BAB I
PENDAHULUAN

Bab ini menjelaskan kerangka dasar dari penelitian yang terdiri dari latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan penelitian, dan sistematika penulisan. Seluruh bagian tersebut disusun untuk memberikan gambaran umum mengenai arah, fokus, serta ruang lingkup penelitian secara menyeluruh.

1.1. Latar Belakang

Evaluasi merupakan bagian krusial dalam proses pembelajaran karena memungkinkan dosen dan mahasiswa menilai sejauh mana capaian pembelajaran telah tercapai. Salah satu bentuk evaluasi yang paling bernilai adalah pemberian *feedback*, yang tidak hanya menunjukkan kekurangan dalam jawaban mahasiswa, tetapi juga memberikan arahan untuk perbaikan dan penguatan pemahaman. *Feedback* yang efektif telah terbukti dapat meningkatkan kualitas belajar mandiri, memperkuat refleksi akademik, dan meningkatkan hasil belajar secara keseluruhan (Williams, 2024).

Namun, dalam praktiknya, pemberian *feedback* secara manual oleh dosen seringkali menemui berbagai kendala. Di lingkungan pendidikan tinggi, terutama pada kelas dengan jumlah mahasiswa besar, dosen menghadapi keterbatasan waktu dan beban kerja tinggi yang dapat berdampak pada keterlambatan, ketidakkonsistenan, atau bahkan penghilangan proses pemberian *feedback* secara menyeluruh (M. Zhang dkk., 2025) Hal ini menimbulkan kebutuhan akan solusi yang dapat membantu proses evaluasi tanpa mengorbankan kualitas instruksionalnya.

Untuk menjawab tantangan tersebut, berkembanglah pendekatan *automatic feedback generation*, yakni pemanfaatan teknologi untuk menghasilkan umpan balik secara otomatis terhadap jawaban mahasiswa. Salah satu teknologi terdepan dalam bidang ini adalah penggunaan *Large Language Models* (*LLM*), yaitu model bahasa besar yang dilatih untuk memahami dan menghasilkan teks secara kontekstual. Model seperti GPT dari OpenAI dan LLaMA dari Meta AI telah menunjukkan kemampuan

dalam berbagai tugas pemrosesan bahasa alami, termasuk *feedback generation* berbasis *prompt* (Tanwar dkk., 2024; M. Zhang dkk., 2025).

Meski menjanjikan efisiensi, masih terdapat celah antara *automatic feedback* dan human *feedback* dari sisi relevansi, kejelasan instruksi, dan konteks akademik. Penelitian menunjukkan bahwa kualitas *output* LLM sangat bergantung pada desain *prompt* yang diberikan. Strategi *prompting* seperti *Zero-Shot*, *One-Shot*, *Few-Shot*, dan *Chain-of-Thought* dapat menghasilkan gaya dan kedalaman *feedback* yang berbeda-beda (Wei et al., 2022). Selain itu, penggunaan persona *prompting*, seperti menyisipkan identitas model sebagai "Dosen" atau "Ahli", turut memengaruhi gaya bahasa dan sudut pandang *feedback* yang dihasilkan (Stahl dkk., 2024)

Beberapa pendekatan telah dikembangkan untuk memanfaatkan LLM dalam *feedback generation*, di antaranya adalah *fine-tuning* model, *retrieval-augmented generation*, dan *prompt engineering*. Fine-tuning memungkinkan adaptasi model terhadap domain tertentu, namun membutuhkan data dalam jumlah besar dan sumber daya komputasi tinggi. *Retrieval-based* cenderung mengandalkan basis data eksternal dan sistem pencocokan, tetapi kurang fleksibel dalam memproduksi respons kontekstual. Sementara itu, *prompt engineering* menawarkan pendekatan yang ringan dan fleksibel, cukup dengan mengubah instruksi masukan tanpa perlu melatih ulang model (Mazzullo & Bulut, 2024)

Oleh karena itu, dalam penelitian ini dipilih pendekatan *prompt engineering*, karena pendekatan ini memungkinkan eksplorasi berbagai strategi instruksi (*prompting*) dan persona tanpa perlu *fine-tuning* model atau integrasi sistem tambahan. Hal ini sangat sesuai dengan keterbatasan data dan infrastruktur yang tersedia, serta memungkinkan evaluasi langsung terhadap pengaruh desain *prompt* terhadap kualitas feedback yang dihasilkan oleh LLM.

Penelitian ini dirancang untuk mengevaluasi kualitas *feedback* yang dihasilkan oleh LLM terhadap jawaban mahasiswa dari dua mata kuliah berbeda, yaitu Sistem Cerdas dan Metodologi Penelitian dan Penulisan Ilmiah (MPI). Kedua mata kuliah tersebut dipilih karena mewakili dua karakteristik tugas akademik yang berbeda: Sistem Cerdas lebih banyak berisi jawaban teknis dan konseptual yang bersifat spesifik

dalam bidang ilmu komputer, sedangkan MPI menuntut penulisan jawaban naratif dan argumentatif yang lebih panjang serta terstruktur. Perbedaan karakteristik ini penting untuk menguji sejauh mana LLM mampu menghasilkan *feedback* yang fleksibel dan relevan dalam konteks disiplin ilmu yang berbeda, baik yang bersifat teknis maupun yang bersifat penulisan ilmiah.

Penelitian dilakukan dengan menerapkan empat strategi *prompting*, tiga konfigurasi persona, dan dua model LLM (GPT-4o mini dan LLaMA 3.1). Evaluasi kualitas dilakukan menggunakan empat metrik otomatis: BLEU, ROUGE-L, METEOR, dan BERTScore, yang mencakup pendekatan leksikal dan semantik. Penelitian ini diharapkan dapat memberikan kontribusi baik secara praktis dalam membantu dosen menghasilkan *feedback* berkualitas, maupun secara teoretis dalam pengembangan riset NLP di bidang pendidikan tinggi.

1.2. Rumusan Masalah

Penelitian ini dilatarbelakangi oleh kebutuhan akan mekanisme pemberian umpan balik yang efisien, relevan, dan berkualitas dalam konteks pendidikan tinggi, khususnya terhadap jawaban mahasiswa yang bersifat terbuka atau naratif. Dengan berkembangnya teknologi *Large Language Model* (LLM), proses *feedback generation* dapat diotomatisasi melalui pendekatan *prompt engineering*. Namun, kualitas *feedback* yang dihasilkan model sangat dipengaruhi oleh strategi *prompting* yang digunakan serta peran atau persona yang dimasukkan dalam *prompt*.

Berdasarkan latar belakang tersebut, rumusan masalah dalam penelitian ini adalah bagaimana kualitas *feedback* yang dihasilkan oleh LLM terhadap jawaban mahasiswa, jika dianalisis berdasarkan strategi *prompting*, jenis persona, model LLM, dan kombinasi strategi *prompting* yang digunakan, serta dievaluasi menggunakan metrik otomatis BLEU, ROUGE-L, METEOR, dan BERTScore dalam konteks pendidikan tinggi.

1.3. Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengevaluasi kualitas *feedback* yang dihasilkan oleh *Large Language Model* (LLM) terhadap jawaban mahasiswa dengan

menganalisis pengaruh strategi *prompting*, jenis persona, dan model yang digunakan. Penelitian ini secara khusus bertujuan untuk menilai efektivitas berbagai strategi *prompting*, yaitu *Zero-Shot, One-Shot, Few-Shot,* dan *Chain-of-Thought*, serta menganalisis bagaimana penggunaan persona seperti *Base*, Dosen, dan Ahli dapat memengaruhi kualitas respons model. Selain itu, penelitian ini juga membandingkan performa dua model LLM, yaitu GPT-4o mini dan LLaMA 3.1, dalam menghasilkan *feedback* akademik yang relevan. Seluruh hasil evaluasi dianalisis menggunakan metrik otomatis BLEU, ROUGE-L, METEOR, dan BERTScore, untuk mengukur sejauh mana *feedback* yang dihasilkan mendekati *feedback* referensi yang telah disusun sebelumnya, baik dalam konteks mata kuliah Sistem Cerdas maupun Metodologi Penelitian dan Penulisan Ilmiah.

1.4. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara teoritis maupun praktis. Secara teoritis, penelitian ini memperkaya kajian di bidang *Natural Language Processing* (NLP) dan penerapan *Large Language Models* (LLM) dalam konteks pendidikan tinggi, khususnya dalam tugas evaluasi otomatis berbasis teks. Dengan mengeksplorasi pengaruh strategi *prompting*, persona, dan jenis model terhadap kualitas *feedback* yang dihasilkan, penelitian ini dapat memberikan kontribusi terhadap pengembangan metode evaluasi berbasis *prompt engineering* dan pemodelan konteks pada sistem kecerdasan buatan.

Secara praktis, hasil dari penelitian ini dapat dimanfaatkan oleh pendidik, institusi pendidikan, maupun pengembang teknologi edukasi dalam mengimplementasikan LLM sebagai alat bantu pemberian umpan balik terhadap tugas mahasiswa. Temuan mengenai kombinasi strategi dan persona yang paling efektif dapat menjadi acuan dalam mendesain sistem *feedback* generator yang lebih adaptif, efisien, dan relevan dengan konteks pembelajaran di kelas. Selain itu, penelitian ini juga dapat membantu institusi dalam mempertimbangkan integrasi LLM sebagai bagian dari solusi pembelajaran berbasis AI, terutama pada mata kuliah yang menuntut analisis naratif atau konseptual.

1.5. Batasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan untuk memperjelas ruang lingkup dan interpretasi hasil, antara lain:

1. Terbatas pada dua mata kuliah

   Penelitian hanya dilakukan pada dua mata kuliah, yaitu Sistem Cerdas dan Metodologi Penelitian dan Penulisan Ilmiah (MPI). Oleh karena itu, temuan yang diperoleh belum tentu dapat digeneralisasikan ke bidang lain yang memiliki jenis tugas atau bentuk jawaban yang berbeda.

2. Evaluasi sepenuhnya otomatis

   Penilaian kualitas *feedback* dilakukan menggunakan empat metrik berbasis teks, yakni BLEU, ROUGE-L, METEOR, dan BERTScore, tanpa disertai evaluasi manual dari manusia (dosen atau mahasiswa). Akibatnya, interpretasi hasil hanya merefleksikan kesamaan leksikal dan semantik dengan *feedback* referensi, bukan nilai pedagogis atau dampak.

3. Terbatas pada dua model LLM dan konfigurasi *default*

   Model yang digunakan terbatas pada GPT-4o mini (API) dan LLaMA 3.1 (lokal via Ollama), sehingga belum mencakup variasi arsitektur model lain atau versi dengan *tuning* berbeda. Selain itu, penelitian hanya menguji empat strategi *prompting* (*Zero-Shot, One-Shot, Few-Shot, Chain-of-Thought*) dan tiga jenis persona (Base, Dosen, Ahli), tanpa mengeksplorasi elemen seperti panjang *prompt*, gaya bahasa, atau pengaturan temperatur model secara sistematis.

4. Tidak membahas perbandingan antar tipe *One-Shot*

   Penelitian ini menggunakan *One-Shot prompting* dengan tipe yang berdasarkan variasi nilai jawaban mahasiswa (rendah, sedang, tinggi) tetapi, belum melakukan analisis terpisah terhadap jenis *One-Shot prompting* berdasarkan variasi nilai jawaban mahasiswa, sehingga belum dapat menyimpulkan apakah efektivitas *One-Shot* dipengaruhi oleh kualitas jawaban yang dijadikan contoh dalam *prompt*.

5. Tidak mengevaluasi dampak nyata terhadap pembelajaran

   Penelitian ini tidak mengukur secara langsung bagaimana *feedback* yang dihasilkan memengaruhi proses belajar atau kualitas revisi mahasiswa. Fokus utama hanya pada kesesuaian *output* dengan *feedback* referensi, bukan pada validitas fungsional atau penerimaan pengguna akhir.

1.6. Sistematika Penulisan

Sistematika penulisan skripsi ini disusun untuk memberikan gambaran yang terstruktur mengenai isi penelitian. Skripsi terdiri dari lima bab utama, yang masing-masing membahas aspek berbeda dari penelitian ini.

BAB I PENDAHULUAN

Bab ini berisi pengantar terhadap topik penelitian, yang meliputi latar belakang permasalahan, rumusan masalah, tujuan dan manfaat penelitian, batasan ruang lingkup, serta sistematika penulisan. Bab ini bertujuan untuk memberikan gambaran awal mengenai konteks dan arah penelitian.

BAB II TINJAUAN PUSTAKA

Bab ini menguraikan teori-teori, konsep, dan hasil penelitian terdahulu yang relevan dengan topik penelitian. Di dalamnya dibahas mengenai *Large Language* Model (LLM), konsep *feedback* dalam pendidikan, strategi *prompting* (*Zero-Shot, One-Shot, Few-Shot,* dan *Chain-of-Thought*), persona dalam *prompting*, serta metrik evaluasi otomatis seperti BLEU, ROUGE-L, METEOR, dan BERTScore.

BAB III METODE PENELITIAN

Bab ini menjelaskan secara rinci rancangan penelitian yang digunakan, termasuk pendekatan penelitian, sumber dan jumlah data, prosedur penyusunan *prompt*, proses generasi *feedback* menggunakan dua model LLM (GPT-4o dan LLaMA), serta teknik evaluasi menggunakan empat metrik otomatis. Selain itu, dijelaskan pula metode analisis data serta perangkat lunak yang digunakan dalam eksperimen.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil eksperimen berupa analisis skor evaluasi dari strategi *prompting*, persona dalam *prompt,* model LLM, dan setiap kombinasi model terhadap jawaban mahasiswa pada hasil keluaran model berdasarkan mata kuliah sistem cerdas dan metodologi penelitian dan penulisan ilmiah. Analisis dilakukan terhadap performa masing-masing strategi, persona, model, dan kombinasi baik dari segi rata-rata maupun sebaran skor.

BAB V PENUTUP

Bab terakhir ini memuat kesimpulan dari temuan utama yang diperoleh dari penelitian, serta memberikan saran untuk penelitian lanjutan atau pengembangan sistem *feedback* otomatis berbasis LLM di masa depan. Peneliti juga merefleksikan keterbatasan penelitian sebagai dasar pertimbangan pengembangan lebih lanjut.