BAB I
PENDAHULUAN

1.1. Latar Belakang

Di era digital saat ini, volume data global terus meningkat secara eksponensial hingga diproyeksikan mencapai 163 *zettabyte* pada 2025, meningkat hampir sepuluh kali lipat dibandingkan 2016 [1]. Pertumbuhan ini menimbulkan tantangan serius berupa *information overload* yang mendorong kebutuhan terhadap sistem *Information Retrieval* (IR) untuk menyeleksi dan mengorganisir informasi relevan dari korpus berskala besar [2]. Secara esensial, IR berfungsi menjembatani kesenjangan antara representasi informasi dalam suatu sistem dan kebutuhan kognitif pengguna yang diekspresikan melalui kueri [3]. Model IR klasik seperti *Term Frequency-Inverse Document Frequency* (TF-IDF) [4] dan Okapi BM25 [5] terbukti efektif untuk pencocokan leksikal, namun gagal mengenali relevansi antara istilah yang berbeda secara leksikal tetapi identik secara konseptual, keterbatasan ini sering kali dikenal sebagai *semantic gap* [6]. Keterbatasan ini mendorong pergeseran menuju *dense retrieval* berbasis *Pre-trained Language Models* (PLM) yang merepresentasikan teks dalam ruang semantik sehingga memungkinkan pencocokan makna dari kueri bukan sekadar kesamaan kata [6]. Kapasitas PLM dalam memahami bahasa secara mendalam selanjutnya menjadi fondasi kemunculan *Large Language Models* (LLM) yang menunjukkan performa luar biasa pada berbagai tugas *Natural Language Processing* (NLP) mulai dari *machine translation*, peringkasan teks, hingga *question answering* [7] namun menyimpan kelemahan fundamental yang membatasi keandalannya pada tugas yang menuntut akurasi tinggi.

Meskipun LLM menunjukkan performa yang luar biasa, model ini menyimpan kelemahan fundamental yang membatasi keandalannya pada tugas-tugas yang menuntut akurasi tinggi. Lebih jauh, model-model ini rentan terhadap halusinasi faktual akibat ketergantungan pada pengetahuan parametrik yang tersimpan selama pelatihan, terikat pada *knowledge cutoff*, dan sulit diverifikasi pada domain yang menuntut akurasi faktual tinggi seperti medis, hukum, dan sains

[8], [9]. Untuk memitigasi kelemahan ini, dikembangkanlah arsitektur *Retrieval-Augmented Generation* (RAG) yang mengintegrasikan memori parametrik LLM dengan memori *non*-parametrik dari korpus eksternal sehingga memungkinkan respons yang dapat dilacak ke sumber konkret [10]. Pendekatan ini terbukti meningkatkan landasan faktual, transparansi, dan kemampuan adaptasi terhadap pengetahuan baru, menjadikannya sangat sesuai untuk tugas-tugas kompleks seperti *open-domain question answering*, penalaran biomedis, dan dialog berbasis pengetahuan [11]. Namun demikian, RAG konvensional masih mengandung limitasi struktural yang signifikan karena sebagian besar implementasi mengandalkan pengambilan potongan teks pendek dan kontinu yang membatasi kemampuannya merepresentasikan struktur wacana pada dokumen panjang di mana informasi relevan tersebar di berbagai bagian [12]. Pendekatan *single-step retrieval* yang umum digunakan juga terbukti tidak memadai untuk pertanyaan kompleks yang membutuhkan integrasi informasi dari berbagai bagian dokumen sekaligus [13], dan hal ini menegaskan bahwa *knowledge selection* merupakan faktor primer yang menentukan kualitas keluaran sistem berbasis *retrieval* [14].

Salah satu pendekatan yang dikembangkan untuk mengatasi keterbatasan RAG konvensional adalah RAPTOR (*Recursive Abstractive Processing for Tree-Organized Retrieval*) [12] yang membangun struktur pohon hierarkis secara rekursif dari bawah ke atas (*bottom up*). Pada setiap *layer*, *node-node* hasil ringkasan dari *layer* sebelumnya menjadi *input* untuk proses *embedding*, *clustering*, dan *summarization* berikutnya hingga terbentuk hierarki abstraksi dari *leaf nodes* hingga *root nodes* [12]. Arsitektur ini memungkinkan sistem mengintegrasikan informasi yang tersebar di berbagai bagian dokumen panjang dengan mengambil konteks pada tingkat abstraksi yang sesuai dengan kebutuhan pertanyaan. RAPTOR mengevaluasi dua strategi *retrieval* utama yaitu *Tree Traversal* berbasis BFS (*Breadth-First Search*) [15] dengan seleksi top-k *node* per level dan *Collapsed Tree* yang memperlakukan seluruh *node* dari semua level secara sejajar dalam satu ruang pencarian [12]. Evaluasi pada NarrativeQA [16], QASPER [17], dan QuALITY [18] menunjukkan bahwa kombinasi RAPTOR dengan *Collapsed Tree* dan GPT-4 mencapai akurasi 82,6% pada QuALITY dan melampaui *state-of-the-art*

sebelumnya sebesar 20,3 poin absolut [12]. Namun fakta bahwa *Collapsed Tree* secara konsisten mengalahkan *Tree Traversal* berbasis BFS mengindikasikan bahwa mekanisme *traversal* RAPTOR itu sendiri bersifat suboptimal secara fundamental karena RAPTOR mencapai performa terbaiknya justru dengan mengabaikan struktur pohon yang dibangunnya sendiri [12]. Di samping itu, RAPTOR menerapkan batas jumlah *node* dan token yang seragam tanpa mempertimbangkan kompleksitas kueri sehingga berpotensi menghasilkan redundansi informasi dan beban komputasi yang tidak proporsional [12].

Menyadari keterbatasan mekanisme *traversal* RAPTOR, HIRO (*Hierarchical Information Retrieval Optimization*) mengusulkan strategi *traversal* yang lebih selektif dengan menggunakan algoritma DFS (*Depth-First Search*) [15] yang dikendalikan oleh dua parameter adaptif yaitu *selection threshold* (τ) sebagai ambang kemiripan semantik minimum agar sebuah *node* layak dieksplorasi dan *delta threshold* (δ) sebagai batas minimum penurunan skor kemiripan semantik antar *node* yang dapat ditoleransi pada perpindahan antar *layer* dalam proses *traversal* [19]. Kombinasi kedua parameter ini memungkinkan pemangkasan cabang yang dinilai tidak relevan secara semantik sehingga *traversal* menjadi lebih efisien dan hemat token dibandingkan *Collapsed Tree* yang memproses seluruh *node* secara sejajar tanpa diskriminasi. Secara empiris, HIRO menunjukkan hasil menjanjikan pada NarrativeQA dengan peningkatan ROUGE-L (*Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence*) dari 0,097 menjadi 0,121 yang mengonfirmasi bahwa strategi *pruning* adaptif efektif untuk dokumen naratif panjang [19]. Namun pada *dataset* QuALITY, metrik akurasi justru menurun dari 0,186 menjadi 0,182 dan F1 turun dari 0,480 menjadi 0,445 [19]. QuALITY sendiri merupakan *dataset* yang memang dirancang khusus untuk menguji penalaran mendalam atas isi dokumen secara menyeluruh, bukan sekedar *retrieval* berbasis kesamaan semantik sederhana. Hal ini tercermin dari QuALITY Hard, di mana bahkan mayoritas anotator manusia pun gagal menjawab dengan benar, menandakan *dataset* ini menuntut kapasitas inferensial yang jauh melampaui pencocokan leksikal maupun semantik sederhana [18]. Pola asimetris ini mengekspos keterbatasan fundamental bahwa *pruning* berbasis kemiripan semantik

statis secara sistematis membuang *node* yang leksikalnya jauh dari kueri namun krusial secara inferensial dalam rantai penalaran. Hal ini dipertegas oleh *benchmark* BRIGHT (*Benchmark for Reasoning-Intensive Retrieval* Tasks) yang secara khusus dirancang untuk menguji *retrieval* pada kueri yang membutuhkan penalaran mendalam, di mana pada model *retrieval* terbaik pada *leaderboard* MTEB (*Massive Text Embedding Benchmark*) [20] yang mencapai skor 59,0 nDCG@10 (*Normalized Discounted Cumulative Gain* pada 10 dokumen teratas) hanya mampu mencapai skor 18,3 nDCG@10 ketika dievaluasi pada BRIGHT [21], mengonfirmasi bahwa pendekatan *retrieval* berbasis kemiripan semantik statis secara inheren tidak memadai untuk tugas yang membutuhkan penalaran mendalam.

Mencermati kedua pendekatan secara menyeluruh, RAPTOR dan HIRO memiliki satu kelemahan struktural yang sama yaitu keputusan *traversal* pada setiap langkah sepenuhnya ditentukan oleh kemiripan semantik statis terhadap representasi kueri awal yang tidak berubah sepanjang proses penelusuran berlangsung. Tidak ada mekanisme yang memungkinkan sinyal *traversal* beradaptasi terhadap kebutuhan penalaran yang berkembang seiring eksplorasi pohon berjalan sehingga sistem tidak mampu membedakan antara *node* yang leksikalnya jauh dari kueri tetapi krusial secara inferensial dengan *node* yang tidak relevan karena keduanya diperlakukan identik oleh fungsi *scoring* statis. Kondisi ini diperparah oleh kenyataan bahwa pembaruan sinyal *retrieval* secara dinamis berbasis penalaran terbukti secara konsisten melampaui formulasi kueri statis pada tugas *question answering* yang kompleks [22], dan adaptasi strategi *retrieval* terhadap kompleksitas kueri juga menghasilkan peningkatan yang konsisten dibandingkan pendekatan *retrieval* yang seragam [23]. Namun demikian, sebagian besar penelitian RAG hingga saat ini masih berkutat pada *retrieval* berbasis kemiripan semantik tanpa menyentuh mekanisme *traversal* yang *reasoning-aware* pada struktur hierarkis, sehingga diperlukan pendekatan yang secara eksplisit mengaitkan proses *retrieval* dengan sinyal penalaran [24].

Bukti empiris dari IRCoT (*Interleaving Retrieval with Chain-of-Thought Reasoning*) menunjukkan bahwa melakukan *interlaving* langkah penalaran CoT

(*Chain-of-Thought*) dengan proses *retrieval* mampu meningkatkan *recall* hingga 21 poin, meningkatkan performa QA hingga 15 poin F1, dan mengurangi kesalahan faktual hingga 50% dibandingkan pendekatan *retrieval* statis [22], yang menegaskan bahwa *retrieval* dan *reasoning* sesungguhnya bisa saling memandu secara dinamis. Temuan serupa juga muncul pada *retrieval* berbasis graf, di mana HopRAG membuktikan bahwa menyuntikkan sinyal penalaran ke dalam proses *traversal node* secara konsisten meningkatkan akurasi pada tugas *multi-hop reasoning* dibandingkan pendekatan yang hanya mengandalkan kemiripan semantik [25]. Konvergensi bukti ini memperkuat keyakinan bahwa integrasi sinyal *reasoning* ke dalam mekanisme *retrieval* adalah arah yang valid dan menjanjikan, namun hingga kini pendekatan semacam itu belum pernah dicoba pada konteks *traversal* pohon hierarkis. Kesenjangan inilah yang mendorong penelitian ini untuk mengintegrasikan sinyal CoT langsung ke dalam fungsi *scoring traversal* DFS pada struktur pohon hierarkis, agar setiap keputusan *traversal* tidak lagi semata-mata bergantung pada kemiripan semantik statis, melainkan juga dipandu oleh konteks penalaran yang terus berkembang.

Berdasarkan motivasi tersebut, penelitian ini mengusulkan mekanisme *traversal* hierarkis yang *reasoning-aware* dengan mengintegrasikan sinyal *Chain-of-Thought* ke dalam fungsi *scoring node* pada setiap langkah DFS. Alih-alih hanya mengandalkan kemiripan semantik terhadap kueri awal, pendekatan yang diusulkan menggunakan *combined scoring function* yang secara simultan mempertimbangkan kemiripan *node* kandidat terhadap kueri awal dan terhadap *embedding* penalaran yang terbentuk dari konteks yang terakumulasi pada langkah-langkah *traversal* sebelumnya. Mekanisme ini dipadukan dengan *adaptive dual-threshold pruning* yang diselingi dengan generasi CoT, sehingga pada setiap langkah DFS sistem terlebih dahulu menghasilkan langkah penalaran baru berdasarkan konteks yang telah terkumpul, memperbarui *embedding reasoning*, lalu baru memutuskan apakah suatu cabang layak dieksplorasi lebih lanjut atau dipangkas. Dengan cara ini, berbeda dari RAPTOR dan HIRO yang mengandalkan representasi kueri statis sepanjang *traversal*, sinyal *traversal* dalam sistem yang diusulkan terus berevolusi seiring penalaran berkembang, sehingga *node* yang leksikalnya jauh dari kueri awal

namun relevan terhadap arah inferensi yang sedang terbentuk tetap dapat teridentifikasi dan dieksplorasi.

Pendekatan ini dievaluasi secara komprehensif pada empat *dataset* yang dipilih secara komplementer, yaitu NarrativeQA, QuALITY, QASPER, dan TyDi QA. NarrativeQA [16] merupakan *dataset* pemahaman bacaan yang mengharuskan sistem menjawab pertanyaan berdasarkan keseluruhan buku atau naskah film, menuntut pemahaman naratif yang tidak dapat diselesaikan melalui pencocokan lokal semata. Karakteristik ini menjadikannya tolok ukur yang tepat untuk mengevaluasi kemampuan *traversal* dalam mengintegrasikan informasi yang tersebar pada dokumen panjang. QuALITY [18] adalah *dataset multiple-choice* dengan konteks panjang rata-rata 5.000 token di mana pertanyaan ditulis oleh pembaca yang telah menelaah keseluruhan teks. Separuh pertanyaannya termasuk dalam *subset* HARD yang dirancang khusus agar tidak dapat dijawab hanya dengan menemukan satu atau beberapa bagian teks tertentu, melainkan menuntut pemahaman dan interpretasi atas keseluruhan dokumen secara holistik. *Dataset* ini digunakan untuk mengevaluasi kemampuan sistem dalam menangani dokumen panjang yang jawabannya tidak terlokalisasi pada satu bagian tertentu. QASPER [17] menyediakan pasangan pertanyaan dan jawaban berbasis makalah penelitian NLP, di mana setiap pertanyaan dirumuskan hanya berdasarkan judul dan abstrak namun harus dijawab dengan bukti dari seluruh isi makalah, konfigurasi ini secara khusus menguji kemampuan sistem melakukan *reasoning* berbasis pada dokumen ilmiah terstruktur. TyDi QA [26], khususnya *split* bahasa Indonesia, digunakan untuk mengevaluasi ketangguhan sistem dalam konteks lintas bahasa, mengingat bahasa Indonesia memiliki karakteristik morfologi dan struktur semantik yang berbeda dari bahasa Inggris. Penelitian ini memberikan tiga kontribusi utama yaitu pertama sebuah *combined scoring function* yang mengintegrasikan kemiripan semantik kueri dan *embedding* penalaran CoT untuk evaluasi *node* adaptif di setiap langkah DFS, kedua mekanisme *adaptive dual-threshold pruning* yang diselingi dengan generasi CoT untuk kontrol *traversal* yang dinamis, dan ketiga analisis empiris yang mengidentifikasi kualitas sinyal *reasoning* sebagai *bottleneck* utama

performa *traversal* hierarkis pada pertanyaan sulit sekaligus memberikan arah yang jelas bagi optimasi sistem RAG di masa mendatang.

1.2. Rumusan Masalah

Rumusan masalah yang menjadi fokus utama pada penelitian ini adalah sebagai berikut:

1. Bagaimana merancang mekanisme *tree traversal* berbasis IRCoT untuk mengoptimasi proses penelusuran pada arsitektur RAG hierarkis?
2. Bagaimana efektivitas mekanisme *traversal* yang diusulkan dibandingkan metode *baseline* pada *dataset* NarrativeQA, QuALITY, QASPER, dan TyDi QA?
3. Bagaimana efisiensi komputasi mekanisme *traversal* yang diusulkan dibandingkan metode *baseline* ditinjau dari jumlah token, latensi, dan jumlah *node* yang ditelusuri pada *dataset* NarrativeQA, QuALITY, QASPER, dan TyDi QA?

1.3. Batasan Masalah

Batasan yang ditetapkan pada masalah yang diamati pada penelitian ini adalah sebagai berikut:

1. Penelitian ini berfokus pada perancangan dan evaluasi mekanisme *traversal reasoning-aware* sebagai optimasi di sisi *retrieval* pada arsitektur RAG hierarkis. Penelitian tidak mencakup perancangan arsitektur LLM baru maupun *fine-tuning* model bahasa.
2. Penelitian tidak mencakup optimasi arsitektur model, *fine-tuning*, maupun eksplorasi infrastruktur komputasi di luar konfigurasi yang ditetapkan. Eksperimen juga tidak mencakup domain di luar *dataset* yang telah ditentukan.
3. Optimasi yang dimaksud dalam penelitian ini berfokus pada peningkatan efektivitas, yaitu kualitas jawaban, bukan efisiensi komputasi, yaitu waktu eksekusi atau jumlah *node* yang ditelusuri. Metrik efisiensi tetap dilaporkan sebagai bagian dari analisis *trade-off*, namun bukan menjadi tujuan utama optimasi.

4. Sistem hanya memproses input berupa teks. Konten nontekstual seperti gambar, tabel, dan formula matematika berada di luar ruang lingkup penelitian ini.
5. Model yang digunakan dalam penelitian ini, antara lain, *multilingual-e5-large-instruct* sebagai model *embedding*, *Qwen2.5-7B-Instruct-AWQ* sebagai model generasi ringkasan dan jawaban melalui inferensi vLLM, serta Qdrant sebagai basis data vektor. Model-model ini digunakan tanpa modifikasi arsitektural.
6. Evaluasi dilakukan pada empat *dataset*, yaitu NarrativeQA, QuALITY, QASPER, dan TyDi QA *split* bahasa Indonesia.

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk menjawab rumusan masalah serta memberikan arahan yang lebih jelas bagi penelitian ini. Adapun tujuan penelitian ini adalah sebagai berikut:

1. Merancang mekanisme *tree traversal* berbasis IRCoT untuk mengoptimasi proses penelusuran pada arsitektur RAG hierarkis.
2. Menganalisis efektivitas mekanisme *traversal* yang diusulkan dibandingkan metode *baseline* pada *dataset* NarrativeQA, QuALITY, QASPER, dan TyDi QA.
3. Menganalisis efisiensi komputasi mekanisme *traversal* yang diusulkan dibandingkan metode baseline ditinjau dari jumlah token, latensi eksekusi, dan jumlah *node* yang ditelusuri pada *dataset* NarrativeQA, QuALITY, QASPER, dan TyDi QA.

1.5. Manfaat Penelitian

Manfaat penelitian yang dapat diperoleh dari hasil penelitian ini adalah sebagai berikut:

1. Bagi bidang keilmuan

   Penelitian ini berkontribusi pada bidang IR dan NLP dengan memperkenalkan metode *traversal* hierarkis berbasis sinyal penalaran eksplisit untuk mengatasi keterbatasan kemiripan semantik statis. Analisis empiris yang disajikan juga

berhasil memetakan kualitas *reasoning* sebagai *bottleneck* performa pada dokumen kompleks, memberikan fondasi baru bagi pengembangan arsitektur RAG di masa depan.

2. Bagi masyarakat

   Penelitian ini menghasilkan sistem RAG yang lebih transparan dan akurat dalam mengekstraksi informasi dari dokumen panjang seperti literatur medis atau dokumen hukum. Hal ini memudahkan masyarakat dalam memperoleh jawaban yang dapat diverifikasi langsung ke sumber aslinya, sehingga meningkatkan literasi dan kepercayaan terhadap informasi berbasis kecerdasan buatan.

3. Bagi Pemerintah

   Penelitian ini menyediakan solusi teknis untuk meningkatkan akuntabilitas layanan publik digital melalui sistem tanya jawab kebijakan yang dapat dilacak. Implementasi teknologi ini mendukung penyediaan informasi peraturan perundang-undangan yang lebih cepat dan transparan, sejalan dengan upaya pemerintah dalam mengoptimalkan layanan digital nasional.

Halaman ini sengaja dikosongkan