ANALISIS PERBANDINGAN *ZERO-SHOT* DAN *FINE-TUNING* PADA LLM UNTUK *AUTOMATED ESSAY SCORING*

Alexandria Felicia Seanne[1], Husnul Hadah[2], Yustika Heti Handal[3], Britney Levina Sukma[4], Budi Tjahyono[5]
Program Studi Magister Ilmu Komputer, Fakultas Ilmu Komputer
Universitas Esa Unggul[1,2,3,4,5]
Jalan Raya Arjuna UTARA No. 9, Tol Tomang Kebun Jeruk, Jakarta Barat 11510
fgelicia@student.esaunggul.ac.id[1], husnul.hadah@student.esaunggul.ac.id[2], handalyustikah@student.esaunggul.ac.id[3], britneylevina2002@student.esaunggul.ac.id[4], budi.tjahyono@esaunggul.ac.id[5]

Abstrak

Automated Essay Scoring (AES) menggunakan *Large Language Model* (LLM) berkembang pesat namun masih terdapat keterbatasan dalam studi yang membandingkan pendekatan *Zero-shot prompting* dan *Fine-tuning* pada model *Open-weight* LLM. Penelitian ini menganalisis perbandingan kedua pendekatan tersebut pada model Qwen 2.5-1.5B-Instruct menggunakan dataset ASAP-AES 2.0. Pendekatan *Zero-shot* menerapkan teknik *Role Prompting* dan *Chain-of-Thought* (CoT), sedangkan *Fine-tuning* menggunakan *Parameter-Efficient Fine-tuning* (PEFT/LoRA). Evaluasi dilakukan menggunakan metrik *Quadratic Weighted Kappa* (QWK) dan *Root Mean Square Error* (RMSE). Hasil penelitian menunjukkan pendekatan *Zero-shot* menghasilkan QWK = 0.0069 dengan *Central tendency bias* ekstrem, sedangkan *Fine-tuning* menghasilkan QWK = 0.3247 (*fair agreement*). Temuan ini mengonfirmasi bahwa adaptasi pendekatan *Fine-tuning* merupakan salah satu cara untuk menghasilkan sistem penilaian esai otomatis yang cukup pada model *open-weight* berukuran kecil.

Kata Kunci: *Zero-shot*, *Fine-tuning*, *Large Language Model* (LLM), Automated Essay Scoring (AES), *Open-weight* LLM

Abstract

Automated Essay Scoring (AES) using Large Language Models (LLMs) is rapidly evolving, yet there remains a lack of studies comparing Zero-shot prompting and Fine-tuning approaches on open-weight LLM models. This study analyzes the comparison of these two approaches on the Qwen 2.5-1.5B-Instruct model using the ASAP-AES 2.0 dataset. The Zero-shot approach employs Role Prompting and Chain-of-Thought (CoT) techniques, while Fine-tuning uses Parameter-Efficient Fine-tuning (PEFT/LoRA). Evaluation was conducted using the Quadratic Weighted Kappa (QWK) and Root Mean Square Error (RMSE) metrics. The results show that the Zero-shot approach yields a QWK of 0.0069 with an extreme central tendency bias, while Fine-tuning yields a QWK of 0.3247 (fair agreement). These findings confirm that adapting a Fine-tuning approach is one way to produce a sufficiently accurate automatic essay grading system on small-scale open-weight models

***Keywords**: Zero-shot, Fine-tuning, Large Language Model (LLM), Automated Essay Scoring (AES), Open-weight LLM*

PENDAHULUAN

Penulisan esai adalah salah satu cara untuk mengukur kemampuan berpikir kritis, argumentasi, dan pemahaman siswa (Liu et al., 2026). Penilaian esai tidak seperti penilaian objektif yang hanya memeriksa aspek mengingat dan memahami saja (Adha et al., 2026; Mahmud et al., 2025). Selain itu, penilaian esai juga penting untuk mendorong perkembangan kemampuan berpikir kritis jika ada umpan balik yang diberikan guru (Martin et al., 2025). Umpan balik ini memungkinkan siswa mengidentifikasi kesalahan dan melakukan perbaikan secara berkelanjutan. Hal tesebut tidak hanya meningkatkan kualitas tulisan, tetapi juga mendorong perkembangan kemampuan berpikir kritis (Bouaziz et al., 2025; Diningrat et al., 2025; Martin et al., 2025). Umpan balik ini juga dapat mendorong proses refleksi diri sehingga dapat meningkatkan kemampuan belajar mandiri siswa (Rizqi Amelia & Tommy S Suyasa, 2025).

Namun proses penilaian secara manual oleh guru seringkali dihadapi tantangan terkait efisiensi dan konsistensi (Pulung Hendro Prastyo et al., 2025). Proses ini membutuhkan waktu yang lama, terutama pada kelas dengan jumlah siswa yang besar. Kondisi tersebut berpotensi menghambat pemberian umpan balik secara tepat waktu (Alawadh et al., 2024). Di samping itu, penilaian manual juga rentan terhadap berbagai bentuk bias seperti bias subjektif, efek halo dan ketidakkonsistenan akibat kelelahan (Faza et al., 2025). Kondisi ini dapat mempengaruhi efisiensi dan berpotensi menurunkan tingkat keadilan dan objektivitas penilaian yang diterima oleh siswa (García-Varela et al., 2025).

Sebagai solusi dari kendala tersebut, dikembangkanlah teknologi Automated Essay Scoring (AES) untuk memberikan penilaian secara otomatis menggunakan kecerdasan buatan (Choi et al., 2026; Pulung Hendro Prastyo et al., 2025). AES mampu memberikan penilaian yang lebih cepat dan konsisten sehingga dapat mengurangi beban kerja administratif guru secara signifikan (Xiao et al., 2025). Pembuatan AES dengan pemanfaatan *Large Language Model* (LLM) dilakukan karena LLM memiliki kemampuan pemahaman Bahasa alami yang sangat maju (AlGhamdi et al., 2026; Liu et al., 2026). Dalam penerapannya, LLM memiliki dua pendekatan utama yang banyak diteliti yakni pendekatan *Zero-shot* dan *Fine-tuning* (Choi et al., 2026; Huang et al., 2026). Dimana pendekatan *Zero-shot* menggunakan instruksi berbasis *prompt* untuk menghasilkan penilaian tanpa pelatihan ulang pada dataset spesifik (AlGhamdi et al., 2026; Liu et al., 2026). Sebaliknya, pendekatan *Fine-tuning* memerlukan penyesuaian parameter model sebelum dapat digunakan untuk menghasilkan penilaian (Faza et al., 2025; Huang et al., 2026).

Hingga saat ini, penelitian yang membandingkan efektivitas *Zero-shot* dan *Fine-tuning* masih terbatas pada LLM *open-weight*. Sehingga penelitian ini bertujuan untuk melakukan analisis perbandingan pendekatan *Zero-shot* dan *Fine-tuning* pada model Qwen 2.5 pada penerapan AES. Proses evaluasi dilakukan dengan *Quadratic Weighted Kappa* (QWK) dan *Root Mean Square Error* (RMSE).

PENELITIAN RELEVAN

Penelitian pada Automated Essay Scoring (AES) telah mengalami perubahan dari penggunaan metode ekstraksi fitur tradisional menuju pemanfaatan *Large Language Model* (LLM) (Pulung Hendro Prastyo et al., 2025). Tabel 1 merangkum penelitian terdahulu yang sudah pernah dilakukan menggunakan pendekatan LLM.

Tabel 1. Rangkuman Penelitian Terdahulu

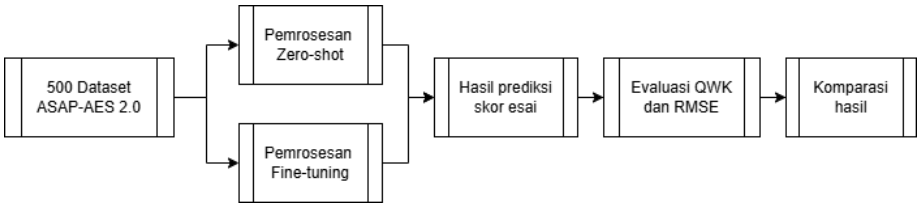
Peneliti (Tahun)	Fokus	Metode	Dataset	Hasil Utama	Gap Penelitian
(Pulung Hendro Prastyo et al., 2025)	Komparasi 10 pretrained transformer untuk AES Bahasa Indonesia	Text embedding + cosine similarity (BERT, ALBERT, SBERT) - tanpa *Fine-tuning*	300 esai kuis Metodologi Penelitian, PNUP (Dataset Indonesia)	IndoBERT-base-p2 terbaik: RMSE 5.664, QWK 0.675	Tidak membandingkan metode *Zero-shot* vs *Fine-tuning*; hanya menggunakan model transformer lama (IndoBERT) tanpa *prompting*
(Xiao et al., 2025)	Framework Fast & Slow Thinking LLM untuk AES + kolaborasi human-AI	*Zero-shot*, *few-shot* GPT-4; *Fine-tuning* GPT-3.5 & LLaMA3-8B (LoRA); dual-process framework	ASAP (12.978 esai, 8 *prompt*) + CSEE (13.372 esai siswa China) (Dataset Inggris)	Korelasi Pearson 0.72; *Zero-shot* kompetitif tanpa data latih	Tidak ada komparasi sistematis *Zero-shot* vs *Fine-tuning* pada arsitektur model yang sama dalam satu dataset; fokus pada model GPT-4 dan LLaMA
(Huang et al., 2026)	Akurasi & fairness GPT-4o untuk AES: *prompting* vs *Fine-tuning* vs feature-based	*Zero-shot* CoT, few-shot, *Fine-tuning* GPT-4o; dibandingkan feature-based AES & human rater	1.768 esai kelas 3–4 SD (Dataset Inggris)	Fine-tuned GPT-4o terbaik (QWK 0.84); *few-shot* kompetitif (0.76); *Zero-shot* lemah (0.47–0.60)	Fokus pada model closed-source (GPT-4o); belum menguji stabilitas penilaian model *open-weight* secara eksplisit serta tidak menggunakan metrik RMSE

Peneliti (Tahun)	Fokus	Metode	Dataset	Hasil Utama	Gap Penelitian
(Mughal et al., 2026)	Eksplorasi LLM (GPT & Gemini) untuk AES dengan analisis bias penilai manusia	*Zero-shot* GPT-3.5-turbo & Gemini-Pro	ASAP-AES (10.685 esai, 6 set), LA-AES (5.143 esai), O-Level Pakistan (21 esai) (Dataset Inggris)	GPT: rata-rata QWK 0.45 (ASAP-AES) dan 0.43 (LA-AES). Objektif: LLM > penilai manusia	Hanya menguji pendekatan *Zero-shot* tanpa menyertakan *Fine-tuning* sebagai pembanding untuk melihat batas optimal performa model
(Choi et al., 2026)	Eksplorasi pendekatan *zero-shot* AES dari *feature-based* hingga LLM-*based*	*Zero-shot* CAnUSe (LLM *pairwise* + *uncertainty self-training*) & EASY (3 fitur: panjang esai, *vocabulary diversity*, *grammar quality*)	ASAP++ (12.978 esai), TOEFL11 (12.100 esai), Ellipse (6.482 esai) (Dataset Inggris)	CAnUSe mendekati performa model supervised (≈10K label); QWK ASAP++ 57.9, TOEFL11 68.1. Ellipse 57.9. EASY melampaui metode *feature-based supervised*	Menggunakan pendekatan *uncertainty self-training* yang kompleks; belum membandingkan efisiensi *prompt engineering* sederhana vs *fine-tuning* pada model Qwen

Berdasarkan Tabel 1, terdapat celah penelitian berupa belum adanya perbandingan pendekatan *Fine-tuning* dan *Zero-shot* pada model *Open-weight* seperti Qwen 2.5 pada dataset ASAP-AES 2.0. Selain itu, evaluasi terhadap stabilitas penilaian dan presisi kesalahan masih belum banyak diteliti.

METODE PENELITIAN

Poin utama pada penelitian ini adalah persiapan dan pengumpulan data, eksperimen perbandingan dua pendekatan adaptasi LLM dan evaluasi dan analisis hasil. Adapun Arsitektur sistem penelitian ini dalam melakukan perbandingan pendekatan *Zero-shot prompting* dan *Fine-tuning* dijalankan pada model Qwen2.5-1.5B-Instruct. Model ini menggunakan model Qwen versi 2.5 dengan ukuran 1.5 miliar parameter yang telah dilatih menggunakan pendekatan Instruction tuning agar mampu mengikuti instruksi manusia.



Gambar 1. Alur Penelitian

Gambar 1 menunjukkan rancangan alur bagaimana keseluruhan sistem penelitian ini dibuat. Dimulai dengan pengumpulan 500 dataset ASAP-AES 2.0 lalu data diproses menggunakan pendekatan *Zero-shot* dan *Fine-tuning* yang menghasilkan prediksi skor esai. Hasil prediksi skor esai dievaluasi menggunakan QWK dan RMSE agar menemukan tingkat keselarasan dan tingkat presisi dari kedua pendekatan tersebut. Sistem ini diakhiri dengan melakukan komparasi hasil dari kedua pendekatan tersebut. Penelitian ini melakukan eksperimen dengan menerapkan parameter teknis yang dapat dilihat pada Tabel 2.

Tabel 2. Parameter Teknis Eksperimen

Parameter	*Zero-shot*	*Fine-tuning* (LoRA)
Model LLM	Qwen2.5-1.5B-Instruct	Qwen2.5-1.5B-Instruct

Parameter	*Zero-shot*	*Fine-tuning* (LoRA)
Jumlah Percobaan per Esai	3x (stabilitas)	3x (epoch training)
Strategi *Prompt*	*Role prompting* + CoT	Parameter-Efficient *Fine-tuning* (PEFT/LoRA)
Metrik Evaluasi Utama	QWK & RMSE	QWK & RMSE
Dataset	ASAP-AES 2.0	ASAP-AES 2.0

Pada pendekatan *Zero-shot*, teks esai diintegrasikan ke dalam *prompt* terstruktur yang memuat identitas peran model (*role prompting* sebagai penilai akademik), langkah-langkah untuk melakukan penilaian dan instruksi eksplisit untuk menghasilkan skor numerik. Model LLM menerima *prompt* ini dan menghasilkan output secara langsung tanpa pembaruan parameter apapun. Agar menghindari sifat stokastik LLM, setiap esai diproses sebanyak 3 kali untuk menganalisis stabilitas skor. Adapun desain *prompt* yang digunakan pada penelitian ini dibagi menjadi dua yakni *system message* yang ada pada Gambar 2 dan *user message* pada Gambar 3.

```
system_msg = (
    'You are an expert essay grader with 15 years of experience evaluating '
    'student academic writing. You score essays fairly, consistently, and '
    'according to standardized analytic rubrics.'
)
```

Gambar 2. *Prompt system message*

Pada desain *prompt system message* digunakan untuk menerapkan prinsip role prompting dengan mendefinisikan identitas model sebagai penilai esai berpengalaman.

```
    user_msg = f"""You are grading a student essay for the prompt: "{row['prompt_name']}".

Writing Assignment:
{assignment[:500] if assignment else '(not provided)'}
{source_ctx}
Score Range: {min_s} (lowest) to {max_s} (highest).

Follow these Chain-of-Thought steps, then give the final score:
STEP1 - CLAIM & FOCUS: Does the essay present a clear central claim relevant to the prompt?
STEP2 - EVIDENCE & SUPPORT: Does the essay cite evidence from the source text effectively?
STEP3 - ORGANIZATION: Is the essay logically structured with introduction, body, conclusion?
STEP4 - LANGUAGE & STYLE: Evaluate vocabulary, sentence variety, and overall writing quality.
STEP5 - MECHANICS: Note grammar, spelling, punctuation errors and their impact.
STEP6 - HOLISTIC JUDGMENT: Synthesize all dimensions into a final holistic score.

Respond EXACTLY in this format:
STEP1: <your analysis>
STEP2: <your analysis>
STEP3: <your analysis>
STEP4: <your analysis>
STEP5: <your analysis>
STEP6: <your analysis>
SCORE: <single integer between {min_s} and {max_s}>

Student Essay:
\"\"\"
{essay_text}
\"\"\"
"""
```

Gambar 3. *Prompt user message*

Pada desain prompt user message digunakan untuk memberikan perintah penilaian esai serta menjelaskan bagaimana urutan penilaian dilakukan beserta instruksi format output yang ketat guna mengurangi variabilitas respons akibat sifat stokastik LLM. Pada prompt ini, variabel {row[‘prompt_name’]}, {assignment}, {source_ctx}, {min_s}, {max_s} dan {essay_text} merupakan variabel yang diganti secara dinamis sesuai dengan topik soal dan isi esai per masing-masing.

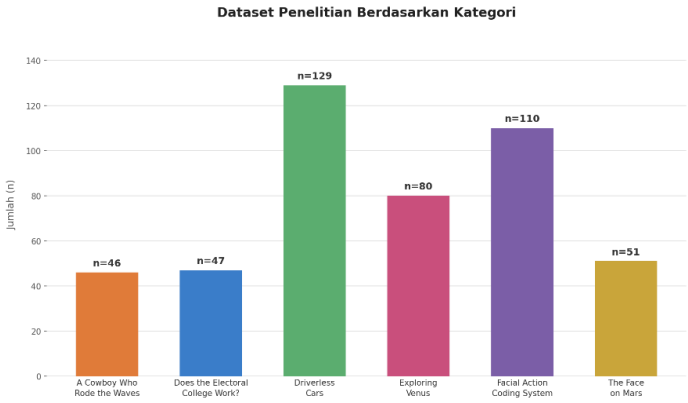
Pada pendekatan *Fine-tuning*, pendekatan ini tidak dilakukan secara penuh melainkan menggunakan pendekatan Parameter-Efficient *Fine-tuning* (PEFT) dengan metode LoRA (Low-Rank Adaptation), sehingga hanya sebagian kecil parameter yang diperbarui selama pelatihan tanpa

mengubah seluruh bobot model. Proses pelatihan dijalankan secara lokal menggunakan GPU mandiri dengan memanfaatkan kuantisasi untuk efisiensi memori. Adapun proses pelatihan dijalankan melalui pendekatan *Supervised Fine-Tuning* (SFT) dengan menggunakan library TRL (*Transformer Reinforcement Learning*) dari Hugging Face dan dijalankan selama 3 *epoch* dengan *learning rate* sebesar 2 x 10^{-4}. Untuk mengatasi keterbatasan kapasitas memori GPU, ukuran batch diatur seminimal mungkin yaitu sebesar 1 (*batch size* = 1). Setelah proses *Fine-tuning* selesai, model yang telah diadaptasi digunakan untuk menilai esai pada data uji secara batch, menghasilkan skor numerik dan umpan balik tekstual yang mencerminkan distribusi penilaian manusia yang telah dipelajari

HASIL DAN PEMBAHASAN

Dataset Penelitian

Penelitian ini dilakukan dengan menggunakan dataset ASAP-AES 2.0 yang didapat dari Kaggle. Dataset ASAP-AES 2.0 terdiri dari 24.728 esai bahasa Inggris yang mencakup berbagai kategori soal dan telah dinilai oleh manusia dengan skala skor 1 hingga 6. Pada penelitian ini menggunakan 500 esai yang diambil secara acak karena keterbatasan spesifikasi GPU. Hal ini menyebabkan keseluruhan dataset tidak memungkinkan untuk diproses. Dari 500 esai tersebut, data dibagi menjadi 400 esai (80%) sebagai data latih dan 100 esai (20%) sebagai data uji untuk pendekatan *Fine-tuning*. Adapun pada penelitian ini jumlah dataset berdasarkan kategori bisa dilihat pada Gambar 4.



Gambar 4. Dataset Penelitian Berdasarkan Kategori

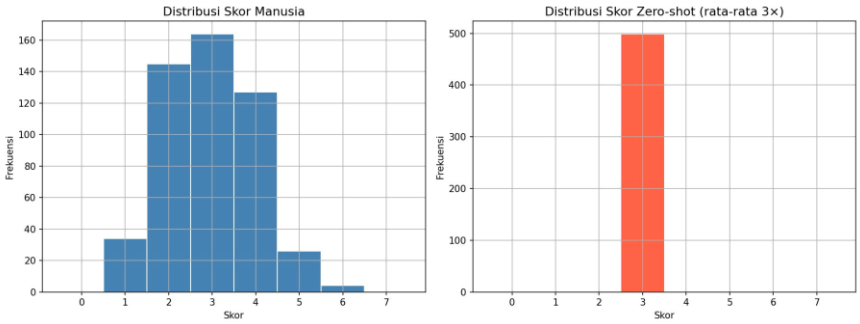
Evaluasi Pendekatan *Zero-shot*

Pendekatan *Zero-shot* dilakukan dengan model Qwen 2.5-1.5B-Instruct tanpa adanya proses training model menggunakan 500 esai. Setiap esai diproses sebanyak tiga percobaan dan skor akhir yang dipakai adalah rata-rata yang dibulatkan dari ketiga percobaan. Tabel 3 menunjukkan hasil QWK dan RMSE dari pendekatan *Zero-shot*.

Tabel 3. Hasil QWK dan RMSE *Zero-shot*

Percobaan	QWK	RMSE
1	0.0069	1.0469
2	0.0069	1.0469
3	0.0069	1.0469
HASIL	**0.0069**	**1.0469**

Berdasarkan Tabel 3, nilai QWK dan RMSE menghasilkan nilai yang identik pada ketiga percobaan. Kesamaan nilai ini menandakan model Qwen2.5-1.5B-Instruct dengan pendekatan *Zero-shot* bersifat deterministik. Nilai QWK sebesar 0.0069 menunjukkan tingkat kesepakatan yang sangat rendah antara skor prediksi dengan penilaian asli. Nilai RMSE sebesar 1.0469 juga menandakan bahwa model hanya menebak satu angka aman untuk semua data. Hal ini mengindikasikan adanya *Central tendency bias* ekstrem. Perbandingan skor manusia dan hasil prediksi yang dihasilkan pendekatan *Zero-shot* dapat dilihat pada Gambar 5.



Gambar 5. Hasil Prediksi Zero-shot dan Skor Manusia

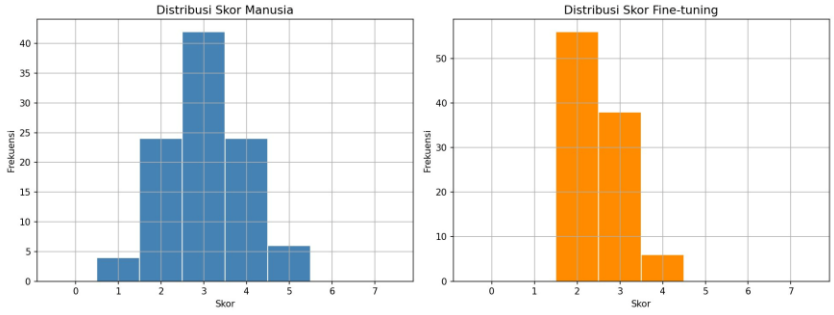
Evaluasi Pendekatan *Fine-tuning*

Pendekatan *Fine-tuning* dilakukan menggunakan metode Low-Rank Adaptation (LoRA). Tabel 4 menunjukkan hasil nilai *Quadratic Weighted Kappa* (QWK) dan RMSE dari pendekatan *Fine-tuning* per kategori esai.

Tabel 4. Hasil QWK dan RMSE *Fine-tuning*

Kategori	QWK	RMSE
A Cowboy Who Rode the Waves	0.4444	0.6325
Car-free cities	-0.1667	1.3229
Does the electoral college work?	0.0909	1.3693
Driverless cars	0.3509	0.9661
Exploring Venus	0.1512	1.1698
Facial action coding system	0.4036	0.8584
The Face on Mars	0.3038	0.6742
KESELURUHAN	**0.3247**	**1.0198**

Berdasarkan Tabel 4, terlihat adanya variasi hasil prediksi di antara masing-masing kategori esai. Pada keseluruhan dataset, model menghasilkan nilai QWK sebesar 0.3247 dan RMSE sebesar 1.0198. Hasil terbaik didapat pada kategori "*A Cowboy Who Rode the Waves*" dengan nilai QWK 0.4444 dan RMSE 0.6325. Sebaliknya, nilai QWK terendah ada pada topik "*Car-free cities*" dengan nilai negatif -0.1667. Nilai ini menandakan prediksi model lebih buruk daripada tebakan acak, kemungkinan besar dipicu oleh ketidakseimbangan data yang ekstrem. Perbandingan skor manusia dan hasil prediksi yang dihasilkan pendekatan *Fine-tuning* dapat dilihat pada Gambar 6.



Gambar 6. Perbandingan Hasil Fine-tuning dan Skor Manusia

Perbandingan *Zero-shot* vs *Fine-tuning*

Perbandingan antara hasil pendekatan *Zero-shot* dan *Fine-tuning* dapat dilihat pada Tabel 5.

Tabel 5. Perbandingan *Zero-shot* vs *Fine-tuning*

Dimensi	*Zero-shot*	*Fine-tuning* (LoRA)
Model	Qwen2.5-1.5B-Instruct	Qwen2.5-1.5B-Instruct + LoRA

Dimensi	*Zero-shot*	*Fine-tuning* (LoRA)
Data yang digunakan	500 esai (inferensi langsung)	400 latih / 100 uji
QWK (keseluruhan)	0.0069 (Poor)	0.3247 (Fair)
RMSE (keseluruhan)	1.0469	1.0198
Ketepatan skor (Δ=0)	32.8%	40.0%
Over-rating	35.6%	9.0%
Under-rating	31.6%	51.0%
Bias utama	*Central tendency bias* (100% skor = 3)	*Under-rating* (*conservative scoring*)
Stabilitas (Std Dev)	0.00 (deterministik)	N/A (inferensi tunggal)

Berdasarkan Tabel 5, dapat dilihat bahwa pendekatan *Fine-tuning* unggul pada seluruh metrik evaluasi. Peningkatan QWK sebesar +0.3178 dan penurunan RMSE sebesar -0.0271 membuktikan bahwa pendekatan *Fine-tuning* dapat digunakan untuk membuat AES yang cukup. Keunggulan ini terjadi karena proses Fine-tuning melalui LoRA berhasil memperbarui bobot parameter model secara spesifik. Hal ini memungkinkan model mengenali pola penilaian esai meskipun tidak mengubah keseluruhan parameter model. Sedangkan pada pendekatan *Zero-shot* meskipun sudah ditambahkan dengan teknik *prompting* canggih seperti *Role Prompting* dan *Chain-of-Thought*, tidak mampu memahami konteks penilaian esai yang spesifik. Kegagalan *Zero-shot* pada model Qwen2.5-1.5B-Instruct ini menandakan adanya *reasoning bottleneck*. Model dengan parameter kecil cenderung mengalami *central tendency bias* yang ekstrem Ketika dipaksa memproses *prompt* berlapis tanpa adanya contoh (few-shot) atau pembaruan bobot parameter model.
Temuan ini sejalan dengan penelitian sebelumnya oleh (Huang et al., 2026) yang mengatakan bahwa pendekatan *Zero-shot* lemah pada model GPT-4o. Penelitian oleh (Mughal et al., 2026) juga memperoleh rata-rata QWK 0.45 pada pendekatan *Zero-shot* dengan GPT-3.5-turbo dan Gemini-Pro. Fakta bahwa LLM *closed-source* seperti GPT-4o dan Gemini-Pro yang menunjukkan keterbatasan pada mode *Zero-shot* mempertegas kesimpulan bahwa mengandalkan *prompt engineering* saja tidak cukup pada model LLM *open-weight* berukuran kecil. Oleh karena itu, diperlukan adanya intervensi *Fine-tuning* untuk memperbaiki hasil prediksi.

SIMPULAN

Berdasarkan penelitian pada implementasi model Qwen 2.5-1.5B-Instruct untuk AES dengan dataset ASAP-AES 2.0, kesimpulan dari penelitian ini adalah sebagai berikut:

1. Pendekatan *Zero-shot* dengan Teknik *Role Prompting* dan Chain-of-Thought menghasilkan performa yang sangat rendah dengan nilai QWK 0.0069 dan RMSE 1.0469. Model ini mengalami *Central tendency bias* ekstrem dengan memberikan skor 3 pada 100% esai, sehingga tidak dapat digunakan pada AES.
2. Pada analisis stabilitas *Zero-shot*, pendekatan ini selalu memberikan skor 3 pada 3x perulangan yang mengindikasikan bahwa model bersifat deterministik total. Stabilitas ini merupakan manifestasi dari *Central tendency bias* dan bukan mencerminkan konsistensi penilaian yang sesungguhnya.
3. Pendekatan *Fine-tuning* dengan PEFT/LoRA meningkatkan QWK secara signifikan menjadi 0.3247 (kategori fair agreement) dengan hanya melatih 0.51% dari total parameter model. Namun, *Fine-tuning* menimbulkan bias *Under-rating* pada 51.0% esai sebagai efek samping dari *conservative scoring*.
4. Perbandingan kedua pendekatan menunjukkan bahwa model LLM *Open-weight* dapat menggunakan pendekatan *Fine-tuning* untuk menghasilkan sistem AES yang cukup.

Saran untuk penelitian selanjutnya adalah untuk mengeksplorasi konfigurasi LoRA yang lebih optimal, menerapkan *few-shot prompting* sebagai alternatif, menggunakan model Qwen 2.5 dengan ukuran parameter yang lebih besar (14B atau 32B) dan menggunakan dataset yang lebih beragam

DAFTAR PUSTAKA

Adha, R., Syaripudin, D., Kurniawan, B. S., Mukharil Bachtiar, A., Maulana, H., Rainarli, E., & Sistem Informasi, D. (2026). Integrasi Gemini AI Berbasis Retrieval-Augmented Generation (RAG) pada Moodle untuk Penilaian Esai Otomatis dengan Pendekatan Human-in-the-Loop di Pendidikan Tinggi. In *AJCSR [Academic Journal of Computer Science Research* (Vol. 8, Number 1). https://doi.org/10.38101/ajcsr.v8i1.16253.

Alawadh, H. M., Meraj, T., Aldosari, L., & Tayyab Rauf, H. (2024). An Efficient Text-Mining Framework of Automatic Essay Grading Using Discourse Macrostructural and Statistical Lexical Features. *SAGE Open*, *14*(4). https://doi.org/10.1177/21582440241300548.

AlGhamdi, E. M., Li, Y., Gašević, D., & Chen, G. (2026). Leveraging *prompt*-based LLMs for automated scoring and feedback generation in higher education. *Computers and Education*, *243*. https://doi.org/10.1016/j.compedu.2025.105511.

Bouaziz, A., Tayaa, K., & Menci, A. (2025). Evaluation and Assessment of Essays in Higher Education: Practices, Challenges, and Insights. *Анали Филолошког Факултета*, *37*(2), 195–218. https://doi.org/10.18485/analiff.2025.37.2.10.

Choi, H., Kang, M. C., Seong, J., & Huang, J. X. (2026). Exploring *Zero-shot* essay scoring: from feature-based to LLM-based approaches. *Data Mining and Knowledge Discovery*, *40*(3). https://doi.org/10.1007/s10618-026-01193-z.

Diningrat, M. S. M., Mabruroh, F., Widya, H., & Nurhamidah, N. (2025). The Application of Machine Learning Algorithms in Analyzing Students' Conceptual Error Patterns in Science Learning. *Jurnal Pendidikan Dan Ilmu Fisika*, *5*(1), 54–65. https://doi.org/10.52434/jpif.v5i1.42586.

Faza, M. N., Purnamasari, P. D., & Ratna, A. A. P. (2025). Defying Data Scarcity: High-Performance Indonesian Short Answer Grading via Reasoning-Guided Language Model *Fine-tuning*. *International Journal of Electrical, Computer, and Biomedical Engineering*, *3*(3). https://doi.org/10.62146/ijecbe.v3i3.148.

García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C., & Bekerman, Z. (2025). ChatGPT as a Stable and Fair Tool for Automated Essay Scoring. *Education Sciences*, *15*(8). https://doi.org/10.3390/educsci15080946.

Huang, Y., Palermo, C., & Wilson, J. (2026). Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, feature-based models, and human raters. *Assessing Writing*, *69*. https://doi.org/10.1016/j.asw.2026.101047.

Liu, T., Ye, L., & Yan, W. (2026). A framework for evaluation of *Large Language Model*s in essay assessment: Reliability, alignment, and causal reasoning. *Computers and Education: Artificial Intelligence*, *10*. https://doi.org/10.1016/j.caeai.2026.100565.

Mahmud, T., Rabbi, M. F., Hossain, M. T., Talukder, A. A., & Hasan, M. K. (2025). Portfolio assessment for developing higher order thinking skills in Bangladeshi undergraduate EFL writing classes. *Language Testing in Asia*, *15*(1). https://doi.org/10.1186/s40468-025-00404-6.

Martin, F., Kim, S., Bolliger, D. U., & DeLarm, J. (2025). Assessment Types, Strategies, and Feedback in Online Higher Education Courses in the Age of Artificial Intelligence: Perspectives of Instructional Designers. *TechTrends*, *69*(6), 1330–1346. https://doi.org/10.1007/s11528-025-01115-8.

Mughal, N., Imran, A. S., Daudpota, S. M., Kastrati, Z., & Noor, W. (2026). Exploring potential of *Large Language Model*s for automated essay scoring in education. *Discover Artificial Intelligence*, *6*(1). https://doi.org/10.1007/s44163-026-01002-y.

Pulung Hendro Prastyo, Eddy Tungadi, & Shaifudin Zuhdi. (2025). Indonesian Automated Essay Scoring: A Comparative Study of Pretrained Transformer Models. *Information Technology Education Journal*, 120–130. https://doi.org/10.59562/intec.v4i2.8069.

Rizqi Amelia, A., & Tommy S Suyasa, P. Y. (2025). THE IMPACTS OF SELF-REGULATED LEARNING ON UNIVERSITY STUDENTS: A SCOPING REVIEW. *International Journal of Application on Social Science and Humanities*, *3*(2), 61–72. https://doi.org/10.24912/ijassh.v3i2.35510.

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2025). Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs. *15th International Conference on Learning Analytics and Knowledge, LAK 2025*, 293–305. https://doi.org/10.1145/3706468.3706507.