BAB I

PENDAHULUAN

1.1 Latar Belakang

Teknologi AI (*Artificial Intelligence*) saat ini mulai mengubah cara bisnis memproses dan menganalisis data [1], [2]. Perubahan ini paling terasa di sektor keuangan, di mana ketepatan angka sangat penting karena menyangkut kepercayaan investor dan kepatuhan terhadap regulasi. Akurasi data menjadi krusial karena kesalahan kecil dapat berdampak pada risiko hukum perusahaan dan hilangnya kepercayaan investor [3]. Namun, analisis keuangan yang menyeluruh membutuhkan pemeriksaan mendalam terhadap laporan tahunan yang biasanya sangat tebal. Proses ini memakan banyak waktu dan tenaga, bahkan untuk analis keuangan yang sudah berpengalaman sekalipun [4]. Oleh karena itu, kebutuhan akan sistem tanya-jawab otomatis yang dapat mengambil dan memverifikasi data spesifik dengan cepat dan akurat semakin meningkat [5]. Tujuannya adalah mengurangi pekerjaan manual yang melelahkan dan meminimalkan risiko kesalahan manusia yang bisa berakibat serius. Salah satu solusi teknis yang berkembang adalah metode RAG (*Retrieval-Augmented Generation*), yang memungkinkan sistem AI mengakses data eksternal yang terpercaya, bukan hanya mengandalkan hafalan dari parameter modelnya [6].

Pendekatan RAG saat ini banyak digunakan karena mampu membuat model AI mengambil informasi faktual dari sumber luar [6]. Namun, ketika diterapkan pada dokumen keuangan, metode RAG masih memiliki beberapa keterbatasan [7], [8]. Sebagai contoh, penelitian [9] melaporkan bahwa model bahasa standar hanya mencapai akurasi 61,24% pada tugas penalaran numerik keuangan, masih jauh di bawah performa ahli manusia yang mencapai 91,16%. Sementara itu, [10] menunjukkan bahwa pendekatan RAG konvensional kesulitan menjawab pertanyaan kompleks yang membutuhkan informasi dari beberapa dokumen sekaligus. Pertama, laporan tahunan perusahaan memiliki struktur dokumen yang rumit. Dokumen ini sering kali tidak terstandarisasi, mencampurkan teks panjang,

tabel angka yang padat, serta judul-judul yang bertingkat [9]. Akibatnya, sistem sering kesulitan mengambil informasi yang tepat dan kehilangan konteks penting [9], [10]. Kedua, sistem sering bingung membedakan identitas perusahaan dan tahun laporan. Penyebabnya, metode pencarian yang digunakan masih mengandalkan pencocokan kata kunci sederhana [5]. Karena itu, sistem bisa keliru mencampur laporan perusahaan induk dengan anak perusahaan, atau laporan tahun 2022 dengan tahun 2023 [5]. Ketiga, model AI cenderung mengalami halusinasi angka. Model sering mengarang angka yang terdengar masuk akal tetapi sebenarnya salah, sehingga tidak dapat diandalkan untuk tugas audit yang membutuhkan kepastian [11].

Dalam mengatasi kegagalan RAG konvensional dalam mengenali struktur pelaporan yang rumit tersebut, diperlukan sebuah arsitektur yang mampu memisahkan dan mengidentifikasi maksud kueri secara spesifik sebelum pencarian dokumen dimulai [6]. Hal ini memerlukan modul *Query Intent Parser* untuk mengekstrak entitas penting seperti kode saham dan tahun laporan secara deterministik [12]. Selanjutnya, pencarian data numerik yang tertanam di dalam matriks tabel tidak dapat hanya mengandalkan satu metode pencarian saja [6]. Sistem membutuhkan pendekatan hibrida yang menggabungkan kekuatan pencarian kata kunci eksak melalui indeks BM25 dan pencarian kedekatan makna semantik berbasis vektor menggunakan pustaka FAISS [13], [14]. Kedua sinyal pencarian yang berbeda skala ini kemudian harus diselaraskan menggunakan algoritma penggabungan peringkat seperti *Reciprocal rank fusion* (RRF) agar dokumen sumber yang relevan tidak tenggelam di tengah konteks dokumen yang panjang [13], [15]. Di sisi lain, mitigasi halusinasi angka pada tahap pemrosesan akhir menuntut adanya model bahasa dengan kapabilitas penalaran matematika yang kuat namun tetap efisien secara komputasi [16]. Oleh karena itu, pemilihan model bahasa berukuran kecil seperti Qwen-3 8B [16] yang dioptimasi menggunakan metode kuantisasi hemat parameter QLoRA [17] menjadi esensial, agar sistem audit otomatis ini dapat dijalankan secara presisi dan deterministik menggunakan perangkat keras kelas konsumen [17].

Dalam mengatasi keterbatasan tersebut, penelitian ini mengusulkan sebuah sistem bernama AuditRAG, yaitu sistem *Entity-Centric* Financial RAG yang dirancang khusus untuk menangani kompleksitas laporan keuangan perusahaan di Indonesia dengan mengintegrasikan seluruh komponen optimasi tersebut. Berbeda dengan pendekatan umum, sistem ini menggunakan metadata (seperti kode saham dan tahun laporan) untuk menyaring hasil pencarian agar lebih tepat sasaran. Sistem ini memiliki tiga komponen utama yang saling terhubung: *Query Intent Parser* (pemroses maksud pertanyaan), *Hybrid retrieval* dengan *Reciprocal rank fusion* (penggabungan dua metode pencarian), dan *Audit-Chain-of-Thought* (metode penalaran ala auditor) [12]. Komponen Parser bertugas mengenali kode saham perusahaan dan tahun laporan yang diminta pengguna sebelum proses pencarian dimulai. Langkah ini penting agar AI tidak mencampuradukkan dokumen antar perusahaan atau antar tahun. Selanjutnya, sistem pencarian tidak bekerja sendiri-sendiri, tetapi menggabungkan pencarian kata kunci (BM25) yang cepat dengan pencarian makna (vektor) yang lebih mendalam. Kombinasi ini bertujuan menemukan data yang tersembunyi di dalam tabel yang rumit, lalu memilih hasil terbaik untuk diproses lebih lanjut [13].

Sebagai lapisan pemeriksa terakhir, metode Audit-CoT diterapkan menggunakan model AI bernama Qwen-3 8B. Pendekatan ini memaksa AI untuk bernalar langkah demi langkah: menemukan baris tabel yang tepat, menyalin angkanya, memeriksa kebenarannya, baru kemudian memberikan jawaban akhir [18]. Mekanisme ini berfungsi sebagai pengaman untuk memverifikasi data dan mengurangi kesalahan dalam membaca angka. Dengan menyempurnakan proses pencarian dokumen dan pembuatan jawaban, sistem ini menunjukkan bahwa analisis keuangan berkualitas tinggi dapat dicapai secara efisien, bahkan hanya dengan perangkat keras komputer biasa sekalipun.

Kontribusi utama dari penelitian ini adalah:

a) Metodologis: Penelitian ini menggabungkan metode pencarian kata kunci (BM25) dan pencarian makna (FAISS) dengan teknik penggabungan RRF (*Reciprocal rank fusion*), serta menambahkan lapisan penyaringan

metadata berdasarkan kode saham dan tahun laporan untuk mengisolasi pencarian.

b) Empiris: Penelitian ini membandingkan kinerja model sebelum dan sesudah *fine-tuning*, dan menunjukkan peningkatan pada metrik *Consistency Score* (CS), *Matched Accuracy* (MA), dan *Audit Reasoning Quality* (ARQ) [13].
c) Efektivitas *Prompting* Audit-CoT: Eksperimental: Pengujian dilakukan pada 15 emiten BEI dengan 90 pertanyaan uji dari tahun 2023-2024 yang belum pernah dilihat model, menggunakan perangkat keras kelas konsumen (*low-resource*) [18].
d) Implementasi: Sistem ini diwujudkan dalam bentuk antarmuka web interaktif yang menampilkan proses penalaran (*reasoning trace*) dan sumber dokumen asli, sehingga analis keuangan dapat memverifikasi langsung jawaban yang diberikan.
e) *Dataset*: Penelitian ini membangun *dataset* instruksi QGA (*Question-Guidance-Answer*) sebanyak 360 pasang dari laporan keuangan 15 emiten BEI periode 2021-2024, yang dipisah secara ketat antara data latih (2021-2022) dan data uji (2023-2024).

Dengan demikian, penelitian ini fokus pada dua tantangan utama: kompleksitas struktur dokumen dan ambiguitas entitas, sementara halusinasi angka dilihat sebagai akibat dari kedua keterbatasan tersebut.

1.2 Rumusan Masalah

Transisi dari pemrosesan data tradisional menuju sistem berbasis AI dalam sektor keuangan menuntut tingkat akurasi yang absolut. Namun, kompleksitas laporan tahunan perusahaan di Indonesia sering kali menjadi penghambat bagi model bahasa standar. Berangkat dari tantangan tersebut, penelitian ini merumuskan masalah sebagai berikut:

1. Sejauh mana peningkatan kualitas jejak logika (*reasoning trace*) dan kepatuhan format *output* model Qwen-3 8B setelah *di-fine-tune* dengan metode Audit-CoT menggunakan teknik QLoRA?

2. Sejauh mana peningkatan stabilitas (*consistency score*) dan ketepatan nominal (*matched accuracy*) dari nilai akhir ekstraksi data numerik laporan keuangan setelah penerapan *fine-tuning* Audit-CoT dibandingkan dengan model *baseline*?
3. Sejauh mana efisiensi komputasi model setelah optimalisasi dengan QLoRA pada perangkat keras kelas konsumen, diukur dari waktu inferensi, konsumsi VRAM, dan jumlah *output token*?

1.3 Batasan Masalah

Agar penelitian ini tetap fokus, terarah, dan mendalam sesuai dengan tujuan yang telah ditetapkan, maka ruang lingkup penelitian dibatasi pada parameter berikut:

1. Batasan Kapabilitas Retrieval (Ruang Lingkup Data): Sistem AuditRAG yang dibangun tidak melakukan pencarian informasi secara daring (*real-time web search*/Google API) pada saat memproses kueri pengguna. Ruang lingkup pencarian sistem dibatasi murni secara *closed-domain* pada korpus data statis yang telah diunduh sebelumnya. Data tersebut berupa dokumen Laporan Tahunan (*Annual report*) dan Laporan Keuangan (*Financial statement*) format PDF dari 15 emiten sampel di Bursa Efek Indonesia (BEI) untuk periode tahun fiskal 2021–2024.
2. Model Generatif: Sistem berfokus menggunakan model dasar Qwen-3 8B yang dioptimasi secara efisien menggunakan teknik QLoRA (kuantisasi 4-bit) pada perangkat keras lokal, tanpa menggunakan API model bahasa berskala masif (seperti GPT-4).
3. Ruang Lingkup Teknis: Pengembangan arsitektur RAG mencakup penerapan *Entity-Centric Filtering* (berbasis *Ticker* dan Tahun) sebagai strategi isolasi ruang pencarian, pencarian dokumen hibrida (BM25 dan FAISS), serta *fine-tuning* dengan metode *Audit-Chain-of-Thought* (Audit-CoT).
4. Definisi Audit Finansial Otomatis: Konteks "audit" dalam penelitian ini dibatasi secara spesifik pada verifikasi dan ekstraksi presisi data numerik faktual (seperti laba bersih, total aset, pendapatan). Sistem bertindak sebagai

asisten yang memvalidasi apakah angka hasil ekstraksi model telah secara konsisten dan akurat merepresentasikan nilai asli yang tertera di dalam tabel/matriks laporan keuangan, bukan melakukan audit investigatif, kepatuhan hukum, atau audit pajak.

5. Fokus Evaluasi: Parameter pengujian menitikberatkan pada kualitas penalaran numerik (diukur dengan ARQ), konsistensi jawaban (diukur dengan CS, yaitu persentase kesamaan jawaban saat pertanyaan yang sama diajukan dua kali), akurasi ekstraksi angka (diukur dengan MA, yaitu ketepatan angka dan satuan mata uang), dan efisiensi komputasi (diukur dengan waktu inferensi dalam detik, konsumsi VRAM dalam GB, dan jumlah *output token*), bukan pada tugas pemrosesan bahasa alami umum seperti ringkasan teks atau analisis sentimen.
6. Bahasa yang Didukung: Sistem hanya memproses dokumen dan pertanyaan dalam Bahasa Indonesia dan Inggris, sesuai dengan korpus laporan keuangan emiten BEI yang umumnya bersifat bilingual. Bahasa lain seperti Arab, Mandarin, Jepang, atau semacamnya tidak didukung.

1.4 Tujuan dan Manfaat Penelitian

Penelitian ini dilakukan dengan tujuan untuk:

1.4.1 Tujuan Penelitian

1. Mengukur peningkatan kualitas jejak logika (*reasoning trace*) dan kepatuhan format output model Qwen-3 8B setelah di-*fine-tune* dengan metode Audit-Chain-of-Thought (Audit-CoT) menggunakan teknik QLoRA.
2. Mengukur peningkatan stabilitas (*consistency score*) dan ketepatan nominal (*matched accuracy*) dari nilai akhir ekstraksi data numerik laporan keuangan setelah penerapan *fine-tuning* Audit-CoT, dengan membandingkan hasil model yang sudah di-*fine-tune* terhadap model *baseline* (Qwen-3 8B tanpa *fine-tuning*).
3. Mengukur efisiensi komputasi model setelah optimalisasi dengan QLoRA pada perangkat keras kelas konsumen.

1.4.2 Manfaat Penelitian

1. Manfaat Teoritis: Memberikan kontribusi pada pengembangan literatur AI, khususnya mengenai bagaimana model bahasa kecil dapat dioptimalkan untuk penalaran numerik tingkat tinggi dalam domain spesifik (keuangan).
2. Manfaat Praktis: Menyediakan prototipe alat bantu bagi analis keuangan yang dapat dijalankan pada perangkat keras komputer standar (tanpa perlu server mahal), untuk melakukan ekstraksi dan verifikasi data laporan keuangan secara lebih cepat (waktu inferensi 22 detik per kueri) dan akurat (tingkat konsistensi 91%).

1.5 Sistematika Penulisan

Penulisan tugas akhir ini disusun secara sistematis ke dalam lima bab untuk memberikan gambaran alur penelitian yang komprehensif sebagai berikut:

BAB I: PENDAHULUAN

Bab ini berfungsi sebagai pengantar yang menguraikan fondasi penelitian. Di dalamnya dijelaskan mengenai urgensi integrasi AI dalam sektor keuangan, tantangan halusinasi pada model bahasa, serta rumusan masalah yang menjadi fokus studi. Selain itu, bab ini menetapkan batasan penelitian agar tetap relevan, serta merinci tujuan dan manfaat yang diharapkan, baik secara teoretis bagi pengembangan ilmu pengetahuan maupun secara praktis bagi industri finansial.

BAB II: LANDASAN TEORI

Bab ini menyajikan kerangka teoretis dan konsep fundamental yang mendasari penelitian. Penjelasan mencakup arsitektur LLM khususnya keluarga model Qwen-3, mekanisme kerja RAG, dan teknik optimasi PEFT. Bab ini juga membedah struktur laporan keuangan Indonesia, teori pencarian hibrida (*Hybrid retrieval*), serta konsep CoT yang dikembangkan menjadi metode Audit-CoT untuk menjaga integritas data numerik.

BAB III: METODOLOGI PENELITIAN

Bab ini menguraikan metodologi dan langkah-langkah teknis yang ditempuh untuk membangun sistem. Pembahasan dimulai dari tahapan pengumpulan data (laporan tahunan BEI), proses preprocessing dokumen, hingga perancangan arsitektur sistem yang meliputi Query Intent Parser dan Score Fusion. Dijelaskan pula desain eksperimen untuk implementasi Audit-CoT, teknik *prompting* yang digunakan, serta metrik evaluasi yang diterapkan untuk mengukur akurasi pengambilan data dan ketepatan kalkulasi numerik.

BAB IV: HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil dari implementasi sistem dan analisis mendalam terhadap performa yang dihasilkan. Di dalamnya dipaparkan perbandingan efektivitas antara RAG standar dengan sistem berbasis entitas dalam menangani query yang ambigu. Analisis juga difokuskan pada kemampuan model Qwen-3 8B dalam melakukan penalaran numerik menggunakan metode Audit-CoT. Selain itu, bab ini membahas efisiensi penggunaan perangkat keras tingkat konsumen dalam menjalankan model yang telah dioptimasi.

BAB V: PENUTUP

Bab terakhir ini berisi kesimpulan yang ditarik dari hasil pengujian dan pembahasan untuk menjawab rumusan masalah yang telah ditetapkan. Di dalamnya juga disampaikan saran-saran strategis bagi pengembangan penelitian selanjutnya, seperti penambahan cakupan data keuangan atau integrasi model yang lebih lanjut, guna menyempurnakan sistem RAG di masa depan.

UNIVERSITAS MULTIMEDIA NUSANTARA