ABSTRAK

Pendidikan merupakan aspek penting dalam pembangunan suatu negara, dengan salah satu tantangannya adalah menciptakan sistem penilaian jawaban esai yang efisien. Penilaian manual sering kali tidak efektif karena memerlukan waktu, tenaga, dan biaya yang besar serta rentan terhadap bias. Penilaian otomatis mulai berkembang dengan pembelajaran mesin, *deep learning*, hingga penggunaan *language model*. Penggunaan *large language model* (LLM) menunjukkan performa yang baik dalam evaluasi teks, tetapi membutuhkan sumber daya komputasi dan biaya yang tinggi. Sebaliknya, *small language model* lebih efisien dalam hal komputasi, tetapi masih memiliki tingkat kesepakatan yang lebih rendah dengan penilai. Penelitian ini mengajukan pendekatan *Atomic Evaluation*, yang memecah konteks panjang menjadi unit-unit kecil sebelum dilakukan penilaian untuk meningkatkan kesepakatan antara *small language model* dan penilai dalam penilaian jawaban esai otomatis. Eksperimen dilakukan dengan membandingkan *Atomic Evaluation* dengan metode *Zero-Shot Chain-of-Thought* (CoT) guna mengevaluasi efektivitas pendekatan yang diajukan. Pengujian dilakukan menggunakan enam model berbeda, yaitu Gemma 2 9B, Llama 3.1 8B, Llama 3 8B, Qwen 2.5 7B, Mistral 7B, dan WizardLM 2 7B. Penelitian ini juga menggunakan dua dataset berbeda, yaitu eLOK dan ASAP-SAS. Pengujian diukur dengan metrik *Quadratic Weighted Kappa* (QWK) yang merepresentasikan tingkat kesepakatan. Hasil eksperimen menunjukkan bahwa *Atomic Evaluation* meningkatkan QWK *Small Language Model* pada dataset eLOK dengan rata-rata peningkatan 0.14772 dibandingkan CoT. Namun, pada dataset ASAP-SAS, peningkatan tidak terjadi pada semua model. Tiga model, yaitu Gemma 2 9B, Llama 3.1 8B, dan Llama 3 8B, mengalami peningkatan, sementara tiga model lainnya, yaitu Qwen 2.5 7B, Mistral 7B, dan WizardLM 2 7B, mengalami penurunan akibat ketidaksesuaian pola distribusi nilai. Berdasarkan rata-rata QWK tertinggi dan konsistensi hasil pada kedua dataset, model Gemma 2 9B dipilih sebagai model terbaik untuk diuji dalam evaluasi keluaran model oleh penilai. Hasil evaluasi menunjukkan peningkatan rata-rata QWK sebesar 0.12026, yang menunjukkan bahwa keluaran model dapat diterima dengan baik oleh penilai. Selain itu, evaluasi penerimaan menggunakan *Technology Acceptance Model* (TAM) yang menunjukkan bahwa keluaran model memiliki nilai *Perceived Usefulness* (PU) sebesar 3.66 dan *Perceived Ease of Use* (PEOU) sebesar 4.06, yang mengindikasikan bahwa keluaran model cukup dapat diterima oleh penilai.

Kata kunci—Penilaian Jawaban Ujian Esai Otomatis, *Atomic Evaluation*, *Small Language Model*, *Quadratic Weighted Kappa*

ABSTRACT

Education is an important aspect of a country's development, with one of its challenges being the creation of an efficient system for evaluating essay answers. Manual scoring is often ineffective because it requires considerable time, effort, and cost, and is prone to bias. Automatic scoring has been advancing through machine learning, deep learning, and the use of language models. The use of large language models (LLMs) has demonstrated strong performance in text evaluation but requires high computational resources and incurs significant costs. In contrast, small language models are more computationally efficient but still exhibit a lower level of agreement with human raters. This study proposes an approach called Atomic Evaluation, which breaks down long contexts into smaller units before evaluation to increase agreement between small language models and human raters in automated essay scoring. Experiments were conducted by comparing Atomic Evaluation with the Zero-Shot Chain-of-Thought (CoT) method to evaluate the effectiveness of the proposed approach. Testing was carried out using six different models: Gemma 2 9B, Llama 3.1 8B, Llama 3 8B, Qwen 2.5 7B, Mistral 7B, and WizardLM 2 7B. The study also utilized two different datasets: eLOK and ASAP-SAS. The evaluation was measured using the Quadratic Weighted Kappa (QWK) metric, which represents the level of agreement. The experimental results indicate that Atomic Evaluation increased QWK of the Small Language Model on the eLOK dataset, with an average improvement of 0.14772 compared to CoT. However, on the ASAP-SAS dataset, improvements did not occur across all models. Three models—Gemma 2 9B, Llama 3.1 8B, and Llama 3 8B—showed improvement, whereas the other three models—Qwen 2.5 7B, Mistral 7B, and WizardLM 2 7B—experienced a decline due to inconsistencies in value distribution patterns. Based on the highest average QWK and consistent results across both datasets, the Gemma 2 9B model was selected as the best-performing model for further evaluation by human raters. The evaluation showed an average QWK increase of 0.12026, indicating that the model output was well received by the raters. Additionally, an acceptance evaluation using the Technology Acceptance Model (TAM) showed that the model output achieved a Perceived Usefulness (PU) score of 3.66 and a Perceived Ease of Use (PEOU) score of 4.06, suggesting that the model output is relatively well accepted by the raters.

Keywords—Automated Essay Exam Scoring, Atomic Evaluation, Small Language Model, Quadratic Weighted Kappa.