1. Pendahuluan

1.1 Latar Belakang

Teknologi kecerdasan buatan, khususnya *Large Language Models* (LLMs), telah membawa perubahan signifikan di berbagai bidang, termasuk pendidikan. Salah satu penerapan yang paling menonjol adalah kemampuan LLMs untuk mengevaluasi esai, yang memiliki potensi besar dalam mendukung pembelajaran bahasa Inggris. Berbagai penelitian telah mengeksplorasi potensi ini, menunjukkan bahwa LLMs dapat memberikan umpan balik yang efektif bagi pembelajar *English as a Foreign Language* (EFL) [21]. Meskipun demikian, validitas dan reliabilitasnya masih menjadi perdebatan utama. Sebuah studi, misalnya, menemukan adanya konsistensi yang buruk antara penilaian esai oleh ChatGPT dan penilai manusia, yang menggarisbawahi perlunya kehati-hatian dalam penggunaannya [1]. Tantangan serupa juga ditekankan dalam penelitian lain yang menyoroti pentingnya menguji validitas dan reliabilitas LLMs secara mendalam, terutama saat penilaian didasarkan pada rubrik yang kompleks [9].

Tantangan ini diperkuat oleh temuan bahwa LLMs seringkali kesulitan dalam memahami elemen linguistik yang halus dan konteks budaya spesifik dalam sebuah tulisan [2], [20]. Studi lain secara eksplisit menemukan bahwa model seperti ChatGPT dan LLaMA cenderung menghasilkan skor yang tidak dapat diandalkan dan tidak berkorelasi baik dengan penilaian manusia [15]. Di sisi lain, model sumber terbuka seperti LLaMA memang menawarkan kapabilitas untuk evaluasi esai, namun kinerjanya terbukti sangat bergantung pada arsitektur model dan data pelatihan yang digunakan [3], [6]. Untuk mengatasi berbagai keterbatasan tersebut, salah satu solusi yang diusulkan adalah kerangka kerja kolaborasi manusia-AI. Pendekatan ini terbukti tidak hanya dapat mengotomatisasi proses penilaian, tetapi juga meningkatkan performa dan efisiensi penilai manusia, terutama untuk esai yang sulit dinilai oleh model *Artificial Intelligence* (AI) [14].

Kebutuhan pendidikan modern menuntut alat penilaian yang efektif untuk membantu para pembelajar meningkatkan keterampilan menulis mereka. LLMs menawarkan potensi sebagai asisten yang bermanfaat, di mana beberapa model terkemuka seperti GPT-4 telah terbukti memiliki validitas dan reliabilitas yang baik dalam menilai tulisan pembelajar bahasa Inggris [5]. Meskipun demikian, evaluasi yang lebih mendalam dan kritis tetap diperlukan untuk memastikan keadilan dan keselarasan dengan penilaian manusia, terutama dalam asesmen skala besar [20]. Perbedaan fundamental dalam arsitektur dan data pelatihan antar LLMs adalah penyebab utama variasi performa ini [11]. Variasi tersebut berdampak langsung pada hasil penilaian esai, terutama dalam tiga aspek krusial: konsistensi penilaian, kualitas umpan balik yang diberikan, dan ketepatan deteksi kesalahan tata bahasa.

Dasar dari penelitian ini dibangun di atas studi ekstensif yang telah ada. ChatGPT, sebagai salah satu model yang paling banyak diteliti, telah menunjukkan hasil yang beragam; beberapa studi menunjukkan kemampuannya menyamai skor manusia dalam kondisi tertentu [18], namun studi lain justru menyoroti konsistensi yang buruk saat dibandingkan langsung dengan penilai manusia [1] dan menekankan perlunya validasi lebih lanjut saat menggunakan rubrik penilaian [9]. Di sisi model sumber terbuka, LLaMA juga telah menjadi subjek penelitian yang signifikan. Beberapa hasil menunjukkan potensinya untuk tugas penilaian dan revisi esai [3], namun perbandingan langsung menunjukkan performanya masih di bawah model closed-source seperti GPT dalam hal konsistensi [13], bahkan ada studi yang menyimpulkan bahwa LLaMA cenderung tidak dapat diandalkan [15]. Dari sini terlihat bahwa meskipun banyak penelitian telah dilakukan, masih ada ketidakpastian performa dan celah signifikan dalam literatur, terutama studi komparatif yang secara spesifik mengevaluasi Gemini.

Oleh karena itu, perbandingan langsung antara model-model terdepan seperti ChatGPT 4o, Gemini 2.0, dan LLaMA 4 menjadi krusial untuk memahami kekuatan dan kelemahan unik masing-masing. Studi ini akan menganalisis dan membandingkan performa ketiganya dalam mengevaluasi esai bahasa Inggris berdasarkan metrik konsistensi, kualitas umpan balik, dan deteksi kesalahan. Dengan metode komparatif, penelitian ini bertujuan untuk mengidentifikasi karakteristik khas dari setiap model dan memberikan rekomendasi praktis bagi para pendidik. Hasilnya, studi ini diharapkan dapat berkontribusi pada pengembangan sistem penilaian otomatis yang lebih andal dan efektif, yang mampu memenuhi kebutuhan kompleks dalam pendidikan bahasa Inggris di era digital.

1.2 Topik dan Batasannya

Penelitian ini menganalisis kinerja tiga LLMs, yaitu ChatGPT 4o, Gemini 2.0, dan LLaMA 4, dalam mengevaluasi esai bahasa Inggris berdasarkan konsistensi penilaian, kualitas umpan balik, dan kemampuan mendeteksi kesalahan tata bahasa. Dengan pendekatan komparatif, penelitian ini bertujuan untuk mengidentifikasi karakteristik unik dari masing-masing model, memberikan wawasan tentang kekuatan dan kelemahan mereka, serta menyusun rekomendasi praktis bagi pendidik dan pengembang teknologi pendidikan. Penelitian ini diharapkan berkontribusi pada pengembangan sistem penilaian esai otomatis yang lebih efektif, yang mendukung kebutuhan pembelajar bahasa Inggris, khususnya mereka yang mempelajari EFL di era digital, dengan memanfaatkan *near-native machine capabilities* dari LLMs untuk simulasi penilaian yang

efisien.

Namun, untuk menjaga fokus dan keterjangkauan, penelitian ini memiliki beberapa batasan. Pertama, penelitian hanya mengevaluasi 71 esai bahasa Inggris dari festival bahasa yang diselenggarakan oleh sebuah universitas di Jawa Barat dengan tema "*Unlocking International Opportunities Through Languages*". Oleh karena itu, hasil penelitian ini tidak mencakup esai dari konteks atau institusi lain di Indonesia, sehingga generalisasi temuan mungkin terbatas pada konteks serupa. Kedua, penelitian ini hanya membandingkan tiga model, yaitu ChatGPT 4o, Gemini 2.0, dan LLaMA 4, tanpa melibatkan model lain, sehingga analisis terfokus pada kinerja ketiga model tersebut. Ketiga, penilaian esai dilakukan menggunakan prompt yang dirancang berdasarkan prinsip "KORAN LAPANGAN", di mana hasil penelitian bergantung pada struktur prompt tersebut, dan penggunaan prompt yang berbeda dapat menghasilkan variasi dalam kinerja model.

Selain itu, penelitian ini tidak melibatkan *human expert* atau perbandingan dengan pendapat *human expert* dalam evaluasi esai, yang menjadi batasan signifikan karena tidak ada validasi eksternal dari pakar bahasa Inggris. Alasan utama tidak menggunakan *human expert* adalah keterbatasan sumber daya dan keterjangkauan, terutama karena kebutuhan akan *native speaker* atau pakar *near-native* yang mahal dan sulit diakses untuk skala penelitian ini. Hal ini mencerminkan *limitation to user*, di mana penelitian lebih menekankan pada perspektif general user yang berinteraksi dengan LLMs untuk mengevaluasi esai, bukan dari sudut pandang pakar. Peran peneliti dalam hal ini bukan sebagai *human expert*, melainkan mewakili general user dengan tambahan latar belakang keilmuan informatika, yang memungkinkan simulasi dan modelisasi proses evaluasi yang berjalan baik secara teknis. Namun, hasil riset ini tetap merupakan model dan simulasi yang efektif dalam konteks terbatas, dengan tetap menjelaskan batasan kemampuan model seperti ketergantungan pada prompt dan potensi bias AI yang tidak sepenuhnya mencerminkan penilaian manusiawi. Dengan batasan-batasan ini, penelitian tetap terarah pada evaluasi kinerja model dalam konteks spesifik, memberikan landasan yang jelas untuk interpretasi hasil dan rekomendasi yang dihasilkan.

1.3 Tujuan

Penelitian ini bertujuan untuk mencapai tujuan utama sebagai berikut:

1. Membandingkan kinerja Large Language Model (LLMs) seperti ChatGPT 4o, Gemini 2.0, dan LLaMA 4 dalam evaluasi esai bahasa Inggris, dengan fokus pada:
    a. Konsistensi
    Menilai stabilitas model dalam memberikan evaluasi yang sesuai dengan format prompt, termasuk skor per komponen, nilai total dan rata-rata, serta deteksi kesalahan bahasa.
2. Menganalisis perbedaan umpan balik yang diberikan oleh ChatGPT 4o, Gemini 2.0, dan LLaMA 4, dengan fokus pada:
    a. Kualitas Umpan Balik
    Membandingkan kelengkapan, kejelasan, dan relevansi umpan balik yang diberikan oleh masing-masing model, berdasarkan rubrik penilaian.
    b. Kegunaan Umpan Balik
    Mengevaluasi seberapa bermanfaat umpan balik dari masing-masing model untuk membantu penulis memperbaiki esai mereka.
3. Menilai kemampuan masing-masing model dalam mendeteksi kesalahan gramatikal pada esai, dengan fokus pada:
    a. Deteksi Kesalahan Gramatikal
    Mengukur akurasi model dalam mengidentifikasi kesalahan tata bahasa.
    b. Saran Perbaikan
    Mengevaluasi relevansi dan spesifisitas saran perbaikan yang diberikan oleh model untuk mendukung perbaikan esai.