Abstract

The English essay assessment in Indonesia is often slow and inconsistent, especially for English as a Foreign Language (EFL) learners, as teachers struggle to provide detailed feedback on a large number of essays, such as those submitted to university language festivals. Artificial intelligence offers a solution, but the performance of language models in assessing essays has not been comprehensively compared. This study uses 71 essays on the theme of international opportunities through language as input, generating scores, feedback, and lists of language errors. However, feedback from current models is often too general, failing to meet the needs of effective English as a foreign language learning. This study aims to compare three artificial intelligence models in essay evaluation, focusing on consistency, feedback quality, and grammar error detection. This topic is important for improving evaluation efficiency, such as in language competitions, but current models are often inconsistent or unclear, creating a gap with educational needs. A total of 71 essays from a language festival at a university in West Java were evaluated by three artificial intelligence models based on criteria such as grammar, word choice, argument logic, writing style, and content appropriateness. A specific prompt guided the evaluation, and the output was analyzed using a rubric for consistency, feedback quality, and error detection. This research produced guidelines for selecting the best model and improving automated evaluation. Results and contributions Gemini excelled in quality feedback (70.42% Very Helpful) and error detection (78.87% Very Accurate), followed by ChatGPT (49.30% Very Helpful, 54.93% Very Accurate), while LLaMA 4 is consistent (85.92% Consistent with Notes) but less specific (43.66% Not Possible). This research supports the use of artificial intelligence as an essay assessment assistant for EFL in Indonesia.

Keywords: ChatGPT, Gemini, LLaMA, essay grading, artificial intelligence, feedback, grammatical errors, consistency