BAB I
PENDAHULUAN

Bab ini menjelaskan latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah, serta sistematika penulisan yang berkaitan dengan penelitian mengenai analisis perbandingan kinerja dan efisiensi metode FFT, LoRA, dan KD-LoRA pada model IndoT5 untuk peringkasan berita bahasa Indonesia.

1.1 Latar Belakang

Pada era digital saat ini, dunia menghadapi peningkatan produksi data yang sangat besar atau dikenal sebagai fenomena ledakan data (*data explosion*). Transformasi digital pada berbagai sektor seperti media sosial, *e-commerce*, dan portal berita daring menyebabkan informasi tekstual diproduksi dalam jumlah sangat besar setiap hari, sehingga menimbulkan kelebihan informasi (*information overload*) bagi manusia karena kemampuan membaca dan menyerap informasi tidak sebanding dengan kecepatan produksi informasi (Kaushal dkk., 2024). Oleh karena itu, diperlukan teknologi yang mampu merangkum informasi secara otomatis melalui *Automatic Text Summarization*.

Dalam pengembangan sistem peringkasan teks, terdapat dua pendekatan utama yaitu ekstraktif dan abstraktif. Pendekatan ekstraktif menghasilkan ringkasan dengan memilih kalimat penting dari teks sumber, sedangkan pendekatan abstraktif menghasilkan ringkasan dengan menyusun kalimat baru yang merepresentasikan isi dokumen secara lebih ringkas dan koheren (Kurniawan & Louvan, 2018). Pendekatan abstraktif dinilai lebih baik karena mampu menghasilkan ringkasan yang lebih padat, koheren, dan mudah dibaca, sehingga pendekatan ini menjadi fokus dalam penelitian ini.

Perkembangan *Natural Language Processing* mengalami kemajuan pesat dengan diperkenalkannya arsitektur Transformer (Vaswani dkk., 2017) yang melahirkan berbagai *pre-trained language model* berbasis *encoder-decoder* seperti BART, PEGASUS, dan T5. BART menggunakan metode *denoising* untuk memulihkan teks yang sengaja dirusak, sehingga unggul dalam mempertahankan koherensi semantik ringkasan. Di sisi lain, PEGASUS dirancang khusus untuk peringkasan abstraktif dengan menyembunyikan kalimat-kalimat penting dalam dokumen dan melatih model untuk memprediksinya.

Sementara itu, Raffel dkk. (2020) memperkenalkan T5 dengan pendekatan *text-to-text* yang menyatukan seluruh tugas pemrosesan bahasa ke dalam satu format seragam, didukung arsitektur *encoder-decoder* dan pelatihan berbasis *denoising* yang terbukti menghasilkan performa tinggi pada berbagai tugas NLP.

Di antara model-model berbasis *encoder-decoder* tersebut, T5 secara konsisten menunjukkan performa yang unggul dalam tugas peringkasan. Bharathi Mohan dkk. (2024) membandingkan T5 *fine-tuned*, PEGASUS, ProphetNet, dan BART menggunakan lima metrik evaluasi pada berbagai dataset (CNN/DailyMail, XLSum, WikiHow, dan DUC2004) dan menemukan bahwa T5 *fine-tuned* mengungguli seluruh model lainnya pada kelima metrik tersebut dengan ROUGE-1 sebesar 0,232. Dharrao dkk. (2024) juga menunjukkan hal serupa pada dataset berita bisnis BBC, di mana T5 meraih ROUGE-1 tertinggi sebesar 0,354 dibandingkan BART maupun PEGASUS. Bhattacharya dkk. (2026) pada domain medis juga mengonfirmasi bahwa T5-base mencapai METEOR tertinggi sebesar 0,39 dan paling disukai oleh evaluator manusia dibandingkan model lainnya.

Keunggulan arsitektur T5 ini kemudian diadaptasi untuk konteks bahasa Indonesia melalui pengembangan IndoT5, sebuah model monolingual yang dilatih khusus pada korpus bahasa Indonesia yang diekstraksi dan difilter secara ketat dari dataset multilingual mC4 sehingga mampu memahami karakteristik linguistik bahasa Indonesia dengan lebih baik dibandingkan model multilingual seperti mT5 (Wikidepia, 2021). Putri dkk. (2025) menunjukkan bahwa IndoT5 mampu mengungguli mBART dengan ROUGE-1 sebesar 0,4309 pada dataset Tempo.co, menegaskan potensi model monolingual untuk tugas peringkasan bahasa Indonesia.

Meskipun model IndoT5 memiliki kemampuan pemahaman bahasa Indonesia yang baik, penerapannya di dunia nyata masih menghadapi kendala biaya komputasi. Metode adaptasi model yang umum digunakan adalah *Full Fine-Tuning* (FFT) yang mengharuskan pembaruan seluruh parameter model yang pada IndoT5-Base berjumlah sekitar 222 juta parameter. Proses *fine-tuning* pada model berukuran besar membutuhkan memori GPU yang tinggi serta menimbulkan tantangan dalam penyimpanan dan *deployment* model (Hu dkk., 2022). Selain itu, model berukuran besar juga menyebabkan proses inferensi menjadi kurang efisien untuk implementasi pada sistem nyata yang membutuhkan respons cepat (Azimi dkk., 2024).

Sebagai solusi terhadap permasalahan tersebut, dikembangkan pendekatan *Parameter-Efficient Fine-Tuning* (PEFT) seperti LoRA (*Low-Rank Adaptation*) yang hanya melatih sebagian kecil parameter model sehingga mengurangi kebutuhan penyimpanan model (Hu dkk., 2022). Hu dkk. (2022) menunjukkan bahwa LoRA mampu mengurangi jumlah parameter terlatih hingga 10.000 kali lipat dan kebutuhan memori GPU hingga 3 kali lipat dibandingkan *fine-tuning* penuh pada GPT-3 175B. Adilova dkk. (2024) mencatat bahwa FFT pada mT5-Base memerlukan pelatihan 582,4 juta parameter, sedangkan penerapan LoRA pada model yang sama hanya membutuhkan sekitar 14 hingga 56 juta parameter, tergantung konfigurasi yang digunakan. Lebih lanjut, Adilova dkk. (2024) menunjukkan bahwa LoRA pada beberapa konfigurasi mampu mencapai bahkan melampaui performa model dengan *fine-tuning* penuh dalam skor ROUGE.

Dalam konteks bahasa Indonesia, Muhammad (2025) membandingkan beberapa arsitektur model untuk peringkasan teks pada dataset IndoSum dan menemukan bahwa IndoT5 menghasilkan performa tertinggi dibandingkan model lain seperti IndoBART dan IdT5, dengan ROUGE-1 sebesar 0,4201. Meskipun demikian, perbedaan arsitektur model, strategi prapemrosesan, serta konfigurasi eksperimen menyebabkan hasil tersebut tidak dapat dibandingkan secara langsung dengan penelitian FFT sebelumnya seperti Satya dkk. (2025) dan Nasution dkk. (2025), sehingga perbandingan langsung antara FFT dan LoRA dalam satu kerangka eksperimen yang sama masih belum dilakukan.

Selain itu, meskipun LoRA efisien saat pelatihan dengan hanya memperbarui matriks dekomposisi berperingkat rendah (Hu dkk., 2022), model yang di-*deploy* tetap berukuran sama dengan model aslinya sehingga tidak menghasilkan penghematan saat inferensi. Oleh karena itu, diperlukan pendekatan yang tidak hanya mengurangi parameter terlatih, tetapi juga menghasilkan model yang lebih kecil dan lebih cepat saat inferensi.

Untuk mengatasi permasalahan tersebut, penelitian ini mengadaptasi metode KD-LoRA yang diusulkan oleh Azimi dkk. (2024), yaitu pendekatan yang mengombinasikan *Knowledge Distillation* dengan LoRA. Pada metode ini, model besar berperan sebagai *teacher* dan mentransfer pengetahuan ke model yang lebih kecil sebagai *student* melalui *soft targets*. Azimi dkk. (2024) menunjukkan bahwa KD-LoRA mampu mempertahankan sekitar 97% performa *Full Fine-Tuning* dan 98% performa LoRA, dengan penggunaan VRAM 30–75% lebih hemat serta waktu inferensi 30% lebih cepat, namun penelitian tersebut masih

difokuskan pada tugas pemahaman bahasa. Penerapannya pada tugas generatif seperti peringkasan abstraktif, khususnya untuk bahasa Indonesia, belum pernah dilakukan.

Untuk mengevaluasi ketiga metode tersebut secara adil dalam satu kerangka eksperimen yang sama, penelitian ini menggunakan dataset IndoSum. Dataset ini dipilih karena merupakan dataset standar untuk peringkasan teks bahasa Indonesia yang ringkasannya disusun secara manual oleh penutur asli bahasa Indonesia, sehingga data yang digunakan bersifat valid dan reliabel (Kurniawan & Louvan, 2018). Selain itu, penggunaan dataset yang sama dengan penelitian sebelumnya memungkinkan hasil penelitian ini dibandingkan secara langsung dengan studi-studi terdahulu, seperti Nasution dkk. (2025) dengan ROUGE-1 sebesar 0,7126 dan Satya dkk. (2025) dengan ROUGE-1 sebesar 0,7352.

Berdasarkan penelitian terdahulu yang menggunakan FFT (Nasution dkk., 2025; Satya dkk., 2025), LoRA (Muhammad, 2025), serta KD-LoRA (Azimi dkk., 2024), masing-masing metode menunjukkan kelebihan dan kekurangan dari sisi kualitas model maupun efisiensi komputasi. Penelitian yang membandingkan secara menyeluruh kualitas ringkasan dan efisiensi komputasi antara FFT, LoRA, dan KD-LoRA pada tugas generatif bahasa Indonesia dalam satu kerangka eksperimen yang sama masih sangat terbatas. Oleh karena itu, penelitian ini bertujuan untuk menganalisis perbandingan kinerja dan efisiensi komputasi metode FFT, LoRA, dan KD-LoRA pada model IndoT5 untuk tugas peringkasan berita bahasa Indonesia menggunakan dataset IndoSum.

1.2 Rumusan Masalah

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

1. Bagaimana perbandingan kualitas ringkasan antara model IndoT5 yang dilatih menggunakan metode *Full Fine-Tuning* (FFT), *Low-Rank Adaptation* (LoRA), dan *Knowledge Distillation* dengan LoRA (KD-LoRA) pada dataset IndoSum, ditinjau dari metrik leksikal (ROUGE), metrik semantik (BERT*Score*), serta tingkat abstraksi ringkasan (*Novelty*)?
2. Seberapa besar efisiensi komputasi yang ditawarkan oleh metode KD-LoRA dibandingkan *Full Fine-Tuning* dan LoRA standar, ditinjau dari aspek penggunaan sumber daya (*peak* VRAM, jumlah parameter, ukuran *checkpoint*), kecepatan komputasi (waktu pelatihan dan waktu inferensi), serta kompleksitas model (GFLOPs)?

1.3 Tujuan dan Manfaat Penelitian

Bagian ini menjelaskan tujuan yang ingin dicapai melalui penelitian serta manfaat yang diharapkan dari hasil penelitian bagi pengembangan ilmu pengetahuan dan penerapannya dalam bidang pemrosesan bahasa alami.

1.3.1 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan, tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Menganalisis perbandingan kualitas ringkasan antara model IndoT5 yang dilatih menggunakan metode *Full Fine-Tuning* (FFT), *Low-Rank Adaptation* (LoRA), dan *Knowledge Distillation* dengan LoRA (KD-LoRA) pada dataset IndoSum berdasarkan metrik leksikal (ROUGE), metrik semantik (BERT*Score*), serta tingkat abstraksi (*Novelty*).
2. Mengukur seberapa besar efisiensi komputasi yang dihasilkan oleh metode KD-LoRA dibandingkan dengan *Full Fine-Tuning* dan LoRA standar, dengan fokus pada penghematan sumber daya (*peak* VRAM, jumlah parameter, ukuran *checkpoint*), percepatan komputasi (waktu pelatihan dan waktu inferensi), dan kompleksitas model (GFLOPs).

1.3.2 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan kontribusi dalam mengatasi permasalahan *information overload* dengan menghasilkan model peringkasan teks berita bahasa Indonesia melalui penerapan metode *Full Fine-Tuning*, LoRA, dan KD-LoRA pada model IndoT5 menggunakan dataset IndoSum. Selain itu, penelitian ini memberikan analisis perbandingan kinerja dan efisiensi komputasi dari ketiga metode tersebut sehingga memberikan gambaran nyata mengenai *trade-off* yang perlu dipertimbangkan dalam pengembangan sistem peringkasan teks otomatis.

Hasil penelitian ini selanjutnya diharapkan dapat menjadi referensi dan acuan bagi penelitian selanjutnya dalam pengembangan model peringkasan teks bahasa Indonesia yang lebih efisien dari sisi komputasi namun tetap mempertahankan kualitas ringkasan, serta bagi pengembang sistem atau praktisi dalam memilih metode *fine-tuning* yang tepat untuk

implementasi sistem peringkasan teks otomatis pada lingkungan dengan keterbatasan sumber daya komputasi.

1.4 Batasan Masalah

Agar pembahasan lebih terarah dan tidak meluas dari topik penelitian, maka ditetapkan batasan masalah sebagai berikut:

1. Penelitian ini difokuskan pada tugas peringkasan teks abstraktif pada dokumen berita berbahasa Indonesia menggunakan dataset IndoSum.
2. Penerapan metode *Knowledge Distillation* dilakukan dengan menggunakan model IndoT5-Base hasil *Full Fine-Tuning* sebagai model *teacher* yang mentransfer pengetahuan kepada model IndoT5-Small berbasis LoRA sebagai model *student*.
3. Pemilihan konfigurasi model terbaik dilakukan melalui eksperimen *grid search* terbatas pada kombinasi *hyperparameter* yang telah ditentukan untuk masing-masing metode pelatihan.
4. Evaluasi kualitas dan karakteristik ringkasan dibatasi pada metrik otomatis, yaitu ROUGE, BERT*Score*, dan *Novelty* untuk mengukur kesamaan leksikal, kesamaan semantik, dan tingkat abstraksi ringkasan.
5. Evaluasi efisiensi komputasi dibatasi pada pengukuran *peak* VRAM, waktu pelatihan, jumlah parameter, ukuran *checkpoint* model, waktu inferensi, serta kompleksitas komputasi (GFLOPs).

1.5 Sistematika Penulisan

Untuk memberikan gambaran menyeluruh mengenai alur pembahasan dalam skripsi ini, laporan tugas akhir disusun dengan sistematika penulisan sebagai berikut:

BAB I PENDAHULUAN

Bab ini menjelaskan latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah, serta sistematika penulisan yang berkaitan dengan penelitian mengenai analisis perbandingan kinerja dan efisiensi metode FFT, LoRA, dan KD-LoRA pada model IndoT5 untuk peringkasan berita Indonesia.

BAB II LANDASAN TEORI

Bab ini menjelaskan tinjauan pustaka dan landasan teori yang mendasari penelitian, meliputi konsep *Natural Language Processing*, peringkasan teks otomatis, prapemrosesan dan tokenisasi data teks, serta arsitektur Transformer. Selain itu, dibahas model *Text-to-Text Transfer Transformer* (T5) dan IndoT5, serta metode pelatihan model yang mencakup *Full Fine-Tuning*, *Parameter-Efficient Fine-Tuning*, *Low-Rank Adaptation*, *Knowledge Distillation*, dan KD-LoRA. Bab ini juga menjelaskan konfigurasi pelatihan dan *hyperparameter*, strategi optimisasi pelatihan, proses inferensi dan *beam search*, serta metrik evaluasi yang digunakan untuk mengukur kinerja dan efisiensi model.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan metodologi penelitian untuk membandingkan performa dan efisiensi metode *Full Fine-Tuning* (FFT), *Low-Rank Adaptation* (LoRA), dan KD-LoRA pada model IndoT5 untuk tugas peringkasan bahasa Indonesia. Alur penelitian mencakup tahapan pengumpulan data, persiapan dan pembagian data, prapemrosesan data, tokenisasi, pembentukan model, serta evaluasi final model.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan hasil pelatihan, evaluasi, dan analisis model IndoT5 pada tiga skema, yaitu *Full Fine-Tuning* (FFT), *Low-Rank Adaptation* (LoRA), dan *Knowledge Distillation*–LoRA (KD-LoRA), yang mencakup hasil pelatihan serta evaluasi dengan ROUGE dan BERT*Score* untuk pemilihan model terbaik, dilanjutkan dengan evaluasi final pada data uji dari sisi performa, efisiensi, dan karakteristik ringkasan.

BAB V PENUTUP

Bab ini menjelaskan kesimpulan penelitian untuk menjawab rumusan masalah serta saran untuk pengembangan penelitian selanjutnya.