Vol. 6, No. 2, Agustus 2026, Hal. 124-132

Jurnal Software Engineering and Information System (SEIS)

https://ejurnal.umri.ac.id/index.php/SEIS/index

e-ISSN: 2809-0950

MITIGASI BIAS LARGE LANGUAGE MODEL MELALUI *HUMAN-IN-THE-LOOP* PADA *AUTOMATED ESSAY SCORING* BERBASIS *ESAI IELTS-LIKE*

Muhammad Abdiel Al Hafiz[1*)], M. Syaiful Amin[2)]
[1,2]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto
email: [1]hafd324@gmail.com, [2]syaifulamin@amikompurwokerto.ac.id
*Corresponding Author

Abstract

This study evaluates the reliability, algorithmic bias, and feasibility of implementing Human-in-the-Loop (HITL) in Large Language Model (LLM)-based Automated Essay Scoring (AES). Two architectures were compared: the Mixture-of-Experts architecture represented by GPT OSS 120B and the Dense architecture represented by Qwen3-32B, using a multiple-run scoring approach on 150 IELTS-like essays written by 50 respondents. Each essay was evaluated five times across four criteria: Grammar, Lexical Resource, Coherence, and Task Achievement. The evaluation employed the Intraclass Correlation Coefficient (ICC), Coefficient of Variation (CV), score range, paired t-test, and confidence-based routing simulation.The results show that GPT OSS 120B achieved higher reliability, with an ICC of 0.94, an average range of 2.40 points, and a CV of 4.69%, while Qwen3-32B obtained an ICC of 0.84, an average range of 5.56 points, and a CV of 9.17%. However, GPT OSS 120B experienced a parsing failure rate of 25.4%, whereas Qwen3-32B demonstrated full format compliance. Comparative analysis also showed that GPT OSS 120B tended to score more strictly, while Qwen3-32B was more lenient. In the Data Report task without visual input, both models exhibited conservatism bias due to contextual limitations. The HITL simulation showed that GPT OSS 120B could automatically approve 55.4% of essays, compared with only 7.8% for Qwen3-32B. These findings highlight the importance of HITL in maintaining the reliability, fairness, and integrity of academic assessment.

Keywords: *Automated Essay Scoring, Algorithmic Bias, Human-in-the-Loop, Large Language Model*

Abstrak

Penelitian ini mengevaluasi reliabilitas, bias algoritmik, dan kelayakan penerapan Human-in-the-Loop (HITL) pada Automated Essay Scoring (AES) berbasis Large Language Model (LLM). Dua arsitektur dibandingkan, yaitu Mixture-of-Experts pada GPT OSS 120B dan Dense pada Qwen3-32B, menggunakan pendekatan multiple-run scoring terhadap 150 esai IELTS-like dari 50 responden. Setiap esai dinilai lima kali pada empat kriteria, yaitu Grammar, Lexical Resource, Coherence, dan Task Achievement. Evaluasi dilakukan menggunakan Intraclass Correlation Coefficient (ICC), Coefficient of Variation (CV), Range, uji t berpasangan, serta simulasi confidence-based routing. Hasil penelitian menunjukkan bahwa GPT OSS 120B memiliki reliabilitas lebih tinggi dengan ICC 0,94, rata-rata Range 2,40 poin, dan CV 4,69%, sedangkan Qwen3-32B memperoleh ICC 0,84, Range 5,56 poin, dan CV 9,17%. Namun, GPT OSS 120B mengalami parsing failure sebesar 25,4%, sementara Qwen3-32B menunjukkan kepatuhan format penuh. Uji komparatif juga menunjukkan bahwa GPT OSS 120B cenderung lebih ketat, sedangkan Qwen3-32B lebih longgar dalam memberikan skor. Pada tugas Data Report tanpa input visual, kedua model menunjukkan conservatism bias akibat keterbatasan konteks. Simulasi HITL menunjukkan GPT OSS 120B mampu mengotomatisasi 55,4% penilaian, sedangkan Qwen3-32B hanya 7,8%. Temuan ini menegaskan pentingnya HITL untuk menjaga reliabilitas, keadilan, dan integritas penilaian akademik.

Kata kunci: *Penilaian Esai Otomatis, Bias Algoritmik, Human-in-the-Loop, Large Language Model*

PENDAHULUAN

Integrasi Kecerdasan Buatan (AI) dalam sistem pendidikan tinggi mengalami perkembangan pesat dalam lima tahun terakhir, tidak hanya pada aspek administrasi, namun juga merambah ke domain evaluasi akademik formatif dan sumatif (Nuong et al., 2024). Di tengah transformasi ini, sistem *Automated Writing Evaluation* (AWE) muncul sebagai solusi teknologi krusial untuk penilaian kemampuan menulis berbahasa Inggris. Berbeda dengan metode konvensional yang intensif secara waktu, AWE menawarkan umpan balik instan dan konsisten (Mat et al., 2024). Namun, penelitian menunjukkan bahwa sistem AWE masih memiliki keterbatasan fundamental dalam menilai *higher-order thinking skills* seperti substansi dan koherensi argumentasi. Tinjauan sistematis terhadap studi empiris terkini menegaskan bahwa AWE efektif meningkatkan *surface features* (akurasi gramatikal dan fluensi), tetapi dampaknya terhadap keterampilan berpikir kritis tetap terbatas, disertai kekhawatiran mahasiswa mengenai *overreliance* terhadap sistem otomatis (Che Mat et al 2024; Du & Nordin, 2025).

Meskipun menjanjikan efisiensi, teknologi AWE menghadapi kekhawatiran fundamental terkait keadilan algoritmik (*algorithmic fairness*) dan keandalan (*reliability*). Penelitian empiris menunjukkan bahwa sistem penilaian otomatis dapat mendemonstrasikan bias demografis signifikan, di mana penutur bahasa asing menerima klasifikasi *false positive* lebih tinggi dibandingkan penutur asli, menciptakan disparitas sistematis dalam penilaian *high-stakes* (Zehner et al., 2025). Fenomena ini diperkuat oleh temuan mengenai *trade-off* akurasi-bias, di mana alat evaluasi teks berbasis AI dengan akurasi tinggi justru menunjukkan bias lebih kuat terhadap penulis non-penutur asli (Pratama, 2025). Kompleksitas ini menggarisbawahi bahwa sistem penilaian otomatis tidak hanya harus akurat secara statistik, tetapi juga adil secara substantif.

Kemajuan *Large Language Model* (LLM) membuka kemungkinan baru bagi pengembangan sistem penilaian yang lebih mutakhir, namun juga memperkenalkan tantangan konsistensi (Pack et al., 2024). Secara arsitektural, LLM untuk pemrosesan bahasa alami saat ini didominasi oleh dua pendekatan komputasional yang berbeda secara fundamental (Patwardhan et al., 2023). Pada arsitektur *Dense* konvensional (seperti Qwen3-32B), seluruh parameter jaringan saraf diaktifkan secara bersamaan setiap kali model memproses sebuah token. Hal ini menuntut beban komputasi besar dan rentan memicu instabilitas saat mengevaluasi data yang kompleks. Sebaliknya, arsitektur *Mixture of Experts* (MoE) seperti GPT OSS 120B beroperasi secara lebih cerdas dan efisien melalui mekanisme perutean (*routing*) yang hanya mengaktifkan sebagian kecil parameter pakar spesifik yang paling relevan untuk setiap tugas komputasi tertentu. Berdasarkan bukti empiris dari literatur komputasional terbaru, arsitektur MoE telah terbukti secara meyakinkan mampu melampaui model *Dense* konvensional dalam hal efisiensi kapasitas dan keseimbangan antara kecepatan dan akurasi komputasi, bahkan ketika dievaluasi menggunakan metrik waktu per langkah yang ketat (Du et al., 2024). Superioritas ini bukanlah ilusi dari sekadar bertambahnya jumlah parameter. Evaluasi metodologis yang sistematis membuktikan bahwa MoE mampu mengungguli model *Dense* bahkan di bawah kendala sumber daya yang sepenuhnya setara. Keunggulan ini tetap dipertahankan ketika total parameter, anggaran komputasi, dan volume data diatur secara identik, asalkan model beroperasi pada tingkat aktivasi yang ideal (Li et al., 2025). Kapasitas arsitektural inilah yang secara empiris mampu meminimalkan kompresi distribusi skor dalam evaluasi teks dibandingkan pendekatan konvensional (Pack et al., 2024; Yavuz et al., 2025).

Untuk mengukur tingkat konsistensi dan instabilitas pada arsitektur tersebut, penelitian ini menerapkan metodologi *multiple-run scoring*, yakni teknik pengujian di mana instrumen esai yang identik dikirimkan ke model yang sama sebanyak beberapa kali iterasi dengan pengaturan suhu (*temperature*) di atas nol guna memetakan distribusi probabilitas keluaran model (Yavuz et al., 2025). Melalui pengukuran berulang ini, berbagai manifestasi bias algoritmik dalam sistem penilaian dapat dideteksi secara presisi. Bias tersebut mencakup *severity bias* (bias keketatan, di mana model secara konsisten bertindak terlalu ketat atau terlalu longgar dalam memberikan nilai), *topic bias* (keberpihakan topik, di mana skor dipengaruhi secara tidak proporsional oleh tema soal), serta *conservatism bias* (bias kehati-hatian, fenomena saat model memampatkan skor ke nilai tengah/rendah secara defensif saat dihadapkan pada defisit informasi).

Mengingat tingginya risiko bias tersebut, penerapan tata kelola AI (*AI governance*) menjadi keharusan mutlak ketika mendelegasikan evaluasi pendidikan kepada algoritma guna menghindari diskriminasi pada tes berdampak tinggi (Nuong et al., 2024). Tata kelola AI tidak hanya mewajibkan pengujian keadilan algoritmik secara berkala (Caton & Haas, 2024; Andersen et al., 2025; Laclau et al., 2024), tetapi juga mewajibkan adopsi pengawasan manusia di dalam siklus keputusan. Tantangan seperti *distribution shifts*—di mana model yang adil di domain sumber menjadi tidak adil di domain target—membuktikan bahwa mitigasi bias secara

murni oleh mesin tidaklah cukup (Shao et al., 2024; Siddique et al., 2023).

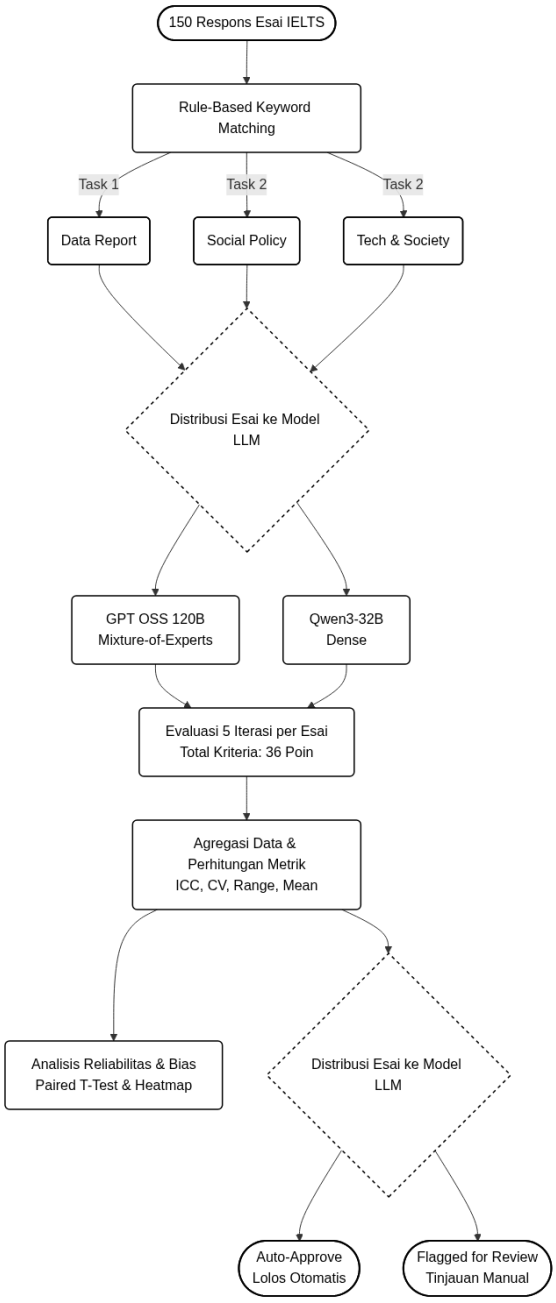
Oleh karena itu, kerangka kerja *Human-in-the-Loop* (HITL) menjadi solusi operasional yang esensial. Dalam konteks penilaian esai, HITL diimplementasikan melalui *confidence-based routing*, yakni mekanisme perutean bersyarat di mana sistem menyeleksi dokumen berdasarkan tingkat keyakinan (stabilitas skor) dari algoritma. Melalui tata kelola berbasis HITL ini, LLM dibatasi perannya sebagai prapemroses analitis semata; mesin dapat meloloskan esai yang dinilai secara konsisten untuk disetujui secara otomatis (*Auto-Approve*), namun diwajibkan untuk mengeskalasi esai dengan variabilitas tinggi (*low-confidence*) kepada pendidik manusia untuk divalidasi ulang (Selvam & González Vallejo, 2025).

Berdasarkan uraian literatur tersebut, teridentifikasi celah (*gap*) penelitian yang krusial: meskipun masalah reliabilitas dan bias LLM telah banyak disorot, ketersediaan kerangka kerja praktis dengan metrik yang definitif untuk deteksi instabilitas secara *real-time* pada skenario HITL masih sangat terbatas. Oleh karena itu, kebaruan (*novelty*) dari penelitian ini terletak pada perumusan parameter empiris baru untuk ambang batas perutean *Human-in-the-Loop* yang direpresentasikan melalui metrik stabilitas variabilitas skor (*Coefficient of Variation* dan *Range*). Secara komprehensif, penelitian ini bertujuan untuk: (1) menganalisis dan membandingkan tingkat reliabilitas intrarater antara arsitektur *Dense* (Qwen3-32B) dan *Mixture-of-Experts* (GPT OSS 120B); (2) mengidentifikasi eksistensi *severity bias*, *topic bias*, dan *conservatism bias* pada masing-masing model; serta (3) mensimulasikan kerangka kerja HITL untuk mengevaluasi kelayakan tingkat otonomi LLM dalam ekosistem penilaian pendidikan.

METODOLOGI PENELITIAN

Penelitian Penelitian ini menggunakan pendekatan kuantitatif komparatif dengan desain pengukuran berulang (*repeated measures design*). Fokus utama penelitian adalah membandingkan tingkat reliabilitas dan bias pada arsitektur *Dense* (Qwen3-32B) dan *Mixture-of-Experts* (GPT OSS 120B). Selain itu, penelitian ini mensimulasikan mekanisme *Human-in-the-Loop* (HITL) untuk menentukan ambang batas kepercayaan (*confidence level*) dalam penilaian esai otomatis. Instrumen utama yang digunakan adalah perangkat lunak berbasis Python pada lingkungan *Jupyter Notebook* yang terhubung dengan *Groq Cloud API*.

Dataset utama terdiri dari 150 respons esai yang dikerjakan oleh 50 responden. Responden merupakan mahasiswa semester enam dengan tingkat kemahiran bahasa Inggris pada level menengah (*intermediate*). Pengumpulan data dilakukan secara terpusat melalui platform penilaian bahasa Inggris berbasis web yang diintegrasikan sebagai bagian dari proyek riset internal mata kuliah bahasa Inggris. Instrumen soal terdiri dari tiga *prompt IELTS-like*, yang mencakup satu tugas deskripsi data (Task 1) dan dua tugas diskusi opini (Task 2). Seluruh esai dikonversi ke format CSV dengan proses anonimisasi identitas mahasiswa. Secara keseluruhan, tahapan pemrosesan data penelitian ini dirancang dalam alur kerja sekuensial yang diilustrasikan pada Gambar 1.



Gambar 1. Alur kerja *multiple-run scoring* dan simulasi *Human-in-the-Loop*

Berdasarkan Gambar 1, alur pemrosesan data terbagi menjadi tiga tahapan utama:

1. ***One-Shot Rule-Based Topic Categorization* (*Preprocessing*)**

Untuk mendeteksi bias algoritmik secara terstruktur tanpa mengintroduksi bias baru dari sistem AI tambahan (*confounding variable*), ketiga instrumen *prompt* esai dikategorikan menggunakan pendekatan *rule-based keyword matching*. Atribut kata kunci dan kategori pada tahap prapemrosesan ini ditentukan berdasarkan kerangka *IELTS like Writing Task* dan divalidasi melalui pertimbangan ahli (*expert judgment*) dari tenaga pengajar bahasa Inggris. Pemetaan deterministik ini mengeksekusi skrip Python untuk mengelompokkan 3 *prompt* esai secara otomatis ke dalam tiga kategori definitif. Rincian instrumen soal dan hasil pemetaan kategorinya disajikan pada Tabel 1.

Tabel 1. Instrumen Soal (*Prompt*) Penulisan Esai IELTS *Like*

Kategori Topik	Klasifikasi Tugas	Deskripsi Soal (Prompt)
Data Report	Task 1	*The charts above show the results of a survey of adult education. The first chart shows the reasons why adults decide to study. The pie chart shows how people think the costs of adult education should be shared. Write a report for a university lecturer, describing the information shown above.*
Social Policy Opinion	Task 2	*Some think that governments should support retired people financially while others believe they should take care of themselves. Discuss both views and give your own opinion.*
Tech & Society Opinion	Task 2	*Some people think that robots are very important to human future development. Other think that they are dangerous and have negative effect on society discuss both view and give your opinion.*

Terkait instrumen pada kategori *Data Report* (Task 1), data visual yang menjadi rujukan soal sengaja tidak didistribusikan ke dalam *prompt* evaluasi sistem. Hal ini dirancang secara khusus sebagai skenario uji stres unimodal (*unimodal stress testing*) untuk mengobservasi respons perilaku model bahasa saat melakukan evaluasi faktual (kriteria *Task Achievement*) dalam kondisi defisit konteks visual.

2. ***Multiple-Run Scoring Execution***

Esai Esai yang telah terklasifikasi dikirimkan ke model GPT OSS 120B dan Qwen3-32B menggunakan skema instruksi JSON yang ketat. Format *system prompt* yang digunakan untuk menginstruksikan model agar mematuhi rubrik disajikan pada Tabel 2.

Tabel 2. *System Prompt* Evaluasi Berbasis Rubrik JSON

Elemen	Deskripsi Instruksi (*System Prompt*)
Role & Output	*You are an objective, rubric-based grader. INSTRUCTIONS: Output EXACTLY one JSON object and NOTHING else. Do NOT include explanations, reasoning, or any additional text. No bullet lists, no code fences, no commentary.*
Schema	*OUTPUT SCHEMA: {"grammar": int, "lexical": int, "coherence": int, "task": int }- Each value MUST be an integer between 0 and 9 (inclusive). Round to the nearest integer if needed.*
Briefs	*- grammar: accuracy and range of grammar, spelling, punctuation. - lexical: vocabulary range and word choice appropriateness. - coherence: organization, logic, and flow of ideas. - task: how well the response addresses the prompt and fulfills requirements.*
Rules	*RULES: If uncertain, choose the closest integer. If the response cannot be scored, set all fields to null. Do NOT reveal chain-of-thought. Keep the JSON compact (no extra spaces or newlines).*

Untuk menguji konsistensi internal model, penilaian dilakukan sebanyak lima kali iterasi (*run*) untuk setiap esai pada masing-masing model dengan pengaturan suhu (*temperature*) 0,7. Penggunaan parameter suhu 0,7 memberikan keseimbangan operasional yang ideal antara determinisme prediktif dan penalaran analitis (Yavuz et al., 2025). Sementara itu, pelaksanaan lima iterasi diterapkan sebagai standar minimum untuk secara akurat menangkap varians stokastik dan distribusi probabilitas yang inheren pada arsitektur LLM (Pack et al., 2024). Skor dievaluasi berdasarkan empat kriteria standar IELTS: *Grammar, Lexical Resource, Coherence,* dan *Task Achievement*.

3. **Data Aggregation & Evaluation**

Hasil dari 1.500 observasi (150 respons esai × 2 model × 5 run) diagregasi menggunakan pustaka Pandas. Evaluasi dilakukan melalui analisis reliabilitas menggunakan *Intraclass*

Correlation Coefficient (ICC) dan *Coefficient of Variation* (CV). Penggunaan ICC dengan spesifikasi efek campuran dua arah (*two way mixed effects*) dan tipe *absolute agreement* diterapkan karena subjek evaluasi (esai) diambil secara acak dari populasi, sedangkan penilainya (LLM spesifik) merupakan parameter yang tetap (Yavuz et al., 2024). Spesifikasi ini memastikan bahwa evaluasi menghitung kesesuaian nilai eksak yang diberikan oleh model pada setiap iterasi, bukan sekadar konsistensi peringkat relatifnya, sehingga menghasilkan estimasi reliabilitas intrarater yang terukur dan objektif (Pack et al., 2024).

Analisis bias kategori dilakukan dengan membandingkan rerata skor antar tiga kategori prompt, sementara indikasi perbedaan ketatnya penilaian relatif antar model (*relative severity bias*) divalidasi menggunakan Uji T Berpasangan (*Paired T Test*). Pendekatan komparatif relatif ini diambil mengingat penelitian tidak melibatkan penguji manusia sebagai standar emas. Terakhir, hasil agregasi metrik digunakan untuk mensimulasikan kerangka HITL. Mengadopsi konsep *confidence based routing* yang diusulkan oleh Selvam dan González Vallejo (2025), penelitian ini mengusulkan parameter empiris baru untuk mendefinisikan *confidence level* berbasis variabilitas skor pada skala 0 hingga 36. Karena model LLM tidak selalu menyediakan probabilitas logit yang transparan, stabilitas skor lintas iterasi digunakan sebagai proksi reliabilitas (Yavuz et al., 2025).

Esai dikategorikan sebagai lolos otomatis (*Auto Approve*) jika memiliki stabilitas tinggi, yang didefinisikan secara operasional dengan *Coefficient of Variation* (CV) < 15% dan rentang perbedaan skor maksimum (*Range*) ≤ 2. Batasan ini diusulkan sebagai proksi empiris awal yang mengadaptasi standar kesepakatan penguji manusia (*human rater agreement threshold*) yang umumnya mensyaratkan korelasi tinggi (misalnya koefisien Pearson ≥ 0,75) sebelum skor dianggap valid dan terkalibrasi dengan baik (Pack et al., 2024). Sebaliknya, esai yang melampaui ambang batas variabilitas ini ditandai (*flagged*) untuk peninjauan manual.

HASIL DAN PEMBAHASAN

Penelitian ini mengevaluasi kinerja dua arsitektur *Large Language Model* (LLM), yakni arsitektur *Mixture-of-Experts* (GPT OSS 120B) dan arsitektur *Dense* (Qwen3-32B), dalam menjalankan fungsi *Automated Essay Scoring* (AES) terhadap 150 respon esai IELTS *like*. Evaluasi dilaksanakan melalui metode *multiple-run scoring* (lima iterasi per esai) untuk mengukur tingkat reliabilitas intrarater, mendeteksi eksistensi bias algoritmik, serta mensimulasikan mekanisme *Human-in-the-Loop* (HITL). Skor akhir yang dianalisis menggunakan skala mentah 0–36 poin, yang merepresentasikan akumulasi dari empat kriteria rubrik resmi IELTS dengan bobot skor maksimal sembilan per kriteria.

Sebanyak 150 observasi yang terdiri dari pasangan dokumen dan model diajukan untuk proses evaluasi otomatis. Jumlah maksimal teoritis ini merupakan hasil perkalian dari 50 responden dengan tiga kategori pertanyaan yang berbeda. Pada pelaksanaannya, sistem mencatat penyusutan sampel (*data dropout*) akibat kendala *time out* koneksi server serta perbedaan tingkat kepatuhan format dari masing masing arsitektur model. Rincian penyusutan sampel observasi pada setiap tahapan evaluasi dapat dilihat pada Tabel 3.

Tabel 3. Rincian Penyusutan Sampel Observasi

Tahapan	GPT OSS 120B	Qwen3-32B
Maksimal teoritis(50 responden dikali 3 soal)	150 observasi	150 observasi
Terkirim sukses ke server API	130 observasi	128 observasi
Menghasilkan struktur format JSON yang valid	97 observasi	128 observasi
Irisan pasangan esai valid (dievaluasi kedua model)	95 observasi	95 observasi

Berdasarkan Tabel 3, model GPT OSS 120B menerima 130 dokumen esai, namun hanya berhasil merilis skor terstruktur yang valid untuk 97 observasi. Sebanyak 33 esai sisanya mengalami kegagalan penguraian format (*parsing failure*) di mana model gagal mematuhi instruksi JSON secara utuh. Di sisi lain, model Qwen3-32B menerima 128 dokumen esai dan mendemonstrasikan kepatuhan format (*format compliance*) secara mutlak dengan menghasilkan skor valid untuk keseluruhan 128 observasi tersebut. Untuk kebutuhan analisis statistik komparatif yang mensyaratkan kedua model mengevaluasi dokumen yang identik, penelitian ini secara ketat hanya menggunakan irisan sampel sebanyak 95 observasi valid. Seluruh perhitungan metrik lanjutan pada bab ini dirancang merujuk secara transparan pada ukuran sampel dari masing masing tahapan tersebut.

Tahap Tahap awal evaluasi pada penelitian ini difokuskan pada pengukuran konsistensi model

dalam memberikan skor pada instrumen esai yang sama secara berulang. Pengukuran reliabilitas pada tingkat model menggunakan metrik Intraclass Correlation Coefficient (ICC), sedangkan stabilitas individu dievaluasi melalui metrik Range serta Coefficient of Variation (CV).

Tabel 4. Ringkasan Metrik Reliabilitas Penilaian Model

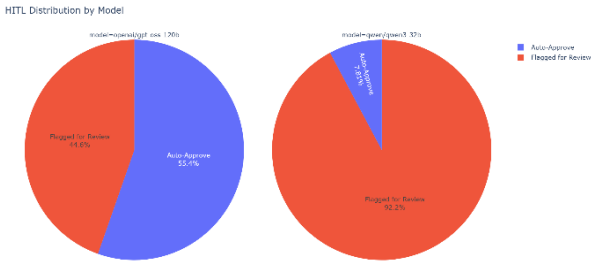
Model	Arsitektur	ICC	Rata-rata *Range*	Rata-rata CV
GPT OSS 120B	Mixture-of-Experts	0.94	2,40 poin	4.69%
Qwen3-32B	Dense	0.84	5,56 poin	9.17%

Berdasarkan hasil pengujian di Tabel 4, model GPT OSS 120B mencatatkan nilai ICC sebesar 0,94 dari 97 observasi valid, yang diklasifikasikan ke dalam kategori reliabilitas sangat tinggi. Sebaliknya, Qwen3-32B mencatatkan ICC sebesar 0,84 dari 128 observasi valid yang tergolong dalam kategori baik, namun mengindikasikan tingkat instabilitas yang signifikan pada metrik tingkat esai. Secara rerata, skor yang didistribusikan oleh Qwen3-32B untuk satu esai yang sama dapat berfluktuasi hingga 5,56 poin di antara lima kali iterasi. Tingkat varians yang tinggi ini secara empiris membuktikan bahwa arsitektur Dense rentan terhadap inkonsistensi operasional saat menangani pemrosesan berulang tanpa pengawasan. Sebaliknya, arsitektur Mixture of Experts pada GPT OSS 120B terbukti secara signifikan lebih stabil dengan rerata deviasi skor absolut hanya sebesar 2,40 poin.

Menyikapi instabilitas metrik pada Qwen3-32B serta kerentanan kepatuhan format pada GPT OSS 120B, penelitian ini langsung mengukur dampak operasionalnya melalui simulasi mekanisme *Human in the Loop* (HITL). Mekanisme ini berfungsi untuk mengklasifikasikan dokumen yang memenuhi syarat otomasi penuh (*auto approve*) dan dokumen yang memerlukan eskalasi untuk tinjauan pakar manusia (*flagged for review*). Sistem menetapkan kriteria gabungan untuk status *auto approve*, yakni rentang deviasi absolut kurang dari atau sama dengan 2 dan *coefficient of variation* kurang dari 15 persen.

Tabel 5. Ringkasan Distribusi Keputusan Human in the Loop

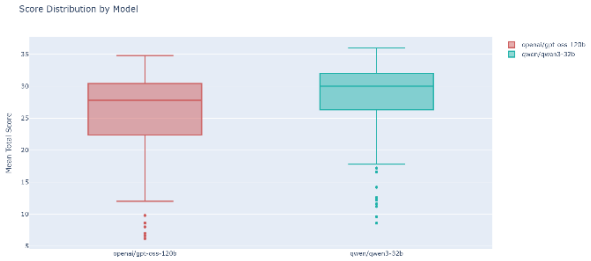
Model	Auto Approve	Flagged for Review	Total Data Valid
GPT OSS 120B	72 esai (55,4%)	58 esai (44,6%)	130 esai
Qwen3-32B	10 esai (7,8%)	118 esai (92,2%)	128 esai



Gambar 2. Distribusi Keputusan Simulasi Kerangka *Human-in-the-Loop*

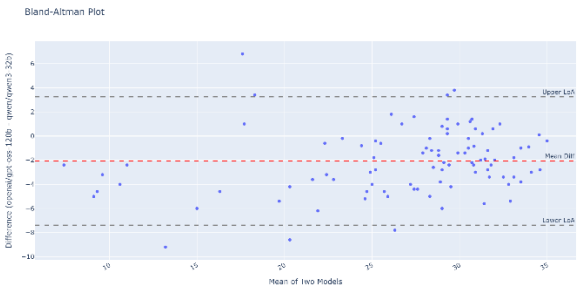
Hasil eksekusi perutean pada Tabel 5 dan Gambar 2 menunjukkan disparitas efisiensi yang sangat signifikan. Sebagai implikasi dari tingkat instabilitas yang tinggi, Qwen3-32B dinilai belum layak diimplementasikan sebagai evaluator mandiri karena tingkat persetujuan otomatisnya hanya sebesar 7,8 persen (10 dari 128 esai). Di sisi lain, GPT OSS 120B menunjukkan performa yang jauh lebih proporsional dengan tingkat persetujuan otomatis mencapai 55,4 persen (72 dari 130 esai). Fakta terpenting dari simulasi ini adalah rasio eskalasi pada GPT OSS 120B yang berjumlah 58 esai. Angka tersebut merupakan akumulasi presisi dari 33 dokumen yang mengalami kegagalan format dan 25 dokumen yang formatnya valid namun melampaui ambang batas variabilitas skor.

Setelah mengevaluasi reliabilitas masing masing model secara mandiri, evaluasi dilanjutkan dengan mengidentifikasi indikasi perbedaan ketatnya penilaian (*relative severity bias*) di antara keduanya. Pengujian ini dianalisis menggunakan uji statistik parametrik uji T berpasangan (*Paired T Test*) secara eksklusif terhadap 95 observasi esai yang beririsan secara valid pada kedua arsitektur. Hasil pengujian menghasilkan nilai t = -7,4253 dengan probabilitas p = 0,000000. Karena nilai p berada jauh di bawah tingkat signifikansi, hipotesis nol ditolak. Hal ini mengonfirmasi eksistensi disparitas standar penilaian yang signifikan antar kedua model. Rerata skor keseluruhan yang diberikan oleh GPT OSS 120B terbukti secara konsisten lebih rendah dibandingkan Qwen3-32B untuk himpunan data yang identik. Visualisasi distribusi skor, yang dikalkulasi berdasarkan rerata nilai dari seluruh lima iterasi pengukuran pada masing masing esai, mengonfirmasi bahwa Qwen3-32B bertindak lebih longgar dengan sering memberikan skor sempurna, sementara GPT OSS 120B memiliki daya pembeda yang lebih ketat.



Gambar 3. Distribusi Skor dan Daya Beda Model

Visualisasi distribusi skor pada Gambar 3, yang dikalkulasi berdasarkan rerata nilai dari seluruh lima iterasi pengukuran pada masing masing esai, mengelaborasi akar presisi dari bias keketatan tersebut. Model GPT OSS 120B menunjukkan kemampuan pembeda (*discriminative power*) yang optimal, dengan rentang antarkuartil (*Interquartile Range/IQR*) selebar 8 poin (Q1 = 22,4 hingga Q3 = 30,4). Sebaliknya, Qwen3-32B mengalami kompresi distribusi (distribution compression) yang terpusat di area skor tinggi dengan rentang IQR yang sempit (Q1 = 26,3 hingga Q3 = 32) serta nilai median 30,0. Kompresi ini membuktikan bahwa Qwen3-32B bertindak sangat longgar (lenient), yang terindikasi dari frekuensi pemberian skor sempurna secara berlebihan, sehingga gagal memetakan kualitas esai secara komprehensif.



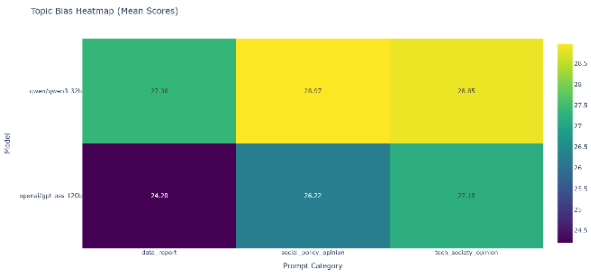
Gambar 4. Uji Kesepakatan Antar-Model (Bland-Altman Plot)

Disparitas kesepakatan antarmodel ini divisualisasikan lebih mendalam melalui Plot Bland Altman pada Gambar 4. Berdasarkan pengujian pada 95 pasangan esai tersebut, garis selisih rerata menunjukkan deviasi negatif sebesar -2,07. Sebaran observasi menjauhi batas toleransi kesepakatan secara asimetris, yang membentang dari batas bawah -7,38 hingga batas atas +3,25. Rentang jarak selebar 10,63 poin ini membuktikan secara visual bahwa meskipun selisih rerata sistematisnya hanya sekitar dua poin, pada kasus esai individu model GPT OSS 120B bisa memberikan skor hingga 7 poin lebih rendah dibandingkan Qwen3-32B. Hal ini menegaskan bahwa kedua arsitektur LLM ini tidak dapat saling menggantikan di dalam ekosistem akademik tanpa kalibrasi yang ekstensif.

Sebagai pendalaman dari analisis bias, penelitian ini juga mengimplementasikan deteksi keberpihakan topik (*topic bias*) untuk mengevaluasi konsistensi model terhadap variasi kompleksitas tugas komputasional (IELTS like Task 1 dan Task 2). Komparasi skor berdasarkan kategori topik tersebut dapat diamati pada Tabel 6.

Tabel 6. Rata-Rata Skor Berdasarkan Kategori Instrumen Uji (*Prompt Category*)

Kategori Instrumen	Klasifikasi Tugas	Mean GPT OSS 120B	Mean Qwen3 32B
Data Report	Task 1 (Analisis Grafik)	24,19	27,36
Social Policy Opinion	Task 2 (Opini Kritis)	26,21	28,97
Tech & Society Opinion	Task 2 (Opini Kritis)	27,17	28,84



Gambar 5. *Heatmap* Bias Keberpihakan Topik

Berdasarkan data pada Tabel 6 dan pemetaan visual Peta Panas di Gambar 5, di mana gradasi warna ungu gelap merepresentasikan intensitas skor terendah dan warna kuning terang menunjukkan probabilitas skor tertinggi, teridentifikasi tren yang konsisten pada kedua model. Kategori *Data Report* selalu memperoleh rerata penilaian terendah, sementara kategori yang berfokus pada penalaran opini berbasis teks mencatatkan probabilitas skor yang lebih tinggi. Fenomena depresiasi skor pada instrumen analisis visual ini berkaitan dengan keterbatasan infrastruktur antarmuka unimodal teks murni yang digunakan. Walaupun data visual secara komputasional dapat ditranslasikan ke dalam matriks teks untuk diinjeksi ke dalam *prompt*, penelitian ini secara sengaja tidak mendistribusikan konteks tersebut sebagai bentuk skenario uji stres (*stress test*). Oleh karena itu, penurunan skor pada kriteria *Task Achievement* di kategori *Data Report* tidak secara mutlak mengindikasikan keberpihakan topik inheren, melainkan merupakan konsekuensi

langsung dari defisit konteks akibat ketiadaan input data visual. Hasil evaluasi membuktikan bahwa tanpa ketersediaan rujukan visual, model tidak merespons dengan halusinasi generatif, melainkan menerapkan bias sikap konservatif (*conservatism bias*). Model secara terstruktur memberikan penilaian yang jauh lebih ketat guna menghindari probabilitas lolosnya informasi yang tidak akurat (*false positive*).

Keunggulan stabilitas arsitektur Mixture of Experts pada GPT OSS 120B atas arsitektur Dense dalam pengujian ini sejalan dengan temuan literatur komputasional mutakhir (Du et al., 2024; Li et al., 2025). Guna mendukung transparansi penelitian dan memungkinkan reproduktibilitas eksperimen (*research reproducibility*), seluruh skrip Python yang memuat logika prapemrosesan data, eksekusi pemanggilan API pada model, agregasi metrik, serta dataset hasil penilaian dari arsitektur GPT OSS 120B dan Qwen3-32B telah didokumentasikan secara publik. Kumpulan data dan kode sumber tersebut dapat diakses melalui repositori GitHub pada tautan berikut: https://github.com/dlzcods/llm-awe-reliability-fairness

SIMPULAN DAN SARAN

Hasil evaluasi menunjukkan bahwa GPT OSS 120B dan Qwen3-32B memiliki kelebihan dan kekurangan yang berbeda dalam penilaian esai otomatis. Qwen3-32B mampu menghasilkan keluaran dengan format JSON yang konsisten tanpa kesalahan struktur, tetapi skor yang dihasilkan cenderung kurang stabil dan lebih longgar. Sebaliknya, GPT OSS 120B memberikan penilaian yang lebih stabil dan mampu membedakan kualitas esai dengan lebih baik, tetapi masih mengalami kegagalan format JSON sebesar 25,4%. Pengujian juga menunjukkan bahwa GPT OSS 120B cenderung memberikan penilaian yang lebih ketat dibandingkan Qwen3-32B. Namun, karena penelitian ini belum menggunakan penilai manusia sebagai acuan utama, perbedaan tingkat keketatan tersebut hanya dapat dibandingkan antar-model. Pada tugas Data Report yang tidak menyediakan input visual, model tidak menghasilkan informasi yang tidak sesuai, tetapi cenderung memberikan penilaian secara lebih hati-hati untuk mengurangi risiko kesalahan. Perbedaan karakteristik kedua model menunjukkan bahwa penilaian otomatis sepenuhnya masih memiliki risiko. Oleh karena itu, pendekatan Human in the Loop tetap diperlukan sebagai mekanisme pengawasan. Sistem dapat digunakan untuk mengidentifikasi dokumen dengan variasi skor yang tinggi agar dapat diperiksa kembali oleh penilai manusia. Dengan cara ini, efisiensi sistem tetap dapat diperoleh tanpa mengurangi keandalan dan keadilan proses penilaian.

Berdasarkan hasil dan keterbatasan penelitian, penelitian selanjutnya disarankan untuk melibatkan penilai manusia sebagai gold standard agar akurasi model dapat diukur secara lebih objektif. Penggunaan model multimodal juga dapat dipertimbangkan untuk menangani tugas yang membutuhkan informasi visual. Selain itu, jumlah data perlu diperbesar agar hasil penelitian lebih representatif. Dari sisi teknis, sistem dapat dikembangkan dengan mekanisme structured output untuk mengurangi kesalahan format JSON serta automatic retry untuk menangani kegagalan koneksi atau batas waktu server. Penelitian selanjutnya juga dapat melakukan analisis sensitivitas terhadap batas Coefficient of Variation dan rentang deviasi skor untuk menentukan aturan perutean yang lebih optimal. Sebagai tahap implementasi, dapat dikembangkan dashboard Human in the Loop agar proses pemeriksaan dan eskalasi penilaian dapat digunakan secara praktis di lingkungan akademik.

TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Amikom Purwokerto dan dosen pembimbing serta pihak terkait yang telah memberikan dukungan penuh terhadap pelaksanaan penelitian ini, khususnya dalam validasi instrumen pakar (*expert judgment*), penyediaan fasilitas, serta bantuan administratif dan teknis. Dukungan tersebut sangat berkontribusi terhadap kelancaran proses penelitian, pengolahan komputasi data, hingga penyusunan artikel jurnal ini sehingga dapat diselesaikan dengan baik.

DAFTAR PUSTAKA

Andersen, N., Mang, J., Goldhammer, F., & Zehner, F. (2025). Algorithmic fairness in automatic short answer scoring. *International Journal of Artificial Intelligence in Education,* 35, 3128–3165 (2025). https://doi.org/10.1007/s40593-025-00495-5

Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, *56*(7), Article 166. https://doi.org/10.1145/3616865

Che Mat, A., Zulkornain, L. H., & Abdul Rahman, N. A. (2024). Automated writing evaluation: Users' perception and expectations. *International Journal of Information and Education Technology*, *14*(2), 183–192. https://doi.org/10.18178/ijiet.2024.14.2.2039

Du, J., & Nordin, N. R. M. (2025). A systematic review of automated writing evaluation (AWE) systems on university students' English writing performance (2020 – 2024). *Forum for Linguistic Studies*, *7*(11), 615–632. https://doi.org/10.30564/fls.v7i11.11764

Du, X., Gunter, T., Kong, X., Lee, M., Wang, Z.,

Zhang, A., Du, N., & Pang, R. (2024). Revisiting MoE and dense speed-accuracy comparisons for LLM training. *arXiv*. https://doi.org/10.48550/arXiv.2405.15052

Laclau, C., Largeron, C., & Choudhary, M. (2022). A survey on fairness for machine learning on graphs. arXiv. https://arxiv.org/abs/2205.05396v2

Li, H., Lo, K. M., Wang, Z., Wang, Z., Zheng, W., Zhou, S., Zhang, X., & Jiang, D. (2025). Can mixture-of-experts surpass dense LLMs under strictly equal resources? ArXiv.org. https://arxiv.org/abs/2506.12119v1

Nuong Deri, M., Singh, A., Zaazie, P., & Anandene, D. (2024). Leveraging artificial intelligence in higher educational institutions: A comprehensive overview. *Revista De Educación Y Derecho*, (30). https://doi.org/10.1344/REYD2024.30.45777

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, *6*, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Palomino, A., Fischer, A., Buschhüter, D., Roller, R., Pinkwart, N., & Paassen, B. (2025). Mitigating bias in item retrieval for enhancing exam assembly in vocational education services. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, 183–193. https://doi.org/10.18653/v1/2025.naacl-industry.16

Patwardhan, N., Marrone, S., & Sansone, C. (2023). Transformers in the real world: A survey on NLP applications. *Information*, *14*(4), 242. https://doi.org/10.3390/info14040242

Pratama A. R. (2025). The accuracy-bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication. *PeerJ Computer science*, *11*, e2953. https://doi.org/10.7717/peerj-cs.2953

Selvam, M., & González Vallejo, R. (2025). human-in-the-loop models for ethical AI grading: Combining AI speed with human ethical oversight. *EthAIca*, *4*, 413. https://doi.org/10.56294/ai2025413

Shao, M., Li, D., Zhao, C., Wu, X., Lin, Y., & Tian, Q. (2024). Supervised algorithmic fairness in distribution shifts: A survey. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 8225–8233). https://doi.org/10.24963/ijcai.2024/909

Siddique, S., Haque, M. A., George, R., Gupta, K. D., Gupta, D., & Faruk, M. J. H. (2024). Survey on machine learning biases and mitigation techniques. *Digital*, *4*(1), 1-68. https://doi.org/10.3390/digital4010001

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. British Journal of Educational Technology, 56, 150–166. https://doi.org/10.1111/bjet.13494