ABSTRAK

Penilaian jawaban singkat secara otomatis Automated Short Answer Grading (ASAG) merupakan tantangan dalam dunia pendidikan karena membutuhkan pemahaman semantik dan konsistensi penilaian yang tinggi. Penelitian ini mengusulkan Metode baru bernama Automated Rubric Generation and Example Selection (ARGES) yang bertujuan meningkatkan akurasi dan keandalan penilaian menggunakan model Large Language Models (LLM). ARGES menggabungkan pembuatan rubrik otomatis dan pemilihan contoh jawaban serupa berbasis kesamaan semantik untuk membentuk prompt evaluasi yang lebih terarah. Eksperimen dilakukan menggunakan dataset Mohler dan dua model LLM (Deepseek dan LLaMA3 70B), dengan membandingkan Metode ARGES terhadap metode prompting few-shot. Evaluasi performa dilakukan menggunakan metrik QWK, MAE, RMSE, dan Pearson Correlation, serta ditambah analisis inkonsistensi dan uji Wilcoxon untuk mengukur signifikansi statistik. Hasil menunjukkan bahwa ARGES secara konsisten mengungguli few-shot, dengan peningkatan QWK sebesar +0,22 pada Deepseek dan +0,11 pada LLaMA3 70B. Selain itu, Metode ini menghasilkan kesesuaian yang tinggi terhadap penilaian manusia (QWK 0,7754, MAE 0,3857, RMSE 0,7313, Pearson 0,8342) dan tingkat inkonsistensi yang sangat rendah (0,37%). Temuan ini menunjukkan bahwa ARGES merupakan Metode yang efektif, stabil, dan layak digunakan untuk mendukung sistem penilaian otomatis berbasis LLM dalam konteks pendidikan.

Kata kunci— ASAG, Large Language Model, Prompt Engineering, ARGES, Penilaian Otomatis

ABSTRACT

Automated Short Answer Grading (ASAG) presents a challenge in educational settings due to the need for semantic understanding and consistent evaluation. This study proposes a novel approach called Automated Rubric Generation and Example Selection (ARGES), aimed at improving the accuracy and reliability of scoring using Large Language Models (LLMs). ARGES combines automatic rubric generation with semantically similar example selection to construct more structured evaluation prompts. Experiments were conducted using the Mohler dataset and two LLMs (Deepseek and LLaMA3 70B), comparing ARGES against the conventional few-shot prompting method. Model performance was evaluated using QWK, MAE, RMSE, and Pearson Correlation metrics, supplemented by inconsistency analysis and the Wilcoxon signed-rank test for statistical significance. Results show that ARGES consistently outperformed the few-shot method, with QWK improvements of +0.22 for Deepseek and +0.11 for LLaMA3 70B. Furthermore, the model achieved high alignment with human scoring (QWK 0.7754, MAE 0.3857, RMSE 0.7313, Pearson 0.8342) and a very low inconsistency rate (0.37%). These findings indicate that ARGES is an effective, consistent, and promising approach for supporting LLM-based automated grading systems in educational environments.

Keywords— ASAG, Large Language Model, Prompt Engineering, ARGES, Automatic Scoring.