

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan *Large Language Models* (LLM) dalam beberapa tahun terakhir telah membawa kemajuan signifikan pada berbagai tugas pemrosesan bahasa alami, seperti *question answering*, *text summarization*, dan *information retrieval* [1]. LLM mampu menghasilkan respons secara fleksibel dan kontekstual. Namun, model generatif murni masih memiliki keterbatasan dalam menghasilkan informasi yang akurat dan dapat dipertanggungjawabkan [2, 3]. Salah satu permasalahan utama yang sering muncul adalah fenomena *hallucination*, yaitu kondisi ketika model menghasilkan jawaban yang tampak meyakinkan tetapi tidak sesuai dengan fakta. *Li et al.* melalui *benchmark* HaluEval melaporkan bahwa ChatGPT cenderung menghasilkan konten terhalusinasi berupa informasi yang tidak dapat diverifikasi pada sekitar 19,5% kueri pengguna [4]. Risiko ini menjadi perhatian serius pada sistem tanya jawab yang menggunakan dokumen resmi sebagai sumber informasi, karena kesalahan jawaban dapat menyebabkan pengguna memperoleh informasi yang keliru atau tidak sesuai dengan isi dokumen yang menjadi rujukan.

Salah satu jenis dokumen resmi yang banyak digunakan sebagai sumber informasi adalah dokumen PDF. Format PDF umum digunakan karena mampu mempertahankan tata letak dan memuat berbagai bentuk informasi, seperti teks naratif, tabel, daftar, struktur bab dan subbab, gambar, lampiran, serta halaman hasil pemindaian. Dokumen PDF yang memuat kombinasi elemen tersebut disebut dokumen PDF multimodal. Istilah *multimodal* dalam penelitian ini merujuk pada keragaman modus penyajian informasi dalam satu dokumen, yaitu teks digital, elemen semi-terstruktur (tabel dan daftar), serta informasi tekstual yang termuat dalam citra hasil pemindaian maupun gambar. Informasi tekstual pada citra hasil pemindaian maupun pada gambar yang memuat teks diekstraksi melalui *Optical Character Recognition* (OCR). Cakupan pemrosesan dibatasi pada perolehan kandungan informasi yang dapat direpresentasikan sebagai teks, dan tidak mencakup analisis konten audio-visual. Karakteristik ini menjadikan dokumen PDF lebih kompleks untuk diproses dibandingkan dokumen teks biasa, karena informasi tidak selalu tersusun dalam paragraf sederhana, melainkan dapat

tersebar dalam struktur hierarkis, tabel, gambar yang mengandung teks, maupun halaman hasil pemindaian yang memerlukan pengenalan karakter secara khusus.

Salah satu contoh dokumen PDF multimodal yang relevan dan banyak digunakan di lingkungan pendidikan tinggi adalah handbook kampus. Handbook kampus merupakan dokumen resmi yang memuat aturan akademik, prosedur administrasi, kurikulum, layanan kemahasiswaan, serta ketentuan lain yang menjadi acuan sivitas akademika. Mahasiswa sering membutuhkan informasi seperti syarat cuti akademik, prosedur pengambilan mata kuliah, atau ketentuan kelulusan yang tersedia dalam handbook. Namun, informasi tersebut umumnya tersebar pada dokumen PDF sepanjang puluhan hingga ratusan halaman, sehingga penelusuran secara manual menuntut waktu dan rawan menghasilkan informasi yang terlewat atau keliru. Kebutuhan ini berlaku pada skala yang besar: berdasarkan data Badan Pusat Statistik tahun 2025, terdapat 2.840 perguruan tinggi negeri dan swasta dengan 9.241.945 mahasiswa di Indonesia [5], yang masing-masing perlu mengakses ketentuan akademik dari dokumen sejenis. Dengan demikian, sistem yang mampu membantu pengguna menelusuri informasi akademik dari dokumen PDF secara otomatis dan akurat memiliki potensi manfaat yang besar.

Salah satu pendekatan yang dapat mengurangi risiko *hallucination* pada sistem tanya jawab berbasis dokumen adalah *Retrieval-Augmented Generation* (RAG). *Lewis et al.* [1] memperkenalkan RAG yang menggabungkan kemampuan generatif LLM dengan mekanisme pencarian informasi dari sumber eksternal, dan menunjukkan efektivitasnya pada berbagai tugas *knowledge-intensive NLP*. Sistem RAG bekerja dengan terlebih dahulu mengambil bagian dokumen yang relevan terhadap pertanyaan, kemudian menggunakannya sebagai konteks bagi model generatif untuk menghasilkan jawaban [6]. Meskipun demikian, pendekatan RAG yang diusulkan *Lewis et al.* [1] menggunakan representasi teks datar tanpa mempertimbangkan struktur hierarkis dokumen, sehingga belum mampu menangani kompleksitas dokumen PDF multimodal secara optimal.

Kualitas jawaban pada sistem RAG sangat bergantung pada kualitas proses *retrieval* [3]. *Gao et al.* [7] dan *Ranjan et al.* [8] menunjukkan bahwa pendekatan RAG konvensional yang membagi dokumen menjadi potongan teks datar berisiko kehilangan konteks lintas bagian dokumen. Permasalahan ini semakin besar pada dokumen PDF multimodal karena informasi dapat tersebar pada bab, subbab, tabel, daftar aturan, dan lampiran yang saling berkaitan. Penelitian mengenai sistem tanya jawab akademik kampus di Indonesia, seperti yang dilakukan *Tribber et al.* [9] dan *Bustomy dan Rosid* [10], juga belum menerapkan representasi hierarkis dan masih

mengandalkan pemecahan dokumen menjadi potongan teks datar, sehingga belum dirancang untuk menangani ketergantungan konteks lintas bagian dokumen.

Selain permasalahan pada mekanisme *retrieval*, proses *preprocessing* dokumen juga menjadi komponen penting dalam sistem tanya jawab berbasis dokumen PDF multimodal. Zhang *et al.* [11] dan Nguyen *et al.* [12] menunjukkan bahwa dokumen PDF dapat memiliki struktur *layout* yang beragam, seperti teks multi-kolom, tabel, *header*, *footer*, gambar, dan halaman hasil pemindaian yang memerlukan OCR. Tripathi *et al.* [13] lebih lanjut menunjukkan bahwa metode *chunking* berbasis teks konvensional gagal menangani tabel lintas halaman dan elemen visual pada dokumen PDF kompleks. Apabila ekstraksi tidak dilakukan secara adaptif sesuai karakteristik setiap halaman, informasi yang masuk ke basis pengetahuan dapat mengandung kesalahan yang berdampak pada kualitas *retrieval*.

Untuk mengatasi keterbatasan representasi datar pada mekanisme *retrieval* sebagaimana diuraikan sebelumnya, Sarthi *et al.* [14] mengusulkan RAPTOR (*Recursive Abstractive Processing for Tree-Organized Retrieval*). RAPTOR mengorganisasi dokumen ke dalam struktur hierarkis berbentuk pohon melalui *clustering* dan peringkasan rekursif, dan menunjukkan peningkatan performa pada tugas *question answering* yang memerlukan pemahaman multitingkat. Beberapa penelitian kemudian mengembangkan pendekatan serupa pada domain multimodal: Drushchak *et al.* [15] mengembangkan *pipeline* RAG yang memproses teks, gambar, tabel, dan video secara terintegrasi, sementara Yasuno [16] mengusulkan RAPTOR-AI yang mengintegrasikan pemahaman gambar ke dalam struktur hierarkis RAPTOR untuk kebutuhan respons bencana. Namun, Sarthi *et al.* [14] menguji RAPTOR pada korpus teks bersih, sedangkan Drushchak *et al.* [15] dan Yasuno [16] berfokus pada domain yang berbeda dari dokumen administratif institusional dan tidak mempertimbangkan kebutuhan *preprocessing* adaptif untuk dokumen PDF dengan struktur bab-subbab, tabel prosedur, daftar regulasi, dan halaman pemindaian.

Berdasarkan tinjauan terhadap penelitian-penelitian tersebut, terdapat celah yang belum banyak dieksplorasi, yaitu belum adanya penelitian yang mengintegrasikan *preprocessing* adaptif dokumen PDF multimodal dengan representasi hierarkis berbasis RAPTOR pada sistem tanya jawab untuk domain dokumen handbook institusional berbahasa Indonesia, celah yang juga ditegaskan oleh survei Gao *et al.* [17] mengenai *multimodal RAG* untuk pemahaman dokumen, yang mencatat bahwa penerapan representasi hierarkis seperti RAPTOR pada dokumen PDF multimodal institusional masih minim.

Berdasarkan celah tersebut, penelitian ini mengusulkan sistem tanya jawab berbasis RAG yang mengintegrasikan *preprocessing* adaptif dengan representasi hierarkis berbasis RAPTOR dalam satu *pipeline*. Dataset yang digunakan berupa 121 dokumen handbook kampus dalam format PDF dengan total 2.698 halaman, yang mencakup panduan akademik, panduan skripsi dan tesis, panduan kemahasiswaan, serta panduan sistem informasi dan layanan kampus. Dengan demikian, sistem yang dikembangkan diharapkan mampu meningkatkan kualitas *retrieval* dan jawaban pada mayoritas metrik dibandingkan pendekatan RAG konvensional, khususnya pada pertanyaan yang menuntut sintesis konteks lintas bagian dokumen.

Kontribusi penelitian ini terdiri atas dua hal. Pertama, secara aplikatif, penelitian ini merancang dan mengevaluasi integrasi *preprocessing* adaptif dan representasi hierarkis RAPTOR pada domain dokumen handbook institusional berbahasa Indonesia yang belum dieksplorasi pada penelitian terdahulu. Kedua, secara empiris, penelitian ini menganalisis kondisi karakteristik dokumen yang memengaruhi efektivitas peringkasan hierarkis RAPTOR pada domain handbook institusional, sebuah aspek yang belum banyak diuji pada korpus teks naratif. Analisis per-kategori pertanyaan pada Bab 4 digunakan untuk mengidentifikasi kondisi di mana pendekatan hierarkis efektif maupun kurang efektif, sehingga bermanfaat sebagai dasar pengembangan lebih lanjut.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana merancang *pipeline* sistem tanya jawab berbasis RAG yang mengintegrasikan *preprocessing* adaptif dokumen PDF multimodal dengan representasi hierarkis berbasis RAPTOR?
2. Bagaimana performa sistem yang diusulkan dibandingkan dengan pendekatan RAG konvensional dalam menjawab pertanyaan berdasarkan isi dokumen PDF multimodal, dengan handbook kampus sebagai studi kasus?

1.3 Batasan Permasalahan

Agar penelitian lebih terarah dan terfokus, batasan permasalahan dalam penelitian ini adalah sebagai berikut.

1. Dokumen PDF multimodal dalam penelitian ini dibatasi pada dokumen yang memuat teks, tabel, daftar, struktur bab dan subbab, lampiran, gambar yang mengandung teks, serta halaman hasil pemindaian. Informasi yang diproses dibatasi pada kandungan yang dapat direpresentasikan sebagai teks, baik melalui ekstraksi langsung maupun OCR; penelitian ini tidak mencakup analisis konten audio-visual.
2. Dataset yang digunakan berupa dokumen handbook kampus dalam format PDF, dengan dataset awal terdiri atas 121 dokumen dan total 2.698 halaman.
3. Proses *preprocessing* difokuskan pada ekstraksi teks, tabel, dan struktur dokumen, serta penerapan OCR apabila terdapat halaman hasil pemindaian maupun gambar yang memuat teks.
4. Evaluasi sistem difokuskan pada kualitas *retrieval* dan kualitas jawaban dengan metrik *Precision@k*, *Recall@k*, *Mean Reciprocal Rank* (MRR), *faithfulness*, dan *answer relevancy*.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut.

1. Merancang dan mengimplementasikan *pipeline* sistem tanya jawab berbasis RAG yang mengintegrasikan *preprocessing* adaptif dokumen PDF multimodal dengan representasi hierarkis berbasis RAPTOR.
2. Mengevaluasi performa sistem yang diusulkan dan membandingkannya dengan pendekatan RAG konvensional berdasarkan kualitas *retrieval* dan kualitas jawaban, dengan handbook kampus sebagai studi kasus.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut.

1. **Manfaat teoritis**, yaitu memberikan kontribusi terhadap pengembangan pendekatan RAG dengan *hierarchical retrieval* berbasis RAPTOR yang diintegrasikan dengan *preprocessing* adaptif pada sistem tanya jawab berbasis dokumen PDF multimodal.

2. **Manfaat praktis**, yaitu menyediakan rancangan sistem tanya jawab yang dapat membantu sivitas akademika dalam memperoleh informasi dari dokumen handbook kampus secara lebih efisien dan akurat. Rancangan sistem ini juga berpotensi diterapkan pada dokumen administratif institusional lain yang memiliki karakteristik serupa.

1.6 Sistematika Penulisan

Sistematika penulisan laporan penelitian ini disusun agar pembahasan dapat disampaikan secara terstruktur dan mudah dipahami. Adapun sistematika penulisannya adalah sebagai berikut.

- **Bab 1 PENDAHULUAN**
Bab ini berisi latar belakang masalah, rumusan masalah, batasan permasalahan, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.
- **Bab 2 LANDASAN TEORI**
Bab ini memuat tinjauan pustaka yang berkaitan dengan konsep-konsep utama penelitian, yaitu RAG, dokumen PDF multimodal, *preprocessing* adaptif dan OCR, representasi dokumen dan *embedding*, *hierarchical retrieval* dan RAPTOR, serta metrik evaluasi sistem. Bab ini juga membahas penelitian terdahulu yang relevan.
- **Bab 3 METODOLOGI PENELITIAN**
Bab ini menjelaskan metodologi penelitian, meliputi alur penelitian, karakteristik dataset, *pipeline preprocessing* adaptif, implementasi RAPTOR, proses *retrieval*, integrasi dengan model generatif, serta metode evaluasi.
- **Bab 4 HASIL DAN DISKUSI**
Bab ini berisi hasil implementasi sistem, pengujian, dan analisis perbandingan antara sistem yang diusulkan dengan pendekatan RAG konvensional berdasarkan metrik evaluasi yang telah ditentukan.
- **Bab 5 KESIMPULAN DAN SARAN**
Bab ini berisi kesimpulan dari penelitian yang telah dilakukan serta saran untuk pengembangan penelitian selanjutnya.