# BAB I
# PENDAHULUAN

Bab pendahuluan berisi uraian awal yang menjadi dasar penelitian mengenai optimasi *hyperparameter* arsitektur *Retrieval-Augmented Generation* (RAG) pada *chatbot* ekstraksi informasi jurnal ilmiah. Bab ini membahas latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, ruang lingkup penelitian, serta sistematika penulisan skripsi.

## 1.1 Latar Belakang

Perkembangan literatur ilmiah di bidang teknologi informasi mengalami pertumbuhan yang sangat pesat setiap tahunnya. Fenomena ledakan informasi ini, sebagaimana ditegaskan oleh Bornmann dan Mutz (2015), membuat volume publikasi tumbuh secara eksponensial. Di Indonesia, SINTA (*Science and Technology Index*) merupakan sistem indeksasi dan pemeringkatan jurnal ilmiah nasional yang digunakan untuk menilai kinerja jurnal, peneliti, dan institusi. Salah satu wadah publikasi ilmiah yang menampung masifnya penelitian di bidang teknologi informasi adalah Jurnal Rekayasa Sistem dan Teknologi Informasi (RESTI). Sebagai jurnal nasional terakreditasi SINTA 2 yang terbit enam kali dalam setahun, Jurnal RESTI menyediakan ratusan artikel ilmiah yang menjadi sumber referensi bagi mahasiswa, dosen, dan peneliti di bidang teknologi informasi di Indonesia.

Meskipun Jurnal RESTI menyediakan sumber pengetahuan yang melimpah, tingginya volume artikel tersebut memunculkan tantangan terkait aksesibilitas dan efisiensi pencarian informasi. Sering kali, mahasiswa atau peneliti mengalami kendala karena proses pencarian informasi spesifik di dalam artikel memakan waktu yang cukup lama. Pengguna harus menemukan artikel yang relevan terlebih dahulu, mengunduh dokumen secara utuh, lalu membacanya secara mendetail dan terperinci. Padahal, dalam praktiknya, pengguna sering kali hanya membutuhkan ekstraksi informasi spesifik tertentu untuk keperluan referensi akademik. Proses pencarian dan pembacaan manual secara menyeluruh ini pada akhirnya memicu inefisiensi waktu yang signifikan bagi para pengguna.

Proses pencarian informasi pada artikel ilmiah juga dipengaruhi oleh keterbatasan fitur pencarian konvensional. Sistem pencarian berbasis leksikal (*keyword search*) sering gagal menemukan dokumen yang relevan apabila kosakata yang digunakan pengguna berbeda dengan istilah teknis di dalam teks, meskipun maknanya sama. Keterbatasan tersebut sejalan dengan studi Karpukhin dkk. (2020), yang menunjukkan bahwa pencarian informasi membutuhkan pendekatan berbasis vektor semantik (*dense retrieval*) agar sistem mampu memetakan konteks pertanyaan secara bermakna.

Pengembangan *chatbot* berbasis *Large Language Model* (LLM) menjadi solusi potensial untuk menjembatani kebutuhan pemahaman teks akademik. Namun, para pengembang dihadapkan pada dilema pemilihan arsitektur. Penggunaan LLM standar memiliki keterbatasan berupa kecenderungan berhalusinasi atau menghasilkan jawaban yang tidak sepenuhnya sesuai dengan dokumen sumber apabila tidak diberikan konteks yang relevan. Untuk meminimalisir halusinasi, arsitektur RAG yang diperkenalkan oleh Lewis dkk. (2020) telah banyak digunakan sebagai pendekatan untuk mengaitkan jawaban dengan dokumen referensi. Gao dkk. (2023) juga menjelaskan bahwa RAG memanfaatkan sumber pengetahuan eksternal, sehingga pembaruan informasi dapat dilakukan melalui dokumen atau basis data tanpa harus selalu melakukan pelatihan ulang model sebagaimana pada *fine-tuning*. Dalam hal penyajian jawaban, Izacard dan Grave (2021) menunjukkan bahwa model generatif yang memanfaatkan hasil *retrieval* mampu memproses beberapa *passage* sebagai sumber informasi untuk menghasilkan jawaban.

Penerapan arsitektur RAG pada domain artikel ilmiah tetap membutuhkan perhatian terhadap keterlacakan jawaban pada dokumen sumber. Lee dkk. (2023) melalui penelitian QASA menunjukkan bahwa *question answering* (QA) pada artikel ilmiah membutuhkan pemilihan *evidence* dan penyusunan jawaban yang didasarkan dari dokumen. Sejalan dengan itu, Dasigi dkk. (2021) melalui QASPER menekankan pentingnya *supporting evidence* dalam penyusunan jawaban berbasis *paper* ilmiah. Namun demikian, penelitian-penelitian sebelumnya umumnya berfokus pada evaluasi arsitektur atau kemampuan model secara umum dan belum ada yang mengkaji secara komprehensif pengaruh konfigurasi parameter RAG terhadap performa ekstraksi informasi pada dokumen ilmiah.

Meskipun RAG menawarkan fleksibilitas, implementasinya pada dokumen akademik yang panjang memunculkan tantangan teknis baru. Penelitian Liu dkk. (2023)

menemukan fenomena *Lost in the Middle*, yaitu penurunan performa model ketika informasi relevan berada pada bagian tengah konteks yang panjang. Fenomena tersebut menegaskan bahwa penentuan strategi pemotongan teks (*chunking*) dan pengelolaan *retrieval method* menjadi fase arsitektural yang esensial untuk menjaga akurasi pemahaman mesin.

Oleh karena itu, penelitian ini bertujuan untuk mengimplementasikan dan mengoptimasi *chatbot* LLM sebagai sistem QA ekstraksi informasi dengan arsitektur RAG pada domain literatur akademik berbahasa Indonesia. Dengan mengambil Jurnal RESTI sebagai studi kasus dan korpus data utama, penelitian ini memberikan kebaruan melalui eksperimen komparatif untuk mencari konfigurasi terbaik antara ukuran potongan teks (*chunk size*), irisan konteks (*chunk overlap*), jumlah dokumen yang diambil (*Top-K*), dan metode penarikan data (*retrieval method*). Kinerja sistem selanjutnya dievaluasi secara komprehensif menggunakan metrik leksikal ROUGE dan semantik BERTScore untuk memastikan *chatbot* dapat mengekstraksi dan merangkum informasi jurnal ilmiah secara akurat, cepat, dan relevan tanpa mengalami degradasi konteks.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, rumusan masalah dalam penelitian ini adalah: bagaimana pengaruh optimasi parameter *Chunk Size, Chunk Overlap, Top-K,* dan *Retrieval Method* pada arsitektur RAG terhadap akurasi *Chatbot* LLM sebagai sistem QA ekstraksi informasi berbasis artikel ilmiah?

## 1.3 Tujuan

Tujuan dari penelitian ini adalah mengimplementasikan dan menemukan nilai parameter *Chunk Size, Chunk Overlap, Top-K,* dan *Retrieval Method* yang paling optimal pada *Chatbot* LLM sebagai sistem QA ekstraksi informasi berbasis artikel ilmiah dengan pendekatan RAG.

## 1.4 Manfaat Penelitian

Manfaat yang dapat diberikan dari penelitian ini adalah memberikan referensi empiris mengenai konfigurasi parameter *Chunk Size, Chunk Overlap, Top-K*, dan *Retrieval Method* yang paling optimal pada arsitektur RAG. Selain itu, sistem yang dibangun diharapkan dapat memberikan kemudahan dalam pencarian informasi spesifik serta

meningkatkan efisiensi waktu pengguna saat mengekstraksi literatur akademik atau artikel ilmiah yang panjang melalui *chatbot* interaktif.

## 1.5 Ruang Lingkup

Untuk memfokuskan penelitian dan menghindari perluasan pembahasan, ruang lingkup pada penelitian ini dibatasi pada:

1. Basis pengetahuan sistem dibangun menggunakan 100 dokumen artikel ilmiah lintas sub-topik dari Jurnal RESTI.
2. Pengujian dan evaluasi akurasi sistem (*Ground Truth*) dibatasi dan difokuskan secara spesifik pada 50 kueri faktual yang bersumber dari 10 dokumen (diambil secara acak) sebagai sampel uji atau Korpus Terkontrol.
3. Parameter arsitektur RAG yang dioptimasi dan dievaluasi dibatasi hanya pada variasi *Chunk Size*, *Chunk Overlap*, *Top-K*, serta *Retrieval Method* (*Semantic Similarity Search* dan *Maximal Marginal Relevance*).
4. Ekstraksi informasi dari dokumen PDF hanya difokuskan pada data berbasis teks (huruf dan angka). Elemen non-teks di dalam artikel ilmiah seperti gambar, diagram, grafik, serta tabel yang tidak terdeteksi sebagai teks, tidak akan diproses oleh sistem dan berada di luar cakupan penelitian ini.
5. *Chatbot* tidak menjawab pertanyaan di luar konteks dari dokumen referensi yang telah disediakan untuk mencegah halusinasi.
6. Metrik Evaluasi diukur secara kuantitatif menggunakan metrik *Response Time*, *Coverage*, ROUGE-1, ROUGE-L, dan BERTScore.

## 1.6 Sistematika Penulisan

Untuk memberikan gambaran yang jelas dan menyeluruh mengenai penyusunan skripsi ini, penulis membagi struktur penulisan ke dalam lima bab utama. Adapun sistematika penulisannya adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini menguraikan latar belakang permasalahan mengenai *information overload* pada literatur ilmiah dan rentannya halusinasi pada LLM. Bab ini juga

memuat rumusan masalah, batasan ruang lingkup penelitian, tujuan yang ingin dicapai, manfaat penelitian, serta sistematika penulisan laporan.

BAB II LANDASAN TEORI

Bab ini membahas tinjauan literatur dari penelitian-penelitian terdahulu (*State of the Art*) yang relevan dengan topik yang diangkat. Selain itu, bab ini juga menjabarkan landasan teori yang menjadi dasar rancang bangun sistem, meliputi konsep LLM, RAG, pemrosesan bahasa alami, teknik *chunking*, basis data vektor, hingga perhitungan metrik evaluasi seperti ROUGE dan BERTScore.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan tahapan dan alur penelitian yang dilakukan secara sistematis. Pembahasan mencakup metode pengumpulan korpus data dari artikel Jurnal RESTI, perancangan arsitektur *pipeline* sistem RAG (*retriever* dan *generator*), pembuatan data uji (*ground truth*), hingga penetapan rancangan eksperimen yang terdiri dari 72 skenario kombinasi *hyperparameter* (*Chunk Size, Chunk Overlap, Top-K*, dan *Retrieval Method*).

BAB IV HASIL DAN PEMBAHASAN

Bab ini merupakan inti dari penelitian yang menguraikan hasil implementasi sistem *chatbot* secara *end-to-end*, mulai dari proses konversi *Knowledge Store* hingga tahap pembangkitan jawaban. Bab ini juga menyajikan visualisasi data dan analisis mendalam terhadap hasil komparasi metrik evaluasi pengujian (Waktu Respon, *Coverage*, ROUGE-1, ROUGE-L, dan BERTScore) guna menentukan konfigurasi arsitektur RAG yang paling optimal.

BAB V PENUTUP

Bab ini berisi kesimpulan akhir dari keseluruhan hasil eksperimen dan analisis yang telah dilakukan untuk menjawab rumusan masalah. Selain itu, bab ini juga memuat saran-saran konstruktif sebagai rekomendasi bagi penelitian atau pengembangan sistem serupa di masa mendatang.