diperlukan strategi yang tepat untuk menjaga stabilitas dan kinerja model (Xia et al., 2025). Di sisi lain, penggunaan LLM berbasis API menimbulkan biaya operasional berulang yang perlu diperhitungkan secara cermat oleh institusi pendidikan, terutama dalam skala penerapan yang besar (Pilicita & Barra, 2025).

Meskipun kedua pendekatan tersebut (*semantic embeddings* berbasis SBERT dan *generative LLMs* berbasis *prompting*) telah diteliti secara terpisah, terdapat sejumlah celah penelitian yang signifikan dalam literatur yang ada. Pertama, belum terdapat kajian komparatif yang secara langsung mempertemukan model SBERT yang telah di-*fine-tune* pada domain spesifik dengan strategi *prompting* GPT untuk tugas ASAG dalam kondisi eksperimen yang terkontrol. Kedua, sebagian besar penelitian sebelumnya tidak mempertimbangkan perbedaan mendasar antara dua tipe pertanyaan yang umum ditemui dalam praktik pendidikan: soal *open-ended* yang menuntut respons eksplanatori dengan variasi parafrasi tinggi, dan soal *close-ended* yang memerlukan penulisan entitas faktual spesifik dan terbatas. Ketiga, pertimbangan *trade-off* antara performa dan biaya operasional jarang dikuantifikasi secara eksplisit dalam penelitian ASAG, padahal faktor ini sangat menentukan kelayakan penerapan di lingkungan institusi nyata.

Penelitian ini menggunakan *dataset* Mohler ASAG (Mohler et al., 2011), sebuah kumpulan data pertanyaan ilmu komputer berbahasa Inggris yang dilengkapi dengan jawaban peserta didik beserta skor dari penilai manusia. *Dataset* ini kemudian disegmentasi menjadi dua *subset* berdasarkan tipe pertanyaan: *subset Open-Ended* (OE) sebanyak 2.273 sampel yang dicirikan oleh variasi parafrase tinggi, dan *subset Close-Ended* (CE) sebanyak 169 sampel yang mensyaratkan *recall* entitas faktual yang presisi. Segmentasi ini dirancang untuk memungkinkan analisis komparatif yang lebih spesifik, sehingga melampaui evaluasi agregat yang umumnya dilaporkan dalam studi-studi sebelumnya.

Penelitian ini mengintegrasikan dua pendekatan teknis utama: pendekatan *semantic embeddings* menggunakan model Sentence-BERT `all-distilroberta-v1` dalam kondisi *zero-shot* dan setelah *fine-tuning*, serta pendekatan *generative LLMs* melalui strategi *zero-shot* dan *few-shot prompting* pada model GPT-5.4-nano dan GPT-5.4, Dengan desain tersebut, penelitian ini dirancang untuk menghasilkan panduan penerapan yang berbasis bukti empiris. Pertanyaan sentral yang dijawab adalah: apakah *fine-tuning* domain-spesifik

pada model yang lebih kecil dan tanpa biaya operasional berulang dapat menandingi atau bahkan melampaui performa LLM generalis yang jauh lebih besar?

Jawaban atas pertanyaan ini memiliki implikasi praktis yang langsung bagi institusi pendidikan dalam memilih solusi ASAG yang optimal, disesuaikan dengan tipe soal, kapasitas infrastruktur, serta kendala anggaran yang dimiliki.

1.2 Rumusan Masalah

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

1. Bagaimana kinerja model Sentence-BERT setelah *fine-tuning* dibandingkan dengan SBERT *zero-shot* dalam penilaian jawaban uraian singkat secara otomatis?
2. Bagaimana pengaruh strategi *few-shot prompting* dibandingkan *zero-shot prompting* terhadap performa model GPT dalam penilaian jawaban uraian singkat, dan apakah pengaruh tersebut bergantung pada kapasitas model?
3. Apakah terdapat perbedaan performa yang signifikan antara pendekatan *fine-tuned* SBERT dan strategi *prompting* GPT pada soal *open-ended* dibandingkan dengan soal *close-ended*?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut.

1. Menghasilkan pengukuran kuantitatif peningkatan kemampuan model Sentence-BERT setelah *fine-tuning* domain-spesifik dalam tugas ASAG dibandingkan dengan kondisi *zero-shot*.
2. Menghasilkan temuan mengenai pengaruh strategi *few-shot prompting* terhadap performa model GPT dalam tugas ASAG, termasuk keterkaitannya dengan kapasitas model yang digunakan.
3. Menghasilkan panduan pendekatan ASAG yang optimal untuk tipe soal *open-ended* maupun *close-ended*, dengan mempertimbangkan *trade-off* antara performa penilaian dan biaya operasional.

1.4 Manfaat Penelitian

Manfaat dilaksanakannya penelitian ini adalah sebagai berikut.

1. Memberikan perbandingan empiris langsung antara *fine-tuned* SBERT dan strategi *prompting* GPT dalam tugas ASAG, sehingga memperjelas kondisi penggunaan optimal masing-masing pendekatan berdasarkan tipe soal dan kapasitas model.
2. Memberikan panduan berbasis data bagi institusi pendidikan dalam memilih solusi ASAG yang sesuai dengan kebutuhan, kapasitas infrastruktur, dan kendala anggaran operasional.
3. Menyediakan dokumentasi desain *prompt engineering* yang tervalidasi dan estimasi biaya operasional yang terkuantifikasi, sehingga dapat langsung direplikasi atau diadaptasi oleh pengembang sistem *e-learning*.
4. Menjadi referensi bagi penelitian lanjutan di bidang ASAG, termasuk pengembangan pada dataset multibahasa maupun pendekatan hibrida.

1.5 Ruang Lingkup

Ruang lingkup penelitian ini ditetapkan untuk menjaga fokus dan kesesuaian dengan tujuan yang telah dirumuskan. Batasan penelitian ini adalah sebagai berikut.

1. *Dataset* yang digunakan adalah Mohler ASAG dalam bahasa Inggris dengan domain soal ilmu komputer, yang disegmentasi menjadi *subset Open-Ended* (OE) sebanyak 2.273 sampel (pembagian 80/20 untuk *train*/*test*) dan *subset Close-Ended* (CE) sebanyak 169 sampel (pembagian 70/30 untuk *train*/*test*). Penggunaan *dataset* berbahasa selain bahasa Inggris atau dari domain yang berbeda berpotensi memengaruhi akurasi penilaian, sehingga generalisasi hasil penelitian ini perlu mempertimbangkan faktor tersebut.
2. Tugas yang diselesaikan adalah klasifikasi biner, dengan label positif (1) diberikan apabila skor rata-rata jawaban $\geq 3{,}5$ dan label negatif (0) untuk skor $< 3{,}5$.
3. Augmentasi data dilakukan secara eksklusif pada *training set* menggunakan model GPT-5.4-nano dengan strategi berbeda berdasarkan nilai `score_avg`: kalimat berlawanan untuk skor 0,0, parafrase `instructor_answer` untuk skor 5,0, dan parafrase `student_answer` untuk skor lainnya. Kualitas hasil augmentasi divalidasi menggunakan METEOR *Score* dan *cosine similarity*.

4. Metode pendekatan *Semantic Embeddings* meliputi dua varian model Sentence-BERT berbasis *checkpoint* `all-distilroberta-v1`: (a) SBERT *zero-shot* dengan *threshold* yang dikalibrasi, dan (b) SBERT yang di-*fine-tune* menggunakan `CosineSimilarityLoss` pada data pelatihan masing-masing *subset*.
5. Metode pendekatan *Generative LLMs* meliputi empat kombinasi model dan strategi *prompting*: GPT-5.4-nano dan GPT-5.4, masing-masing dengan strategi *zero-shot* dan *few-shot prompting*, menggunakan desain *system prompt* dan *user prompt* yang didokumentasikan secara penuh.
6. Matriks evaluasi yang digunakan adalah *F1-Score* sebagai matriks utama, didukung oleh *Accuracy*, *Precision*, dan *Recall*. Matriks regresi seperti *Pearson Correlation*, *Quadratic Weighted Kappa* (QWK), atau RMSE tidak digunakan karena tugas ini merupakan klasifikasi biner.
7. Analisis operasional mencakup perbandingan latensi pengujian dan estimasi biaya API per sejumlah submisi tertentu. Perlu dicatat bahwa estimasi biaya yang dihitung hanya meliputi biaya token untuk penggunaan API GPT, dan tidak memperhitungkan biaya operasional lainnya seperti biaya komputasi, infrastruktur server, maupun biaya pengembangan sistem.
8. Penelitian ini tidak mencakup: (a) *dataset* berbahasa Indonesia atau bahasa lain selain bahasa Inggris; (b) *fine-tuning* model GPT; serta (c) evaluasi pada tipe soal selain *open-ended* dan *close-ended* sebagaimana didefinisikan dalam segmentasi *dataset*.

1.6 Sistematika Penulisan

Dokumen skripsi ini disusun dengan sistematika penulisan sebagai berikut.

BAB 1 PENDAHULUAN

Bab ini membahas latar belakang permasalahan yang mendorong dilaksanakannya penelitian, rumusan masalah, tujuan penelitian, manfaat penelitian, ruang lingkup yang menjadi batasan penelitian, serta sistematika penulisan laporan skripsi ini.

BAB 2 TINJAUAN PUSTAKA

Bab ini memaparkan teori dan konsep yang menjadi landasan ilmiah penelitian, mencakup *state of the art* penelitian ASAG, konsep *Automated Short Answer*

Grading (ASAG), pra-pemrosesan teks (pembersihan teks, pelabelan biner, pembagian dataset, dan augmentasi data), arsitektur Transformer dan BERT, Sentence-BERT beserta komponen-komponennya, *cosine similarity*, *transfer learning*, *fine-tuning*, *grid search*, *Large Language Model* (LLM), *prompt engineering*, serta evaluasi model klasifikasi biner.

BAB 3 METODOLOGI PENELITIAN

Bab ini menguraikan rancangan alur penelitian secara menyeluruh. Pembahasan diorganisasikan ke dalam gambaran umum penelitian yang mencakup pengumpulan data, pra-pemrosesan data, pembagian data dan pelabelan biner, augmentasi dan penyeimbangan *training set*, desain pendekatan *Semantic Embeddings* (SBERT *zero-shot* dan *fine-tuned*), desain pendekatan *Generative LLMs* (strategi *prompting* GPT), serta prosedur evaluasi. Bab ini juga memuat pengaturan lingkungan eksperimen yang digunakan.

BAB 4 HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil implementasi dan eksperimen dari keenam metode yang dievaluasi pada *subset Open-Ended* dan *Close-Ended*, disertai analisis mendalam terhadap empat skenario perbandingan utama: pengaruh *fine-tuning* pada SBERT, pengaruh strategi *prompting* pada GPT, perbandingan performa lintas tipe soal, serta identifikasi model terbaik berdasarkan pertimbangan performa dan efisiensi biaya operasional.

BAB 5 PENUTUP

Bab ini merumuskan kesimpulan yang menjawab langsung setiap rumusan masalah yang telah ditetapkan, serta menyampaikan saran untuk penelitian lanjutan di bidang ASAG yang dapat dikembangkan dari hasil penelitian ini.