM *cloud*	Tidak
Saman et al. [16]	Informasi akademik	Tidak	LLaMA 3B	Ya
GastroBot [17]	Gastrointestinal	Tidak	GPT (*cloud*)	Tidak
Hladěna et al. [18]	*General-purpose IR*	Ya (*chunk size*)	Tidak spesifik	—
Smole et al. [19]	Information retrieval	Ya (*chunk size*, *overlap*)	Tidak spesifik	—
Lyu et al. [20]	Question answering	Ya (*chunk size*, *top-k*)	Tidak spesifik	—
Penelitian ini	**Gym & nutrisi**	**Ya (*chunk size*, *overlap*, *top-k*)**	**Mistral-7B**	**Ya**

Implementasi sistem RAG memerlukan evaluasi yang terukur untuk memastikan kualitas jawaban yang dihasilkan. *Framework* RAGAS (*Retrieval-Augmented Generation Assessment*) hadir sebagai solusi evaluasi otomatis yang mengukur performa sistem RAG melalui metrik yang bersifat *reference-free* sehingga tidak membutuhkan data jawaban referensi yang telah dianotasi secara manual [21]. Dalam penelitian ini digunakan tiga metrik utama yaitu *faithfulness*, *response relevancy*, dan *context precision* yang tersedia pada versi terbaru *framework* RAGAS [21].

Meskipun RAG terbukti efektif, performa sistem RAG sangat dipengaruhi oleh konfigurasi parameter *retrieval*-nya. Parameter seperti *chunk size* (ukuran potongan dokumen), *chunk overlap* (irisan antar potongan), dan *top-k* (jumlah potongan yang diambil *retriever*) memiliki pengaruh signifikan terhadap efisiensi pengambilan informasi dan kualitas respons [18]. Hladěna et al. [18] dalam penelitiannya di Springer menunjukkan bahwa konfigurasi *chunk size* berpengaruh signifikan terhadap efisiensi *retrieval* dan kualitas respons sistem RAG, di mana ukuran potongan yang terlalu kecil maupun terlalu besar dapat menurunkan akurasi pencarian. Smole et al. [19] melalui *benchmark* terhadap 90 konfigurasi *chunker* yang dipublikasikan di IEEE Conference on Artificial Intelligence menemukan bahwa *sentence-based splitter* dengan *window* 512 token dan *overlap* 200 token mencapai nilai *Intersection-over-Union* (IoU) tertinggi sebesar 0.099 sambil tetap

efisien secara komputasi. Li et al. [22] dalam penelitiannya di COLING 2025 menunjukkan bahwa granularitas *chunking* yang optimal bergantung pada karakteristik dokumen dan jenis pertanyaan yang diajukan. Lyu et al. [20] dalam CRUD-RAG yang dipublikasikan di *ACM Transactions on Information Systems* menunjukkan bahwa *chunk size* yang terlalu kecil maupun terlalu besar dapat mengurangi akurasi pencarian atau menghilangkan konten penting, dan bahwa nilai *top-k* mempengaruhi kekayaan serta keragaman konteks yang diberikan kepada LLM. Lee et al. [23] dalam penelitiannya di Cambridge University Press juga menunjukkan bahwa kombinasi parameter *chunking* dan *retrieval* yang tepat secara signifikan mempengaruhi efisiensi dan akurasi sistem RAG.

Penggunaan model LLM *open-source* yang bersifat *lightweight* memungkinkan sistem dijalankan secara lokal dengan beban komputasi yang lebih ringan [24, 14]. Hal ini tidak hanya efisien secara sumber daya, tetapi juga krusial dalam menjaga privasi data pengguna, terutama dalam konteks aplikasi kebugaran yang semakin berkembang [25, 26].

Perlu dicatat bahwa penelitian ini berfokus pada penyediaan informasi umum berbasis pedoman resmi, bukan rekomendasi yang dipersonalisasi berdasarkan kondisi kesehatan atau kebutuhan gizi individu pengguna. Pendekatan generik ini dipilih karena personalisasi (misalnya mempertimbangkan riwayat penyakit, usia, atau tujuan kebugaran spesifik) memerlukan mekanisme tambahan seperti manajemen profil pengguna dan validasi medis yang berada di luar cakupan evaluasi parameter RAG yang menjadi fokus utama penelitian ini. Fokus penelitian diarahkan pada akurasi dan relevansi jawaban berbasis dokumen terverifikasi secara ilmiah, sebagai fondasi sebelum pengembangan lapisan personalisasi pada penelitian selanjutnya.

Berdasarkan permasalahan dan celah penelitian yang telah diidentifikasi, penelitian ini berfokus pada implementasi sistem *chatbot* berbasis RAG untuk domain *gym* dan nutrisi makanan, dengan mengevaluasi pengaruh parameter *chunk size*, *chunk overlap*, dan *top-k retrieval* menggunakan *framework* RAGAS.

1.2 Rumusan Masalah

Rumusan masalah dari penelitian ini adalah sebagai berikut.

1. Bagaimana pengaruh kombinasi parameter *chunk size*, *chunk overlap*, dan *top-k retrieval* terhadap kualitas jawaban sistem *chatbot* berbasis *Retrieval-Augmented Generation* (RAG) pada domain *gym* dan nutrisi makanan, yang

diukur menggunakan metrik *faithfulness*, *response relevancy*, dan *context precision*?

2. Konfigurasi parameter manakah yang menghasilkan performa optimal berdasarkan rata-rata skor ketiga metrik RAGAS dari lima belas kombinasi eksperimen yang diuji?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut.

1. Sistem *chatbot* hanya dapat menjawab pertanyaan dalam domain *gym* dan nutrisi makanan berdasarkan sepuluh dokumen pedoman resmi yang tersedia dalam *knowledge base*. Pertanyaan di luar cakupan tersebut tidak dapat dijawab secara akurat oleh sistem.
2. Sistem tidak memiliki kemampuan memori percakapan (*conversational memory*), sehingga setiap pertanyaan diproses secara independen tanpa mengingat konteks pertanyaan sebelumnya.
3. Sistem tidak dapat memberikan rekomendasi yang dipersonalisasi berdasarkan kondisi kesehatan atau kebutuhan gizi spesifik pengguna. Jawaban yang diberikan bersifat umum sesuai panduan yang tersedia dalam *knowledge base*.
4. Waktu respons sistem bergantung pada spesifikasi perangkat keras karena model LLM dijalankan secara lokal, dengan rata-rata waktu respons berkisar antara 40 hingga 60 detik per pertanyaan.
5. Sistem tidak menjamin akurasi jawaban sebesar 100% karena nilai *faithfulness* rata-rata yang dicapai adalah 0.8422, sehingga masih terdapat kemungkinan sebagian pernyataan tidak sepenuhnya dapat diverifikasi dari dokumen sumber.
6. Evaluasi sistem dibatasi pada tiga metrik RAGAS yaitu *faithfulness*, *response relevancy*, dan *context precision* terhadap lima belas konfigurasi eksperimen menggunakan 50 pertanyaan evaluasi.

1.4 Tujuan Penelitian

Penelitian ini dilakukan dengan tujuan sebagai berikut.

1. Mengevaluasi pengaruh kombinasi parameter *chunk size* (64, 128, 256, 512, 1024 token), *chunk overlap* (0%, 10%, 20%), dan *top-k retrieval* (3, 5, 10) terhadap kualitas jawaban sistem *chatbot* berbasis RAG pada domain *gym* dan nutrisi makanan, yang diukur menggunakan metrik *faithfulness*, *response relevancy*, dan *context precision*.
2. Mengidentifikasi konfigurasi parameter RAG yang menghasilkan performa optimal berdasarkan rata-rata skor ketiga metrik RAGAS dari lima belas kombinasi eksperimen yang diuji.

1.5 Manfaat Penelitian

Hasil penelitian ini diharapkan memberikan kontribusi nyata baik secara praktis maupun teoretis.

- Bagi pengguna dan fasilitas kebugaran, penelitian ini menyediakan sistem *chatbot* yang dapat memberikan informasi latihan fisik dan nutrisi makanan secara akurat berbasis dokumen pedoman resmi yang telah terverifikasi secara ilmiah.
- Bagi pengembangan ilmu informatika, penelitian ini memberikan bukti empiris mengenai pengaruh konfigurasi parameter RAG terhadap performa sistem pada domain terspesialisasi, yang dapat menjadi referensi teknis bagi peneliti dan pengembang dalam merancang sistem RAG yang optimal.
- Bagi penelitian selanjutnya, penelitian ini menyediakan data eksperimen yang sistematis mengenai kombinasi *chunk size*, *chunk overlap*, dan *top-k* pada model Mistral-7B, yang dapat dijadikan dasar penelitian lanjutan dengan model atau domain yang berbeda.

1.6 Sistematika Penulisan

Sistematika penulisan laporan penelitian ini adalah sebagai berikut.

- Bab 1 PENDAHULUAN

  Bab ini membahas latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

- Bab 2 LANDASAN TEORI

  Bab ini membahas teori-teori yang menjadi landasan penelitian, meliputi *Large Language Model* (LLM), *Retrieval-Augmented Generation* (RAG), Mistral-7B, Ollama, ChromaDB, *framework* RAGAS, serta dataset yang digunakan dalam penelitian.

- Bab 3 METODOLOGI PENELITIAN

  Bab ini menjelaskan metodologi penelitian yang digunakan, meliputi alur penelitian, teknik pengumpulan data, desain eksperimen, dan teknik analisis data.

- Bab 4 HASIL DAN DISKUSI

  Bab ini menyajikan hasil implementasi sistem, hasil eksperimen dari lima belas kombinasi konfigurasi parameter RAG, serta analisis dan pembahasan terhadap skor RAGAS yang diperoleh dari setiap konfigurasi.

- Bab 5 SIMPULAN DAN SARAN

  Bab ini menyajikan simpulan dari hasil penelitian berdasarkan rumusan masalah yang telah ditetapkan, serta saran untuk pengembangan penelitian selanjutnya.