Abstract

The development of Large Language Models (LLMs) such as ChatGPT, Gemini, and LLaMA holds significant potential in supporting calculus learning. However, their performance varies greatly depending on the complexity of the problem. This study compares the performance of ChatGPT 4o, Gemini 2.0, and LLaMA 4 on 90 calculus problems (limits, derivatives, integrals) of varying difficulty levels. Responses were evaluated by experts based on metrics of correctness, clarity, and representation. Scores were normalized using Min-Max scaling, combined through Manual Weighting (0.5, 0.3, 0.2), and grouped using K-Means clustering. The results show that Gemini 2.0 and ChatGPT 4o dominate Cluster 1 (Optimal Performance), while LLaMA 4 often falls into Cluster 0 (High Correctness, Low Representation and Clarity). This study recommends Gemini 2.0 and ChatGPT 4o as effective calculus learning tools, with caveats regarding notation limitations and answer consistency.

Keywords: Large Language Models, ChatGPT, Gemini, LLaMA, Calculus, K-Means clustering