

Abstrak

Perkembangan *Large Language Models* (LLM) seperti ChatGPT, Gemini, dan LLaMA memiliki potensi besar dalam mendukung pembelajaran kalkulus. Namun, performa mereka sangat bervariasi tergantung kompleksitas soal. Penelitian ini membandingkan kinerja ChatGPT 4o, Gemini 2.0, dan LLaMA 4 pada 90 soal kalkulus (limit, turunan, integral) dengan tingkat kesulitan beragam. Respons dievaluasi oleh ahli berdasarkan metrik *correctness*, *clarity*, dan *representation*. Skor dinormalisasi dengan *Min-Max scaling*, dikombinasikan melalui *Manual Weighting* (0.5, 0.3, 0.2), dan dikelompokkan menggunakan *K-Means clustering*. Hasil menunjukkan bahwa Gemini 2.0 dan ChatGPT 4o mendominasi Cluster 1 (Performa Optimal), sedangkan LLaMA 4 sering berada di Cluster 0 (*Correctness* Tinggi, *Representation* dan *Clarity* Rendah). Penelitian ini merekomendasikan Gemini 2.0 dan ChatGPT 4o sebagai alat bantu pembelajaran kalkulus yang efektif, dengan catatan pada keterbatasan notasi dan konsistensi jawaban.

Kata kunci: Large Language Models, ChatGPT, Gemini, LLaMA, Kalkulus, K-Means clustering.