ABSTRAK

Perkembangan *Large Language Model* (LLM) membuka peluang baru dalam pengembangan sistem penilaian esai otomatis, khususnya untuk dokumen ilmiah yang kompleks dan multimodal. Namun, penerapan penilaian otomatis pada esai panjang multimodal berbahasa Indonesia, seperti *paper Systematic Literature Review* (SLR), masih menghadapi tantangan terkait kompleksitas struktur dokumen, keterbatasan pemrosesan konteks, serta minimnya riset pada dokumen yang memadukan teks dan elemen visual. Kondisi ini menuntut pendekatan yang tidak hanya mengandalkan kemampuan generatif LLM, tetapi juga mengintegrasikan mekanisme penalaran yang selaras dengan rubrik penilaian akademik. Penelitian ini mengkaji kinerja multimodal LLM dalam penilaian paper SLR berbahasa Indonesia melalui eksperimen komparatif berbagai teknik *prompting*, yaitu *zero-shot, chain-of-thought, multi-trait, dan multi-chain prompting*. Teknik terbaik diterapkan pada evaluasi 104 paper mahasiswa menggunakan model GPT-4o dan GPT-5.x. Penilaian dilakukan berdasarkan rubrik akademik yang mencakup aspek konseptual, metodologis, struktural, dan visual. Evaluasi performa dilakukan pada level komponen rubrik dan skor agregat menggunakan metrik *Quadratic Weighted Kappa* (QWK), akurasi, *Pearson Correlation Coefficient* (PCC), *Mean Absolute Error* (MAE), dan *Root Mean Squared Error* (RMSE). Hasil penelitian menunjukkan bahwa *multi-trait prompting* menghasilkan keselarasan tertinggi dengan penilaian manusia, sementara model GPT-5.x secara konsisten menunjukkan performa yang lebih baik dibandingkan GPT-4o, dengan GPT-5.2 sebagai model dengan performa paling optimal. Performa model meningkat secara signifikan pada skenario evaluasi yang berfokus pada elemen visual, yang mengindikasikan bahwa panjang konteks dan kompleksitas dokumen berpengaruh terhadap kualitas penilaian otomatis. Penelitian ini menunjukkan bahwa kinerja multimodal LLM dalam penilaian esai panjang bersifat selektif dan sangat dipengaruhi oleh desain *prompt*, karakteristik rubrik, serta struktur dokumen. Temuan ini memberikan kontribusi empiris dan konseptual terhadap pengembangan sistem penilaian esai multimodal otomatis berbasis LLM pada konteks Bahasa Indonesia, sekaligus menegaskan perlunya strategi *prompt* dan pengelolaan konteks yang terstruktur untuk mendekati pola penilaian manusia.

Kata kunci—penilaian esai otomatis, multimodal LLM, *prompt engineering*, GPT, *systematic literature review* (SLR)

ABSTRACT

The rapid advancement of Large Language Models (LLMs) has opened new opportunities for developing automated essay scoring systems, particularly for complex and multimodal academic documents. However, the application of automated assessment to long multimodal essays in Indonesian, such as Systematic Literature Review (SLR) papers, still faces significant challenges related to document structural complexity, limitations in context processing, and the scarcity of research on documents that integrate textual and visual elements. These conditions require approaches that not only rely on the generative capabilities of LLMs but also integrate reasoning mechanisms aligned with academic assessment rubrics. This study investigates the performance of multimodal LLMs in assessing Indonesian SLR papers through a comparative experimental evaluation of several prompting techniques, including zero-shot, chain-of-thought, multi-trait, and multi-chain prompting. The best-performing technique was subsequently applied to the evaluation of 104 student papers using GPT-4o and GPT-5.x models. The assessment was conducted based on an academic rubric encompassing conceptual, methodological, structural, and visual aspects. Model performance was evaluated at both the rubric component level and the aggregate score level using Quadratic Weighted Kappa (QWK), accuracy, Pearson Correlation Coefficient (PCC), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The results indicate that multi-trait prompting achieves the highest alignment with human assessment, while GPT-5.x models consistently outperform GPT-4o, with GPT-5.2 demonstrating the most optimal performance. Model performance improves significantly in evaluation scenarios focused on visual elements, suggesting that context length and document complexity substantially influence the quality of automated assessment. Overall, the findings reveal that the performance of multimodal LLMs in long-essay assessment is selective and highly dependent on prompt design, rubric characteristics, and document structure. This study contributes empirical and conceptual insights into the development of multimodal LLM-based automated essay scoring systems in the Indonesian context and underscores the importance of structured prompting strategies and context management to approximate human evaluation patterns.

Keywords—automated essay scoring, multimodal LLM, prompt engineering, GPT, systematic literature review (SLR)