1. Pendahuluan

Latar Belakang

Teknologi kecerdasan buatan (AI), terutama *Large Language Models* (LLM) seperti ChatGPT (OpenAI), Gemini (Google DeepMind), dan LLaMA (Meta AI), telah mengubah semua sektor secara signifikan, termasuk pendidikan. LLM telah menunjukkan kemampuannya sebagai alat belajar, terutama dalam mendukung pemahaman matematika dan pemecahan masalah [1]. Kalkulus, khususnya pada topik limit, turunan, dan integral, seringkali menjadi tantangan bagi pelajar. Oleh karena itu, menganalisis potensi LLM untuk memfasilitasi pembelajaran kalkulus sangat relevan untuk meningkatkan efektivitas pendidikan.

Studi sebelumnya menunjukkan bahwa LLM dapat menghasilkan solusi matematika yang terstruktur, tetapi kinerjanya bervariasi secara signifikan tergantung pada kompleksitas masalah [2], [3]. Dalam ChatGPT, meskipun fleksibel, ia menunjukkan ketidakkonsistenan dengan masalah kalkulus yang kompleks, sementara Gemini unggul dalam langkah-langkah sistematis [4], [5], [6]. Kemudian, LLaMA, meskipun belum diuji secara khusus dalam konteks kalkulus, telah menunjukkan potensi dalam menghasilkan respons terstruktur [7]. Variasi ini memerlukan analisis komparatif komprehensif untuk memahami kekuatan dan keterbatasan masing-masing model dalam pembelajaran kalkulus.

Studi ini mengusulkan pendekatan baru untuk mengatasi hal ini dengan membandingkan kinerja ChatGPT 4o, Gemini 2.0, dan LLaMA 4 dalam menyelesaikan soal kalkulus. Ketiga LLM ini dipilih secara spesifik karena mewakili LLM terkemuka dari ekosistem teknologi AI yang berbeda, yaitu OpenAI, Google DeepMind, dan Meta AI. Perbandingan akan berdasarkan tiga metrik: *correctness*, *clarity*, dan *representation*. Data evaluasi akan diproses menggunakan *Min-Max scaling* untuk normalisasi, *Manual Weighting* untuk integrasi skor, dan *K-Means clustering* untuk pengelompokan kinerja [8], [9], [10].

Data akan dikumpulkan sebanyak 90 soal kalkulus (30 soal masing-masing untuk limit, turunan, dan integral) dari buku teks yang umum digunakan, dengan distribusi yang merata antara soal mudah, sedang, dan sulit. Setiap soal akan diformat secara seragam untuk masukan LLM yang konsisten. Ahli kalkulus akan mengevaluasi respons LLM berdasarkan *correctness*, *clarity*, dan *representation*, dengan data dicatat dalam Excel. Data ini akan menjalani *Min-Max scaling* dengan rentang 0.00 – 1.00 dan *Manual Weighting* dengan bobot yang digunakan dalam penelitian ini yaitu *correctness* (0.5), *clarity* (0.3), dan *representation* (0.2). Akhirnya, *K-Means clustering* akan diterapkan pada data yang sudah dibobot untuk mengelompokkan kinerja, dengan jumlah *cluster* optimal ditentukan oleh metode *Elbow*.

Penelitian ini bertujuan untuk memberikan wawasan mendalam tentang kemampuan teknis masing-masing LLM dalam menyelesaikan masalah kalkulus dan implikasi pendidikannya untuk meningkatkan pengajaran tingkat perguruan tinggi [11]. Dengan menggunakan pendekatan sistematis dan berbasis data, penelitian ini diharapkan dapat menghasilkan rekomendasi yang objektif dan dapat diterapkan bagi akademisi. Penelitian ini juga akan mengidentifikasi potensi penuh dan batasan LLM dalam pendidikan matematika, membimbing pendidik untuk secara efektif mengintegrasikan AI guna meningkatkan kualitas dan efisiensi pembelajaran.

Topik dan Batasannya

Penelitian ini bertujuan untuk melakukan analisis komparatif performa *Large Language Models* (LLM) seperti ChatGPT 4o, Gemini 2.0, dan LLaMA 4 dalam menyelesaikan soal kalkulus. Permasalahan utama yang akan dijelaskan adalah sejauh mana model-model LLM ini dapat diandalkan untuk menyelesaikan soal kalkulus secara akurat dan edukatif, mengingat tantangan pemahaman konsep limit, turunan, dan integral yang sering dihadapi mahasiswa. Selain itu, penelitian ini juga menganalisis bagaimana data evaluasi performa ketiga LLM tersebut dapat diproses menggunakan teknik *Min-Max scaling* dan *Manual Weighting*, kemudian dikelompokkan dengan *K-Means clustering* untuk mengidentifikasi pola kinerja spesifik pada setiap model. Dalam konteks ini, input penelitian ini adalah 90 soal kalkulus yang mencakup tiga topik (limit, turunan, integral) dengan tiga tingkat kesulitan (mudah, sedang, dan sulit). Output yang diharapkan adalah evaluasi performa masing-masing LLM berdasarkan tiga metrik utama, yaitu *correctness*, *clarity*, dan *representation*, serta pengelompokan performa model menggunakan algoritma *K-Means clustering*. Sebagai contoh singkat, penelitian ini akan menguji bagaimana setiap LLM merespons soal turunan kompleks dan menganalisis kualitas penjelasannya serta keakuratan jawabannya dibandingkan dengan dua LLM lainnya.

Batasan dalam penelitian ini diterapkan untuk menyederhanakan ruang lingkup agar sesuai dengan karakteristik Tugas Akhir. Pertama, analisis perbandingan dibatasi pada tiga model LLM spesifik: ChatGPT 4o, Gemini 2.0, dan LLaMA 4, karena popularitas dan mewakili LLM terkemuka dari ekosistem teknologi AI yang berbeda. Kedua, fokus penelitian adalah kemampuan LLM dalam memahami dan memecahkan soal kalkulus standar (limit, turunan, integral) dengan variasi tingkat kesulitan (mudah, sedang, sulit), sehingga tidak mencakup topik matematika lainnya. Ketiga, jumlah data soal dibatasi pada 90 soal, yaitu 30 soal per topik (limit, turunan, integral) dengan distribusi 10 soal per tingkat kesulitan. Keempat, penelitian ini menggunakan *prompt* dasar dan seragam untuk menguji setiap model, tanpa menggunakan teknik *prompt engineering* yang kompleks untuk memastikan konsistensi input antar model. Terakhir, evaluasi hasil respons terbatas pada penilaian oleh tiga ahli kalkulus, yang merupakan dosen kalkulus, di mana setiap ahli hanya

bertanggung jawab mengevaluasi 90 respons (30 soal dari 3 LLM) untuk satu topik kalkulus spesifik (limit, turunan, atau integral). Pembatasan jumlah ahli dan cakupan evaluasi ini disesuaikan dengan ketersediaan ahli dan efisiensi proses penilaian.

Tujuan

Penelitian ini memiliki beberapa tujuan utama yang akan dicapai melalui serangkaian eksperimen. Pertama, akan dilakukan evaluasi dan membandingkan performa ChatGPT 4o, Gemini 2.0, dan LLaMA 4 dalam menyelesaikan 90 soal kalkulus yang mencakup topik limit, turunan, dan integral, dengan variasi tingkat kesulitan. Performa ini akan diukur secara terukur menggunakan metrik *correctness*, *clarity*, dan *representation*. Selanjutnya, penelitian ini juga bertujuan untuk menganalisis data evaluasi dengan menerapkan *Min-Max scaling* untuk metrik *clarity* dan *representation*, serta *Manual Weighting* untuk menghasilkan skor performa terintegrasi yang lebih komprehensif. Setelah itu, akan dilakukan pengelompokan performa ketiga model LLM menggunakan *K-Means clustering* berdasarkan data yang telah dibobot, dengan menentukan jumlah *cluster* optimal melalui metode *elbow*. Tujuan akhir dari penelitian ini adalah memberikan rekomendasi berbasis data mengenai model LLM yang paling efektif untuk mendukung pembelajaran kalkulus di perguruan tinggi, berdasarkan hasil analisis klasterisasi dan visualisasi data yang telah dilakukan.

Organisasi Tulisan

Penelitian ini terbagi menjadi lima bagian utama yang saling berkaitan. Pendahuluan menguraikan latar belakang, perumusan masalah, batasan, tujuan, dan sistematika penulisan. Studi Terkait membahas teori pendukung dan penelitian terdahulu yang relevan, termasuk model, metode, pendekatan, serta metrik evaluasi. Bagian Sistem yang Dibangun menjelaskan deskripsi *dataset*, proses *clustering*, *prompting*, pengujian model, evaluasi performa, *Min-Max scaling*, *Manual Weighting*, dan *K-Means clustering*. Selanjutnya, Evaluasi memuat hasil pengujian performa model, hasil *preprocessing* (*Min-Max scaling* dan *Manual Weighting*), serta hasil *K-Means clustering*. Terakhir, Kesimpulan merangkum temuan utama penelitian dan memberikan rekomendasi LLM sebagai alat pembelajaran kalkulus yang efektif.