BAB V
PENUTUP

5.1. Kesimpulan

Berdasarkan penelitian yang telah dilakukan mengenai optimasi *tree traversal* berbasis *Interleaving Chain-of-Thought* pada arsitektur *Retrieval-Augmented Generation* hierarkis, diperoleh beberapa kesimpulan sebagai berikut:

1. Perancangan mekanisme *tree traversal* berbasis IRCoT pada arsitektur RAG hierarkis berhasil diwujudkan melalui integrasi sinyal penalaran ke dalam algoritma DFS. Terdapat tiga komponen utama yang dibangun, yaitu fungsi skor gabungan yang menggabungkan kemiripan terhadap kueri awal dan terhadap *embedding* penalaran yang berkembang sepanjang *traversal*, pemangkasan adaptif dengan dua ambang batas (*selection_t* dan *delta_t*) untuk menentukan apakah suatu *node* dipangkas, di-*backtrack*, atau dieksplorasi, serta generasi penalaran yang diselingi di setiap langkah sebelum evaluasi *node* turunan. Parameter optimal hasil studi ablasi bervariasi antar *dataset*, misalnya QASPER memerlukan bobot penalaran yang besar karena sinyal penalaran sangat membantu, sedangkan NarrativeQA dan TyDi QA memerlukan bobot kueri yang lebih dominan.
2. Efektivitas mekanisme yang diusulkan bersifat asimetris tergantung karakteristik *dataset*. Pada NarrativeQA, mode *reasoning* meningkatkan ROUGE-L dari 0,1144 menjadi 0,1275 dan Token F1 dari 0,1043 menjadi 0,1191. Pada QASPER, mode *reasoning* meningkatkan *Answer* F1 dari 0,3135 menjadi 0,3288, meskipun *Evidence* F1 turun dari 0,1671 menjadi 0,1353. Sebaliknya, pada QuALITY, mode *reasoning* menurunkan akurasi total dari 54,98% menjadi 54,49% dan pada *subset hard* dari 44,75% menjadi 44,55%. Pada TyDi QA, mode *reasoning* menghasilkan Token F1 lebih rendah yaitu 0,3862 berbanding 0,3922, dengan *Exact Match* identik di angka 0,1265. IRCoT efektif untuk *dataset* dengan informasi tersebar dan struktur kausal seperti NarrativeQA dan QASPER, tetapi tidak efektif untuk *dataset* yang

menuntut pemahaman holistik seperti QuALITY maupun untuk *dataset* pendek yang menghasilkan pohon dangkal seperti TyDi QA.

3. Efisiensi komputasi mode *reasoning* secara konsisten lebih lambat daripada *baseline* HIRO (*non-reasoning*) pada seluruh *dataset*. Total waktu *pipeline reasoning* pada NarrativeQA mencapai 20,70 detik, sementara *non-reasoning* hanya 4,84 detik, sehingga *reasoning* 4,28 kali lebih lambat. Pada QASPER waktu *reasoning* membutuhkan 10,43 detik berbanding 2,63 detik pada *non-reasoning*, atau 3,97 kali lebih lambat. Pada QuALITY waktu *reasoning* mencapai 10,69 detik sedangkan *non-reasoning* 4,23 detik, atau 2,53 kali lebih lambat. Pada TyDi QA waktu *reasoning* mencapai 5,06 detik sementara *non-reasoning* 2,43 detik, atau 2,08 kali lebih lambat. Total *token cost* juga meningkat drastis. Pada NarrativeQA mencapai 13.266 token sedangkan *non-reasoning* 1.934 token, atau 7 kali lipat. Pada QASPER *token cost reasoning* 8.007 token berbanding 2.703 token pada *non-reasoning*, atau 3 kali lipat. Pada QuALITY *token cost reasoning* 5.931 token sementara *non-reasoning* 1.836 token, atau 3,2 kali lipat. Pada TyDi QA *token cost reasoning* 3.334 token sedangkan *non-reasoning 2*.281 token, atau 1,5 kali lipat. Jumlah *node* yang dikumpulkan pada mode *reasoning* meningkat paling signifikan di NarrativeQA, yaitu dari 8,6 *node* menjadi 50,1 *node*. Dengan demikian, *trade-off* efektivitas dan efisiensi terjustifikasi pada NarrativeQA dan QASPER karena peningkatan performa sebanding dengan tambahan biaya, tetapi tidak terjustifikasi pada QuALITY dan TyDi QA karena tidak ada peningkatan efektivitas yang berarti.
4. Pada NarrativeQA, peningkatan pada METEOR dan Token F1 signifikan secara statistik dengan *p-value* masing-masing 0,005 dan 0,037. Pada QASPER, penurunan *Evidence* F1 sangat signifikan dengan *p-value* $7{,}10\times10^{-42}$, sementara peningkatan *Answer* F1 tidak signifikan dengan *p-value* 0,097. Pada TyDi QA dan QuALITY, seluruh perbedaan yang teramati tidak signifikan secara statistik, dengan *p-value* jauh di atas 0,05. Hal ini memperkuat kesimpulan bahwa mekanisme *reasoning* memberikan keunggulan yang meyakinkan hanya pada NarrativeQA untuk metrik

METEOR dan Token F1, serta pada QASPER untuk pola pergeseran bukti, tetapi tidak memberikan keunggulan statistik pada QuALITY maupun TyDi QA.

5.2. Saran Pengembangan

Berdasarkan hasil penelitian dan keterbatasan yang ditemukan, berikut adalah saran untuk pengembangan lebih lanjut:

1. Efisiensi komputasi pada tahap *traversal* masih perlu dioptimalkan mengingat waktu pemrosesan pada mode *reasoning* mencapai hingga 20,70 detik per kueri di NarrativeQA. Beberapa optimasi yang dapat diterapkan seperti *early stopping* ketika skor gabungan suatu *node* telah melampaui batas keyakinan tertentu.
2. Penelitian selanjutnya dapat mengadaptasi strategi *traversal* untuk *dataset* dengan struktur pohon dangkal seperti TyDi QA yang memiliki kedalaman rata-rata hanya 1,81 *layer* sehingga mekanisme *reasoning* tidak memberikan manfaat. Sistem dapat dirancang untuk secara adaptif melewati proses penalaran jika kedalaman pohon di bawah ambang tertentu atau jika jumlah *node* yang tersedia terbatas, sehingga komputasi tidak terbuang sia-sia.
3. Untuk *dataset* dengan pertanyaan holistik seperti QuALITY yang menuntut pemahaman seluruh dokumen, pendekatan penalaran inkremental terbukti kurang efektif. Penelitian lanjutan disarankan mengembangkan strategi *global-first reasoning*, yaitu sistem meringkas seluruh dokumen dari *root node* terlebih dahulu sebelum melakukan eksplorasi ke bagian yang lebih spesifik.
4. Penelitian ini terbatas pada pemrosesan teks, padahal dokumen di dunia nyata seperti makalah ilmiah atau artikel Wikipedia sering menyertakan gambar, tabel, dan diagram yang mengandung informasi penting. Pengembangan ke depan disarankan untuk mengintegrasikan model *vision language* ke dalam arsitektur, sehingga *node* dapat merepresentasikan konten visual sekaligus

Halaman ini sengaja dikosongkan