



Evaluasi Kualitas, Akurasi, dan Stabilitas Penilaian Pemrograman Berbantuan Model Bahasa Besar

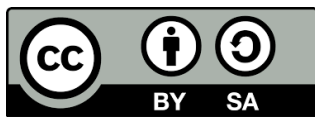
Naufal Raihan Saputra*, Ahmad Faqod Kurnia, Jenny Agustina Rahman

Institut Teknologi PLN, Indonesia

Email: naufal2411106@itpln.ac.id*

Keywords:

Automated Assessment; ICC; LLM; Programming Practicum; Systematic Bias.



This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

Copyright © 2026 by Author. Published by Publikasiku Publisher

Kata Kunci:

Bias Sistematis; ICC; LLM; Penilaian Otomatis; Praktikum Pemrograman

Abstract

Large language models (LLMs) offer opportunities to automate programming assessment, but their suitability requires evidence of scoring accuracy, stability, and feedback quality. This study evaluates these dimensions using data from four programming laboratory classes involving 197 students and 4,664 responses. A Design Science Research approach guided system development and evaluation. Accuracy was examined using 163 paired scores against historical assistant assessments, employing correlation, error metrics, bias analysis, Bland–Altman agreement, and boundary errors. Stability was investigated through repeated assessments and configuration comparisons, while feedback quality was evaluated using automated textual measures and three LLM judges. Results showed strong correlation with assistant scores ($r = 0.854$), alongside systematic negative bias of 8.97 points, MAE of 10.09, and RMSE of 13.39. False failure classifications affected 17.2–18.4% of cases at the examined thresholds. Repeated scoring showed high consistency ($ICC = 0.9943$), although assessor version changes shifted scores by 2.49 points and setting temperature to zero increased variability 4.7-fold. Feedback received higher factual correctness and language ratings than actionability, which averaged 2.633. These findings indicate that consistency alone does not establish assessment validity. LLMs can support ranking and screening, but final grading requires calibration, version control, and human oversight, particularly near consequential academic decision thresholds.

Abstrak

Model bahasa besar (LLM) menawarkan peluang otomatisasi penilaian pemrograman, tetapi kelayakannya memerlukan bukti akurasi nilai, stabilitas, dan kualitas umpan balik. Penelitian ini bertujuan mengevaluasi ketiga dimensi tersebut menggunakan data empat kelas praktikum pemrograman yang melibatkan 197 mahasiswa dan 4.664 jawaban. Pendekatan Design Science Research digunakan untuk pengembangan dan evaluasi. Akurasi diperiksa melalui 163 pasangan nilai dengan penilaian asisten historis sebagai pembanding menggunakan korelasi, metrik galat, analisis bias, kesepakatan Bland–Altman, dan kesalahan batas. Stabilitas diuji melalui penilaian berulang dan perbandingan konfigurasi, sedangkan kualitas umpan balik dinilai menggunakan ukuran tekstual otomatis dan tiga penilai LLM. Hasil menunjukkan korelasi kuat dengan nilai asisten ($r = 0,854$), tetapi disertai bias negatif sistematis 8,97 poin, MAE 10,09, dan RMSE 13,39. Kesalahan klasifikasi lulus menjadi

gagal mencapai 17,2–18,4% pada ambang yang diperiksa. Penilaian berulang menunjukkan konsistensi tinggi ($ICC = 0,9943$), tetapi perubahan versi penilai menggeser nilai 2,49 poin dan pengaturan *temperature* nol meningkatkan variasi 4,7 kali. Umpan balik memperoleh penilaian kebenaran faktual dan kebahasaan lebih tinggi daripada keterlaksanaan saran, yang memiliki rerata 2,633. Temuan ini menunjukkan bahwa konsistensi belum menjamin validitas penilaian. LLM dapat mendukung pemeringkatan dan penyaringan, tetapi penetapan nilai akhir memerlukan kalibrasi, pengendalian versi, serta pengawasan manusia, terutama pada keputusan akademik dekat ambang kelulusan.

PENDAHULUAN

Penilaian otomatis pada pendidikan pemrograman telah lama menjadi kebutuhan karena volume jawaban yang harus dikoreksi tumbuh lebih cepat daripada kapasitas pengajar, sementara koreksi manual dibatasi waktu sekaligus rawan tidak konsisten (Cisneros-González et al., 2025). Kemunculan *Large Language Model* (LLM) memperluas cakupan penilaian otomatis dari sekadar pencocokan output program menjadi penilaian atas penjelasan, penalaran, dan gaya kode, sekaligus menghasilkan feedback sebagai penjelasan (Yousef et al., 2025). Kemampuan ini mendorong adopsi LLM sebagai penilai pada praktikum pemrograman untuk mengurangi beban koreksi dan menegakkan rubrik secara konsisten (Jukiewicz, 2026a). Namun kemudahan itu menyembunyikan satu pertanyaan mendasar: apakah LLM memang layak dipercaya menetapkan nilai akademik, dan atas dasar bukti apa kelayakan itu dinyatakan.

Banyak penelitian penilaian berbantuan LLM masih menjadikan korelasi terhadap penilai manusia sebagai bukti utama kelayakan (Cisneros-González et al., 2025), padahal korelasi tinggi tidak cukup untuk mengambil keputusan nilai. Tiga keterbatasan mengikuti. Pertama, korelasi tidak menangkap bias sistematis maupun letak *error*. Dua penilai dapat berkorelasi hampir sempurna sambil satu di antaranya konsisten lebih ketat, sehingga peringkatnya benar tetapi nilai absolutnya salah, dan korelasi juga tidak menunjukkan apakah *error* menumpuk tepat di sekitar ambang kelulusan. Perbedaan tingkat kelonggaran antarmodel pada tugas yang sama telah terdokumentasi (Jukiewicz, 2026a), dan metrik kesepakatan yang lazim dipakai memvalidasi penilai LLM terbukti melebih-lebihkan kemampuan penilaiannya (Norman et al., 2026). Kedua, konsistensi model, yaitu kesamaan nilai ketika jawaban yang sama dinilai ulang pada konfigurasi sama, masih jarang dilaporkan berdampingan dengan akurasi meski merupakan syarat minimum agar dapat berkelanjutan. Konsistensi justru diidentifikasi sebagai persoalan inti yang belum terselesaikan pada pemikiran *LLM-as-a-judge* (Gu et al., 2026a), dan pengujian berulang memperlihatkan bahwa konsistensi penilai LLM bervariasi antar sampel serta tidak sepenuhnya hilang dengan menurunkan *temperature* (Yagubyan, 2026). Ketiga, kualitas *feedback* jarang diaudit. penilaian sejenis kerap berhenti pada angka yang dihasilkan LLM, bukan pada kebenaran teks penjelasan yang justru dibaca mahasiswa (Raj et al., 2025a), padahal *feedback* yang keliru lebih berbahaya daripada tiadanya *feedback*.

Kebaruan penelitian ini bukan pada satu pengukuran yang belum pernah ada, melainkan pada pengujian ketiga syarat kelayakan tersebut secara serentak, akurat, stabil, dan berguna, pada data operasional satu mata kuliah praktikum yang nyata dengan penilaian asisten sebagai pembanding. Pembanding manusia diperlakukan sebagai rujukan operasional, bukan baku,

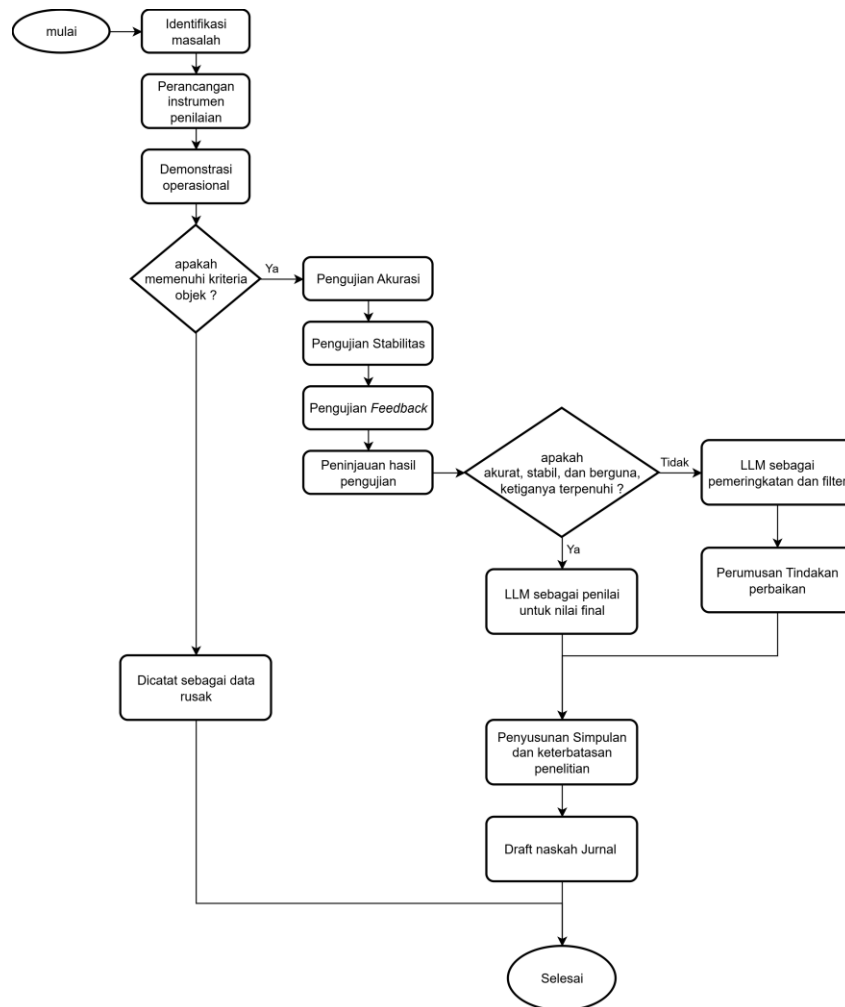
mengingat penilai manusia sendiri terbukti kerap tidak konsisten terhadap dirinya sendiri saat menilai tugas pemrograman (Messer et al., 2025). Ketiga syarat itu diurai menjadi tiga objektif. Objektif akurat mengukur seberapa dekat nilai LLM dengan asisten pada jawaban identik dan di titik mana error terkonsentrasi, melalui metrik kesepakatan, *bias analysis*, serta *boundary error* yang memetakan *error* ke konsekuensi pada kategori rubrik dan ambang kelulusan. Objektif stabil menguji apakah LLM memberi nilai yang sama pada jawaban dan *prompt* yang sama, melalui metrik simpangan dan kesepakatan antar pengulangan. Objektif berguna menilai apakah *feedback* benar secara faktual dan dapat ditindaklanjuti, melalui metrik korpus tanpa penilai dan rubrik multi variabel. Jika salah satu syarat gagal, LLM belum layak menentukan nilai final. Penelitian ini mengukur kelayakan LLM sebagai penilai, bukan mengklaim LLM lebih baik daripada manusia. Selain ketiga tekanan volume dan konsistensi di atas, adopsi LLM juga sering dimotivasi harapan menekan bias relasi antara penilai dan praktikan. Motif ini tetap relevan, tetapi diperlakukan sebagai risiko yang belum terverifikasi, bukan temuan, karena sistem penilaian yang diamati tidak mencatat relasi penilai-praktikan sehingga besarnya bias tersebut tidak terukur pada penelitian ini.

Kontribusi penelitian ini ada tiga. Pertama, kerangka evaluasi tiga syarat kelayakan penilai LLM (akurat, stabil, berguna) yang diterapkan serentak pada data operasional nyata. Kedua, pemisahan dua sumber ketidakstabilan penilai LLM, yaitu *noise* antar panggilan yang kecil dan pergeseran antar versi model yang jauh lebih besar, yang jarang dibedakan pada literatur (Yagubyan, 2026). Ketiga, pendefinisian *boundary error* tiga tingkat yang menghubungkan besaran statistik dengan konsekuensi nyata pada kategori rubrik, ambang kelulusan, dan batas skala.

Dengan pendekatan Design Science Research, penelitian ini bertujuan mengevaluasi kualitas, akurasi, dan stabilitas penilaian pemrograman berbantuan LLM menggunakan data operasional serta pembandingan manual historis. Secara khusus, penelitian mengukur kedekatan nilai LLM dengan nilai asisten pada jawaban yang sama, mengidentifikasi arah dan konsekuensi galat, menguji konsistensi penilaian ulang, serta menilai kebenaran dan keterlaksanaan saran dalam umpan balik. Kontribusi metodologisnya berupa kerangka evaluasi terpadu yang memungkinkan kelayakan penilai diperiksa melalui beberapa dimensi yang saling melengkapi. Secara praktis, hasil penelitian diharapkan membantu pengelola praktikum menentukan kebutuhan kalibrasi, pengendalian versi model, penyempurnaan rubrik, dan peninjauan manusia pada jawaban berisiko. Bagi mahasiswa, manfaat yang diharapkan adalah penilaian yang lebih dapat dipertanggungjawabkan dan umpan balik yang lebih terarah, tanpa mengganggu peningkatan hasil belajar telah terbukti melalui evaluasi sistem ini.

METODE PENELITIAN

Penelitian memakai Design Science Research (DSR), pendekatan yang dipilih karena luaran risetnya berupa rancangan sistem nyata yang dievaluasi dengan data operasional, bukan data simulasi (Vom Brocke et al., 2020). Proses penelitian berjalan dari identifikasi masalah dan celah keilmuan, perancangan rubrik beserta pipeline penilaian LLM, demonstrasi pada empat kelas praktikum selama satu semester, evaluasi tiga objektif secara statistik, hingga komunikasi hasil. Gambar 1 menyajikan alur penelitian tersebut secara keseluruhan. Penekanan naskah ini berada pada tahap evaluasi.



Gambar 1. Diagram alir metode penelitian

Pola berpasangan ini tidak selalu melibatkan manusia. Pada pengujian stabilitas, pembandingnya adalah LLM terhadap hasilnya sendiri pada pengulangan yang berbeda.

Instrumen Penelitian

Instrumen penilaian dalam penelitian ini berperan sebagai alat pengambilan data, bukan sebagai kontribusi utama. Jawaban mahasiswa dikirim ke model tanpa disertai identitas apa pun; prompt hanya memuat soal, kunci jawaban, rubrik, dan teks jawaban. Sifat buta identitas ini ditegakkan secara struktural pada tingkat prompt sehingga tidak bergantung pada kepatuhan prosedural penilai. Keluaran model berupa JSON yang memuat kategori atau sub-kriteria, poin, dan umpan balik, lalu dilewatkan melalui lima lapis pengaman sebelum disimpan:

1. pembersihan pembungkus (wrapper) markdown
2. penguraian (parsing) JSON dengan kegagalan diperlakukan sebagai galat, bukan diisi nilai sembarang
3. pemotongan (clamping) tiap sub-kriteria ke rentang rubrik
4. penskalaan ke poin soal
5. pemotongan akhir ke rentang 0 sampai poin maksimum. Nilai disimpan dengan prefiks penanda AI dan otoritas akhir atas nilai tetap berada pada pengelola praktikum.

Rubrik jalur esai mengharuskan model memilih satu kategori, lalu poin diambil dari tabel rubrik. Rancangan ini membuat model tidak menetapkan angka secara bebas sehingga risiko halusinasi numerik ditekan, sejalan dengan temuan bahwa keandalan skor yang dihasilkan LLM secara langsung masih menjadi persoalan terbuka (Gu et al., 2026b). Pendekatan rubrik yang dapat disesuaikan pengajar juga telah dipakai pada sistem penilaian pemrograman berbantuan LLM sebelumnya (Raj et al., 2025b). Kategori pada rubrik ini bersifat terurut dengan jarak poin antar kategori yang tidak seragam. Pada *pre-test* tingkat *medium*, jarak kategori “benar_singkat” ke “salah_penjelasan” hanya 3 poin, sedangkan pada *post-test* tingkat *medium* jarak “benar_penjelasan” ke benar_singkat” mencapai 15 poin. Satu langkah kesalahan kategori karena itu bernilai 3 sampai 15 poin, dan ketidakseragaman inilah yang menjadi dasar kuantitatif definisi *boundary error* tingkat pertama pada Subbab B. Rubrik jalur *coding* memakai tiga sub-kriteria, yaitu kesesuaian dengan petunjuk, keberhasilan dijalankan tanpa *error*, dan ketepatan waktu pengumpulan, dengan total 45 atau 85 poin bergantung pada jenis *course*.

Penilaian untuk keperluan penelitian dijalankan pada *endpoint* LLM tersendiri yang kompatibel dengan OpenAI, terpisah dari layanan produksi, agar pemilihan model tidak terikat pada keputusan operasional sistem. Model yang digunakan adalah *gpt-oss-120b*, yang pada pertengahan periode penelitian berganti menjadi *mimo-v2.5-pro-hermes*, keduanya diakses melalui *openrouter* pada 2025. Pergantian ini dicatat secara eksplisit karena menjadi salah satu variabel yang dianalisis pada pengujian stabilitas, dan karena model yang berbeda diketahui memiliki gaya penilaian serta tingkat kelonggaran yang berbeda meskipun meAnilai tugas yang sama (Jukiewicz, 2026b). Prompt dan rubrik yang dipakai identik dengan yang berjalan pada sistem produksi. Konfigurasi pemanggilan ditetapkan sebagai format keluaran JSON pada tingkat antarmuka melalui parameter *response_format* : *json_object*, mode *streaming* dinonaktifkan, dan batas waktu HTTP 50 detik. Parameter *temperature* sengaja tidak ditetapkan pada kondisi pertama sehingga mengikuti nilai bawaan penyedia, dan justru parameter inilah yang menjadi variabel perlakuan utama pada pengujian stabilitas. Perlu dicatat bahwa nilai bawaan tersebut tidak didokumentasikan penyedia, sehingga kondisi pembanding pada pengujian *temperature* dinyatakan sebagai kondisi bawaan, bukan sebagai nilai numerik tertentu.

Matriks Evaluasi

Sebelum metrik diuraikan, seluruh simbol didefinisikan lebih dulu. Indeks i menandai satu pasangan nilai, dengan a_i sebagai nilai penilai LLM pada jawaban ke- i dan m_i sebagai nilai asisten (manual) pada jawaban yang sama. n adalah jumlah pasangan nilai yang dianalisis. Rata-rata selisih kedua penilai dilambangkan d dan simpangan bakunya s_d . Ambang kelulusan dilambangkan τ . Untuk pengujian stabilitas, indeks j menandai satu jawaban dan indeks t menandai satu pengulangan penilaian, dengan k sebagai jumlah pengulangan per jawaban dan $x_{j,t}$ sebagai nilai jawaban ke- j pada pengulangan ke- t . rata-rata nilai jawaban ke- j lintas pengulangan dilambangkan $\bar{x}_j$, simpangan bakunya σ_j , dan poin maksimum soal ke- j dilambangkan P_j . Lambang N adalah jumlah jawaban yang diuji, $df(t)$ adalah jumlah jawaban yang memuat token t (frekuensi dokumen), dan Persamaan (1) adalah fungsi indikator yang bernilai 1 bila syarat di dalamnya terpenuhi dan 0 bila tidak.

1. Pengujian Akurasi

Akurasi diukur melalui persamaan (1) sampai (7). Bias adalah kesalahan sistematis yang berarah, yaitu rata-rata selisih nilai LLM terhadap nilai asisten, dan dihitung melalui persamaan (1).

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n (a_i - m_i) \quad (1)$$

Simpangan baku selisih, s_d , mengukur besar variasi error di sekitar bias dan dihitung melalui persamaan (2), dengan $d_i = a_i - m_i$

$$s_d = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2} \quad (2)$$

Mean Absolute Error (MAE) mengukur besar error rata-rata tanpa memperhatikan arahnya, sebagaimana persamaan (3).

$$MAE = \frac{1}{n} \sum_{i=1}^n |a_i - m_i| \quad (3)$$

Root Mean Square Error (RMSE) memberi bobot lebih besar pada error yang besar, sebagaimana persamaan (4).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (a_i - m_i)^2} \quad (4)$$

Ketiga besaran ini terhubung melalui identitas pada persamaan (5).

$$RMSE^2 = Bias^2 + s_d^2 \quad (5)$$

Identitas tersebut memungkinkan total error kuadrat diurai menjadi komponen sistematis, yang dapat dikoreksi lewat kalibrasi, dan komponen acak yang tidak dapat dikoreksi. Rasio RMSE terhadap MAE menunjukkan seberapa besar kontribusi error ekstrem terhadap total error, karena rasio ini meningkat ketika terdapat pencilan yang memperbesar RMSE di atas MAE. Kesepakatan linear antara kedua penilai diukur dengan koefisien korelasi Pearson r , sedangkan kesepakatan peringkat diukur dengan koefisien korelasi Spearman ρ . Signifikansi bias diuji dengan uji t berpasangan pada persamaan (6).

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \quad (6)$$

Korelasi dan uji signifikansi saja belum memadai untuk menyimpulkan kesepakatan dua metode pengukuran, karena keduanya tidak menunjukkan seberapa jauh dua nilai dapat berbeda pada satu individu (Gerke, 2020). Karena itu dihitung pula *Bland-Altman Limits of Agreement* (LoA) pada persamaan (7), yang memberi rentang selisih yang masih wajar untuk satu jawaban (Gerke & Möller, 2021).

$$LoA = \bar{d} \pm 1.96 \cdot s_d \quad (7)$$

2. Boundary Error

Boundary error memetakan error nilai ke konsekuensinya pada tiga tingkat batas, yang seluruhnya diukur dari 163 pasangan nilai yang sama dengan pengujian akurasi.

a. Batas kategori rubrik

Proporsi kesalahan yang hanya berjarak satu langkah kategori pada rubrik beserta biaya poin rata-rata yang ditimbulkannya.

b. Batas keputusan nilai

Proporsi mahasiswa yang berpindah sisi ambang kelulusan τ ketika nilai LLM menggantikan nilai asisten. Proporsi ini dipecah menjadi *false-fail*, yaitu mahasiswa yang lulus menurut penilaian manual tetapi gagal menurut penilaian LLM, dan *false-pass*, yaitu mahasiswa yang gagal menurut penilaian manual tetapi lulus menurut penilaian LLM. Pengujian dilakukan pada dua ambang, yaitu $\tau = 50$ dan $\tau = 60$.

c. Batas skala rubrik

Floor Effect, yaitu proporsi nilai yang menumpuk di angka 0, dan *ceiling function*, yaitu proporsi nilai yang menumpuk di poin maksimum. Ambang penanda sebesar 15 persen ditetapkan sebagai batas ketika skala rubrik dipandang tidak lagi mampu membedakan kualitas jawaban pada ujung sebarannya.

3. Pengujian Stabilitas

Pada pengujian stabilitas, satu jawaban ke- j dinilai sebanyak k kali dengan prompt yang identik, sehingga menghasilkan deret nilai $x_{j,1}$ sampai $x_{j,k}$. Simpangan baku nilai per butir dihitung melalui persamaan (8).

$$\sigma_j = \sqrt{\frac{1}{k-1} \sum_{t=1}^k (x_{j,t} - \bar{x}_j)^2} \quad (8)$$

Agar dapat dibandingkan antar soal yang berbeda poin maksimumnya, simpangan baku dinyatakan secara relatif terhadap poin maksimum soal sebagaimana persamaan (9).

$$\sigma_{\text{relatif}} = \frac{\sigma_j}{P_j} \times 100\% \quad (9)$$

Proporsi butir yang deterministik, yaitu butir yang menghasilkan nilai identik pada seluruh pengulangan, dihitung melalui persamaan (10).

$$P_0 = \frac{1}{N} \sum_{j=1}^N \mathbf{1}[\sigma_j = 0] \quad (10)$$

Kesepakatan mutlak antar pengulangan diringkas dengan Intraclass Correlation Coefficient dua arah acak, ICC(A,1) (Koo & Li, 2016), sedangkan kesepakatan pada tingkat kategori rubrik lintas k pengulangan diringkas dengan *Fleiss κ* (Zapf et al., 2016). Pemisahan antara κ kategori dan σ nilai penting dilakukan karena sebuah model dapat stabil pada angka nilainya tetapi berpindah kategori rubrik, atau sebaliknya stabil pada kategori tetapi bergeser pada angka.

4. Pengujian Feedback

Kualitas *feedback* dinilai dalam dua lapis. Lapis pertama dihitung otomatis oleh program, tanpa melibatkan penilai manusia maupun LLM. Lapis ini menilai dua sifat. Pertama, seberapa spesifik *feedback*, yaitu apakah *feedback* memakai kata yang khas untuk satu jawaban tertentu dan bukan kalimat umum yang cocok untuk jawaban mana pun. Sebuah kata dianggap khas bila muncul pada paling banyak 5 dari 100 jawaban ($df(t) \leq 5$). Kedua, seberapa mirip *feedback* satu sama lain, diukur dengan indeks *Jaccard*, yaitu ukuran kesamaan dua teks yang membagi jumlah kata yang dipakai bersama dengan jumlah seluruh kata yang muncul di keduanya. Nilainya 0 bila tidak ada kata yang sama dan 1 bila identik, dihitung atas 4.950 pasangan *feedback*. Ambang 5 tersebut dipilih langsung, tidak diturunkan dari standar baku.

Lapis kedua menilai *feedback* dengan rubrik lima kriteria berskala 1 sampai 5, yaitu kebenaran faktual, konsistensi skor, akurasi, dapat ditindaklanjuti, dan kualitas bahasa, dan dikerjakan oleh tiga penilai LLM dari model yang berbeda. Kriteria dapat ditindaklanjuti dipilih karena *feedback* baru benar-benar berguna apabila memberi arah perbaikan, bukan sekadar menyatakan letak kesalahan (Wisniewski et al., 2020). Karena penilai juga berupa LLM dan berisiko bias, ketiganya sengaja diambil dari model yang berbeda (Gu et al., 2026b).

Kesepakatan antar penilai diukur dengan dua angka, yaitu *Fleiss* κ (Zapf et al., 2016) untuk kesepakatan bertiga sekaligus dan *Cohen* κ_w untuk tiap pasangan penilai (Zapf et al., 2016). Nilai κ (kappa) adalah ukuran kesepakatan yang sudah dikoreksi terhadap peluang cocok secara kebetulan, nilai mendekati 0 berarti kesepakatan tidak lebih baik daripada menebak, sedangkan mendekati 1 berarti para penilai hampir selalu sepakat. Kedua angka ini dilaporkan bersama karena mengukur hal yang berbeda. *Fleiss* κ hanya menghitung sepakat bila ketiga penilai memberi skor yang sama persis, sedangkan κ_w berbobot masih memberi nilai sebagian bila skor mereka hanya berselisih satu tingkat. Besar kecilnya κ dibaca memakai rentang penggolongan yang umum dipakai (Zapf et al., 2016).

5. Protokol dan Sumber Data

Tabel 1. Protokol Pengujian Per Objektif

Pengujian	Protokol
Akurasi	163 pasangan nilai LLM dan manual (dirinci per jenis <i>course</i> dan tingkat kesulitan). <i>boundary error</i> dari data yang sama, matriks 351 butir berpasangan
Stabilitas (data arsip)	1.279 pasangan dari arsip <i>regrade</i> dua iterasi berkonfigurasi sama
Stabilitas (data uji terkendali)	200 jawaban (50 per jenis <i>course</i>). Tiap jawaban dinilai 4 kali pada 2 kondisi <i>temperature</i> (bawaan dan 0.0), total 1.592 panggilan tanpa gagal
Kualitas <i>Feedback</i>	100 keluaran JSON dengan struktur jelas. 3 penilai LLM lintas model (deepseek-v4-pro, glm5-2, kimi-k3) dengan protokol dua tahap. tahap 1 menilai tanpa melihat skor, tahap 2 baru membuka skor untuk menilai konsistensinya

Data berasal dari dua sumber. Basis operasional adalah PostgreSQL produksi (Supabase) berisi 4.664 jawaban dari 197 mahasiswa pada 4 kelas, 175 soal, dan 1.849

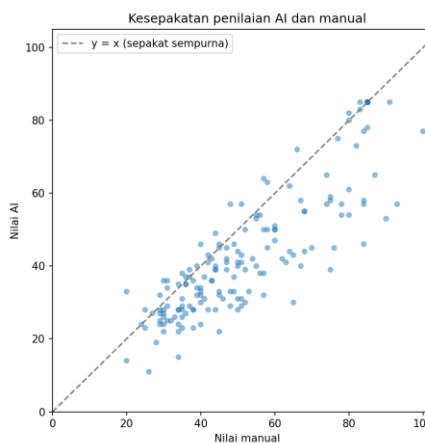
pengerjaan. Satu-satunya pembandingan manual diambil dari arsip Firebase sistem lama, yaitu 163 dari 1.735 jawaban yang berlabel penilaian asisten (sisanya 1.572 atau 90,6% dinilai LLM). Karena nilai manual ini berasal dari sistem dan periode berbeda (Maret-April 2026, berbanding Juni-Agustus 2026 untuk data LLM), perannya adalah historical control, bukan eksperimen paralel.

HASIL DAN PEMBAHASAN

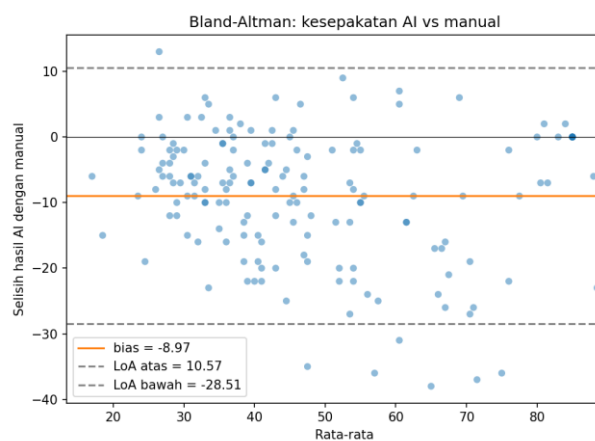
1. Korelasi dan Bias

Hasil analisis 163 pasangan menunjukkan korelasi kuat terutama untuk IK (Indek Kepercayaan). *Pearson* $r = 0,854$ (IK 95%: 0,806-0,891), *Spearman* $\rho = 0,851$. Artinya peringkat mahasiswa relatif terjaga. Namun bias rata-rata -8,97 poin (nilai dari LLM dikurangi nilai manual) sangat signifikan, yaitu $t(162) = -11,49, p < 0,001$, IK 95%, bias: -10,51 sampai -7,43. Interval tidak memuat nol, membuktikan bias bukan kebetulan *sampling*. LLM cenderung menilai lebih ketat daripada asisten.

Nilai RMSE adalah 13,39 sementara nilai MAE adalah 10,09, sehingga rasio antara RMSE dengan MAE adalah 1,33 yang menandakan adanya *error* ekstrem yang memperbesar RMSE di atas MAE. Penguraian RMSE menunjukkan bias kuadrat sebesar 80,46 dan varians selisih sebesar 98,79, sehingga sekitar 45 persen error kuadrat bersumber dari bias sistematis dan 55 persen dari variasi acak. Ini berarti sebagian besar error dapat dihapus lewat kalibrasi (pergeseran konstan atau regresi linear), sedangkan komponen acak tidak. *Bland-Altman* LoA selebar sekitar 39 poin (-28,51 sampai +10,57), terlalu lebar untuk keputusan kelulusan per individu.



Gambar 2. Scatter plot kesepakatan penilaian LLM dan manual



Gambar 3. Bland-Altman plot kesepakatan penilaian LLM dan manual

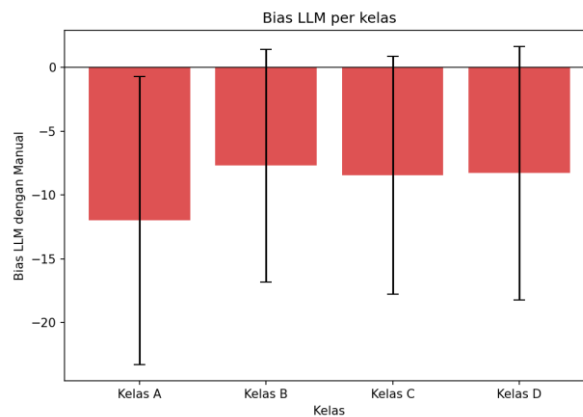
Error tidak tersebar merata. Rincian per jenis *course* :

Tabel 2. Bias dan Korelasi LLM dan Manual Per Jenis Penilaian

Course	N	Avg. manual	Avg. LLM	Bias	MAE	r
Pretest	61	52.43	40.36	-12.07	13.18	0.786
Posttest	76	43.08	36.66	-6.42	7.82	0.800
praktik	26	76.65	67.50	-9.15	9.46	0.849

Bias terburuk pada *pretest* (-12,07 poin, MAE 13,18). Hal ini masuk akal karena *pretest* diambil sebelum materi diajarkan, sehingga jawaban cenderung pendek dan setengah benar. Kondisi tersebut persis wilayah abu-abu tempat kategori *salah_penjelasan* dan *benar_singkat* paling rawan tertukar, dan pada *pre-test* medium jarak keduanya hanya 3 poin sedangkan pada *pre-test* hard 5 poin. Sebaliknya, praktik justru paling selaras karena rubriknya paling terstruktur.

Kesetaraan antar kelas terkonfirmasi. ANOVA satu arah menghasilkan $F(3,159)$ sebesar 1,480 ($p > 0,05$) dan *Kruskal-Wallis H* sebesar 5,634 dengan derajat bebas 3 ($p > 0,05$). Artinya besar *error* tidak bergantung pada kelompok mahasiswa. LLM memperlakukan keempat kelas dengan standar yang sama.



Gambar 1. Garis Simpangan Baku Bias LLM

2. Boundary Error

Batas kategori rubrik, yaitu kesepakatan penggolongan kategori pada matriks 351 butir, menunjukkan 64 butir (18,2%) berbeda kategori, dengan arah asimetris - 46 butir LLM lebih salah dibanding 18 butir LLM lebih benar. Penilaian benar/salah dasarnya justru kuat (sesuai 89,9% untuk kategori salah dan 89,2% untuk benar). Kategori penilaian yang lemah adalah apakah penjelasan mahasiswa cukup. Hanya 29,4% sesuai pada kategori *benar_penjelasan* (5 dari 17 butir), sisanya diturunkan ke *benar_singkat* (7 butir) atau *salah_penjelasan* (5 butir). Proporsi kesalahan ini naik seiring rubrik makin halus. butir *easy* 3 kategori berada pada 0,025-0,066. Butir *medium/hard* 5 kategori pada 0,171-0,329, tanpa tumpang tindih. Oleh karena itu proporsi tersebut sebaiknya dilaporkan per level, dengan 0,182 sebagai angka ringkas.

Batas keputusan nilai, yaitu migrasi mahasiswa terhadap ambang kelulusan pada 163 pasangan, diuji pada dua ambang. Pada ambang τ sebesar 50, proporsi migrasi 0,190 (*false-fail* 18,4%, *false-pass* 0,6%). Pada ambang τ sebesar 60, proporsi migrasi 0,184

(*false-fail* 17,2%, *false-pass* 1,2%). Kesalahan LLM hampir seluruhnya berarah *false-fail*. Dengan rata-rata kelas 48,78-53,05 dan bias -8,97, ambang 50-60 berada tepat di zona paling rawan.

Batas skala rubrik, yaitu *ceiling effect* dan *floor effect*, diukur dari 4.547 jawaban. *Ceiling effect* paling tajam pada praktik (24,1% jawaban nilai maksimum 85), *post-test* 8,5%, *pre-test* 9,4%. *Floor effect* minimal di semua jenis (di bawah 2%). Ujian_praktik/essay menunjukkan keanehan. Kondisi 19 jawaban bernilai identik 18 (simpangan baku nol), menandakan rubrik tidak cocok untuk jenis soal tersebut.

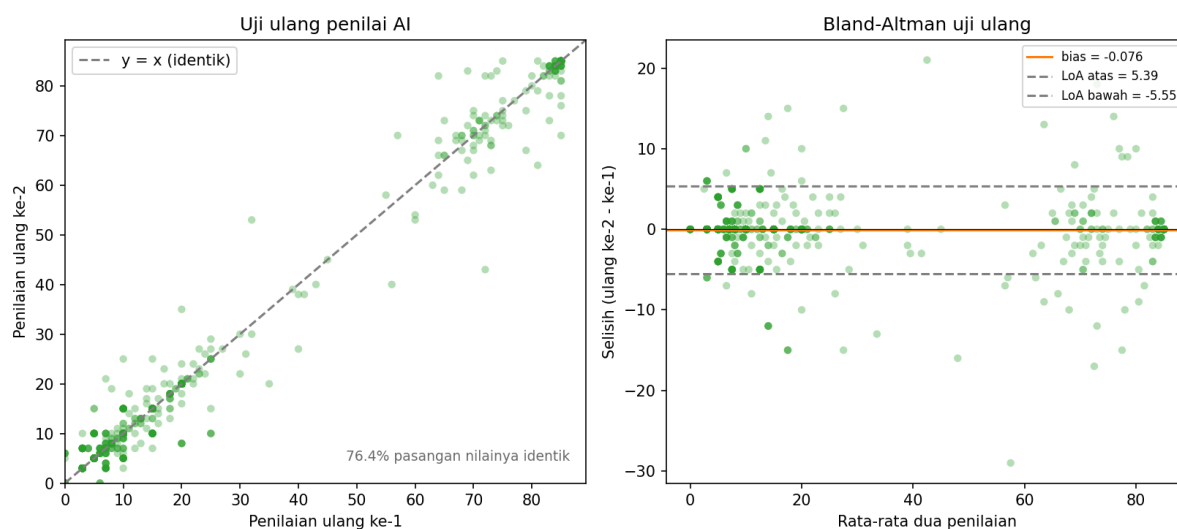
3. Stabilitas

Berlawanan dengan dugaan awal, konsistensi LLM terhadap dirinya sendiri pada konfigurasi yang sama ternyata sangat tinggi. Dua sumber bebas memberikan angka yang praktis identik:

Tabel 3. Konsistensi LLM dari dua sumber data

Sumber	n	r	ICC (1,1)	MAE	Identik
Penilaian ulang dari arsip	1279	0.9943	0.9943	1.01	76.4%
Eksperimen ulang	199	-	0.996	-	-

Selisih rata-rata pasangan arsip hanya -0,08 poin (t sebesar -0,97), tidak berbeda nyata dari nol, memenuhi syarat penilaian ulang dengan konfigurasi sama. Batas kesepakatan -5,55 sampai +5,39 poin. Pada uji terkendali *temperature* default, simpangan baku antar-panggilan sebesar 0,484 poin (1,001% dari poin maksimum), proporsi butir deterministik sebesar 0,809, MMR sebesar 0,933, *Fleiss κ* kategori sebesar 0,973, dan *ParseRate* sebesar 1,000.



Gambar 2. Konsistensi LLM pada penilaian ulang: (kiri) sebaran nilai penilaian ke-1 dan ke-2; (kanan) plot Bland-Altman selisih antar penilaian

Dugaan awal bahwa konsistensi LLM lebih rendah daripada kesepakatannya dengan manusia ditolak dengan arah sebaliknya. 0,9943 jauh di atas 0,854. Uji beda dua korelasi bebas (*Fisher z*, yaitu uji statistik untuk membandingkan dua nilai korelasi) menghasilkan *z* sebesar 19,77 ($p \ll 0,001$), meski dilaporkan sebagai indikasi arah

karena kedua korelasi berasal dari kumpulan jawaban berbeda.

Risiko sesungguhnya adalah perubahan versi penilai, bukan variasi antar-panggilan. Pergantian versi penilai (termasuk pergantian model *gpt-oss-120b* ke *mimo-v2.5-pro-hermes* pada periode yang sama) menggeser nilai -2,49 poin secara tetap ke satu arah pada 1.279 jawaban (t sebesar -16,44), yaitu dua sampai tiga kali lipat besarnya variasi antar-panggilan. Pada arsip yang berbeda versi (*iterasi_1* dan *iterasi_3*), r sebesar 0,9792, MAE 2,82, dan identik 61,1%.

Lebih mengejutkan, menyetel *temperature* nol bukan solusi melainkan justru memperburuk. Hasil uji konsistensi terkendali (199 jawaban, 4 pengulangan, 2 kondisi).

Tabel 4. Perbandingan stabilitas LLM pada *temperature* bawaan dan *temperature* 0,0

Metrik	Temp. Default	Temp 0
Rata-rata σ (poin)	0,484	2,283
σ relatif	1,001	4,726
$P_o(\sigma = 0)$	0,809	0,422
ICC (A,1)	0,996	0,981
Fleiss κ kategori	0,973	0,847
ParseRate JSON	1,00	1,00
Panggilan Gagal	0/796	0/796

Menyetel *temperature* 0.0 memperburuk σ 4,7 kali lipat, memangkas kategori yang nilainya tetap dari 80,9% ke 42,2%, dan menggeser nilai +3,73 poin (t sebesar +7,66, 109 naik, 22 turun, 68 sama). Dari 161 jawaban stabil di *default*, hanya 79 (49,1%) tetap stabil di *temperature* 0. *Temperature* 0 merusak 82 jawaban dan menstabilkan hanya 5, sehingga rugi bersih. ICC (ukuran konsistensi dua penilaian atas objek yang sama) menyembunyikan kemunduran ini (hanya turun dari 0,996 ke 0,981) karena sebaran nilai antar jawaban lebar (SD sekitar 25 poin) sehingga *noise* antar-panggilan hanya bergerak dari 1,99% ke 8,96%. Oleh karena itu, σ , $\sigma_{relatif}$, dan P_o wajib dilaporkan berdampingan dengan ICC. Pemisahan dua sumber ketidakstabilan (*noise* panggilan yang kecil dan pergeseran versi yang besar) itulah penyumbang utama dari sisi metode pada penelitian ini.

4. Kualitas Feedback

Evaluasi 100 sampel *feedback* oleh 3 penilai LLM lintas model (1.500 penilaian aspek)

Tabel 5. Penilaian kualitas *feedback* LLM oleh tiga LLM Berbeda

Aspek	Rata-rata	Simpangan	Skor 1-2 (%)	Fleiss κ	choen k_w
Kebenaran faktual	3,933	1,352	21,0	0,157	0,411
Konsistensi skor	4,230	0,916	5,7	0,246	0,324
Kedetailan	3,053	1,457	39,0	0,213	0,614
Dapat ditindaklanjuti	2,633	1,385	52,7	0,240	0,680
Kualitas bahasa	4,403	0,727	0,7	0,190	0,330

Temuan pentingnya pada aspek dapat ditindaklanjuti, jauh tertingga. rata-rata 2,633 dengan 52,7% penilaian di skor rendah (1-2). *Feedback* menjelaskan apa yang salah, bukan apa yang perlu diperbaiki, tepat pada aspek yang menentukan apakah

feedback berfungsi (Wisniewski et al., 2020). Yang perlu diperbaiki rubrik dan *prompt*, bukan model.

Kesepakatan harus dibaca hati-hati, *Fleiss κ* rendah (0,157-0,246, tergolong "sangat lemah" (Han et al., 2025)) sedangkan Cohen κ_w pada kriteria sama jauh lebih tinggi (0,680 pada kriteria dapat ditindaklanjuti). Perbedaan ini konsekuensi definisi *Fleiss* hanya setuju bila ketiga penilai pilih kategori persis sama, κ_w nilai selisih 1 tingkat. Artinya penilai sepakat besaran mutu tetapi beda letak batas. kategori 5 titik terlalu halus untuk disepakati harfiah, bukan penilaiannya acak.

Penilaian LLM ternyata sejalan dengan ukuran lapis 1 (ukuran otomatis tanpa penilai) pada keempat arah yang diuji. Keberadaan hasil langsung berkaitan dengan aspek kedetailan (r sebesar 0,533), kedetailan terukur berkaitan dengan aspek kedetailan (r sebesar 0,396), kesamaan istilah dengan jawaban berkaitan dengan aspek kebenaran faktual (r sebesar 0,390), dan jumlah kalimat berkaitan dengan aspek dapat ditindaklanjuti (r sebesar 0,439). Karena ukuran lapis 1 dihitung program tanpa penilai, kesesuaian arah ini menunjukkan penilai LLM benar-benar melacak sifat teks yang nyata dan terukur, bukan sekadar menebak. Kecenderungan menilai *feedback* yang panjang lebih tinggi tergolong sedang, yakni korelasi antara panjang *feedback* dan rata-rata lima aspek sebesar 0,477.

Pengaruh model penilai yang juga diuji, dengan *feedback* dari model *deepseek* dinilai oleh satu sekeluarga dan dua penilai dari luar. Pada rata-rata lima aspek, selisihnya kecil, yaitu +0,104 (t sebesar 2,02, $p < 0,05$). Namun selisih ini terpusat pada bagian kebenaran faktual sebesar +0,505 (t sebesar 4,27, $p < 0,001$) dan justru berbalik arah pada aspek kualitas bahasa sebesar -0,245 (t sebesar -3,87, $p < 0,001$). Pola ini bisa dijelaskan sebab-akibatnya. Menilai kebenaran faktual menuntut penilai menyetujui klaim yang dibuat *feedback* itu sendiri, sedangkan menilai kualitas bahasa hanya menyentuh permukaan teks. Artinya untuk praktik, penilaian oleh model yang sejenis sebaiknya tidak dipakai untuk aspek kebenaran faktual, sementara untuk aspek kebahasaan relatif aman.

KESIMPULAN

Penelitian ini mengevaluasi kelayakan model bahasa besar (LLM) sebagai penilai praktikum pemrograman melalui tiga syarat alat ukur, yaitu akurat, stabil, dan berguna, pada data operasional satu semester penuh. Ketiga syarat tidak terpenuhi secara seragam, dan pola itulah yang menjadi simpulan utamanya. Pada syarat akurat, penilaian LLM berkorelasi kuat dengan asisten ($r = 0,854$) sehingga peringkat mahasiswa relatif terjaga, tetapi mengandung bias sistematis -8,97 poin yang sangat signifikan ($t(162) = -11,49$; $p < 0,001$) dan batas kesepakatan selebar 39 poin, dan pemetaan boundary error menunjukkan bahwa 17,2% sampai 18,4% mahasiswa berisiko berpindah dari lulus menjadi gagal pada ambang kelulusan 50 sampai 60. Karena sekitar 45% galat kuadrat bersumber dari bias sistematis, sebagian besar galat masih dapat dihapus melalui kalibrasi, sedangkan komponen acak tidak. Pada syarat stabil, dugaan awal ditolak: konsistensi LLM terhadap dirinya sendiri sangat tinggi ($r = 0,9943$; ICC = 0,9943; 76,4% nilai identik), sehingga risiko utamanya adalah pergeseran versi penilai yang menggeser nilai -2,49 poin secara sistematis, sedangkan derau antar panggilan relatif kecil, dan

menyetel temperature nol bukan solusi karena memperburuk simpangan baku 4,7 kali lipat sekaligus menggeser nilai +3,73 poin. Pada syarat berguna, umpan balik benar secara faktual (3,933) dan bersih secara bahasa (4,403), tetapi lemah pada dimensi dapat ditindaklanjuti (2,633 dengan 52,7% penilaian di skor rendah), sehingga perbaikan yang paling mendesak berada pada rubrik dan prompt, bukan pada penggantian model. Simpulan keseluruhannya adalah bahwa LLM layak dipakai sebagai alat pemeringkatan dan penyaring, tetapi belum layak menetapkan nilai final tanpa kalibrasi dan penguncian versi penilai. Keterbatasannya adalah pembandingan manual yang hanya 163 pasangan dari sistem dan periode berbeda serta penilaian umpan balik oleh model, bukan manusia; penelitian lanjutan disarankan untuk menguji kalibrasi nilai, menajamkan definisi kategori rubrik yang bertetangga, menerapkan penilaian ganda selektif pada jawaban yang berisiko berpindah status kelulusan, dan mengevaluasi kebergunaan umpan balik secara langsung bersama mahasiswa (AlGhamdi et al., 2026).

PENDANAAN

Penulis menyatakan bahwa tidak ada dukungan finansial yang diterima untuk pelaksanaan penelitian ini dan/atau penerbitan artikel ini.

KONFLIK KEPENTINGAN

Penulis menyatakan bahwa penelitian ini dilakukan tanpa adanya hubungan komersial atau finansial yang dapat ditafsirkan sebagai potensi konflik kepentingan.

PERNYATAAN PENGGUNAAN KECERDASAN ARTIFISIAL GENERATIF

Penulis menyatakan bahwa kecerdasan artifisial (AI) generatif digunakan dalam penelitian dan penyusunan manuskrip ini. Pada tahap penelitian, model bahasa besar berperan sebagai instrumen penilaian yang dievaluasi, yaitu gpt-oss-120b dan mimo-v2.5-pro-hermes sebagai penilai utama serta deepseek-v4-pro, glm5-2, dan kimi-k3 sebagai penilai umpan balik lintas model. Pada tahap penyusunan manuskrip, AI generatif digunakan secara terbatas untuk membantu meningkatkan kejelasan bahasa dan merapikan struktur tata bahasa. Seluruh analisis data, interpretasi hasil, dan simpulan dirumuskan dan diverifikasi oleh penulis berdasarkan data operasional dan keluaran statistik yang sebenarnya. Setiap penulis memberikan kontribusi intelektual yang substansial terhadap karya ini. Seluruh isi manuskrip telah ditelaah secara cermat untuk memastikan keakuratannya, dan penulis bertanggung jawab sepenuhnya atas integritas serta isi akhir manuskrip.

REFERENSI

- AlGhamdi, E. M., Li, Y., Gašević, D., & Chen, G. (2026). Leveraging prompt-based LLMs for automated scoring and feedback generation in higher education. *Computers & Education*, 243, 105511. <https://doi.org/10.1016/j.compedu.2025.105511>
- Cisneros-González, J., Gordo-Herrera, N., Barcia-Santos, I., & Sánchez-Soriano, J. (2025). JorGPT: Instructor-aided grading of programming assignments with large language models (LLMs). *Future Internet*, 17(6), 265.
- Evaluating the capability of large language models in characterising relational feedback: A comparative analysis of prompting strategies. (2025). *Computers and Education*:

- Artificial Intelligence*, 8, 100427. <https://doi.org/10.1016/j.caeai.2025.100427>
- Gerke, O. (2020). Reporting standards for a Bland--Altman agreement analysis: A review of methodological reviews. *Diagnostics*, 10(5), 334.
- Gerke, O., & Möller, S. (2021). Bland--Altman limits of agreement from a Bayesian and frequentist perspective. *Stats*, 4(4), 1080–1090.
- Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., & others. (2026a). A survey on llm-as-a-judge. *The Innovation*, 7(6).
- Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., & others. (2026b). A survey on llm-as-a-judge. *The Innovation*, 7(6).
- Han, S. S., Takashima, Y., Shen, S. Z., Liu, C., Liu, Y., Thuo, R. K., Knowlton, S., Piskac, R., Shapiro, S. J., & Cohan, A. (2025). COURTREASONER: Can LLM Agents Reason Like Judges? *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 35279–35294.
- Jukiewicz, M. (2026a). A systematic comparison of large language models for automated assignment assessment in programming education: Exploring the importance of architecture and vendor. *Computers and Education Open*, 10, 100364.
- Jukiewicz, M. (2026b). A systematic comparison of large language models for automated assignment assessment in programming education: Exploring the importance of architecture and vendor. *Computers and Education Open*, 10, 100364.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163.
- Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2025). How consistent are humans when grading programming assignments? *ACM Transactions on Computing Education*, 25(4), 1–37.
- Norman, J. D., Rivera, M. U., & Hughes, D. A. (2026). Reliability without Validity: A Systematic, Large-Scale Evaluation of LLM-as-a-Judge Models Across Agreement, Consistency, and Bias. *ArXiv Preprint ArXiv:2606.19544*.
- Raj, J. N., Muppa, A., Nirmal, R., Kamath, T. W., Boicu, M., Dipukumar, A., & Laddha, A. (2025a). BOARD# 97: WIP: The Effectiveness of Rubric-Based LLM Feedback for Programming Assessments. *2025 ASEE Annual Conference \& Exposition*.
- Raj, J. N., Muppa, A., Nirmal, R., Kamath, T. W., Boicu, M., Dipukumar, A., & Laddha, A. (2025b). BOARD# 97: WIP: The Effectiveness of Rubric-Based LLM Feedback for Programming Assessments. *2025 ASEE Annual Conference \& Exposition*.
- Vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to design science research. In *Design science research. Cases* (pp. 1–13). Springer.
- Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087.
- Yagubyan, A. (2026). The Coin Flip Judge? Reliability and Bias in LLM-as-a-Judge Evaluation. *ArXiv Preprint ArXiv:2606.13685*.
- Yousef, M., Mohamed, K., Medhat, W., Mohamed, E. H., Khoriba, G., & Arafa, T. (2025). BeGrading: large language models for enhanced feedback in programming education. *Neural Computing and Applications*, 37(2), 1027–1040.
- Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data--which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, 16(1), 93.