ABSTRAK

PENERAPAN *SUPERVISED FINE-TUNING* PADA *OPEN-SOURCE LARGE LANGUAGE MODEL* UNTUK PENILAIAN OTOMATIS ESAI BERBAHASA INDONESIA

Oleh

AHLAN SAYYID ALGHIFFARI

Penilaian esai otomatis sangat penting untuk mengurangi subjektivitas serta beban kerja penilai manusia dalam dunia pendidikan. Penelitian ini bertujuan untuk mengimplementasikan *Supervised Fine-Tuning* (SFT) pada model *open-source Large Language Model* (LLM), yaitu Meta Llama 3.1 8B *Instruct*, untuk penilaian otomatis esai berbahasa Indonesia. Penelitian ini menggunakan desain eksperimen dengan memanfaatkan dataset UKARA 1.0 *Challenge* yang terdiri dari *Problem* A dan *Problem* B. Metodologi yang digunakan adalah *Parameter-Efficient Fine-Tuning* (PEFT) dengan pendekatan *4-bit Quantized Low-Rank Adaptation* (QLoRA) melalui *framework* Unsloth untuk mengatasi keterbatasan komputasi dan pemakaian memori. Empat skema data dievaluasi, yaitu data *Raw*, *Misspelling Correction* & *Normalization*, augmentasi *back-translation*, dan data kombinasi dari ketiga skema tersebut. Ketidakseimbangan kelas ditangani menggunakan pendekatan *weighted loss*, selain itu berbagai strategi *prompting* juga dilakukan pengujian. Hasil penelitian menunjukkan bahwa SFT secara signifikan meningkatkan performa dibandingkan pendekatan tanpa *fine-tuning*. Performa SFT terbaik diperoleh menggunakan data mentah dengan strategi *prompt template* menggunakan Llama *special tokens*, akurasi mencapai 89,93% pada *Problem* A dan 72,13% pada *Problem* B, dengan F1-*score* terbaik pada kelas positif masing-masing sebesar 93,16% dan 75,68%.

Kata kunci: Penilaian Otomatis Esai, SFT, LLM, QLoRA, Llama 3.1.

ABSTRACT

APPLICATION OF SUPERVISED FINE-TUNING ON AN OPEN-SOURCE LARGE LANGUAGE MODEL FOR AUTOMATED INDONESIAN ESSAY SCORING

By

AHLAN SAYYID ALGHIFFARI

Automated essay scoring plays a crucial role in reducing subjectivity and the workload of human raters in educational settings. This study aims to implement Supervised Fine-Tuning (SFT) on an open-source Large Language Model (LLM), namely Meta Llama 3.1 8B Instruct, for automated scoring of Indonesian-language essays. The research adopts an experimental design utilizing the UKARA 1.0 Challenge dataset, which consists of Problem A and Problem B. The methodology employs Parameter-Efficient Fine-Tuning (PEFT) using the 4-bit Quantized Low-Rank Adaptation (QLoRA) approach through the Unsloth framework to address computational and memory constraints. Four data schemes are evaluated: raw data, misspelling correction and normalization, back-translation augmentation, and a combination of these three schemes. Class imbalance is handled using a weighted loss approach, and various prompting strategies are also evaluated. The results indicate that SFT significantly improves performance compared to approaches without fine-tuning. The best SFT performance is achieved using raw data with a prompt template strategy utilizing Llama special tokens, achieving an accuracy of 89.93% on Problem A and 72.13% on Problem B, with the best F1-scores on the positive class reaching 93.16% and 75.68%, respectively.

***Keywords**: Automated Essay Scoring, SFT, LLM, QLoRA, Llama 3.1.*