DOI: https://doi.org/10.26618/3zghgz16 Published by University of Muhammadiyah Makassar

Equilibrium: Jurnal Pendidikan
Vol. XIV. Issu 2. Mei-September 2026

Automated Tri Dharma Quality Assessment for Academic Accreditation Using Qwen, Mistral, and DeepSeek

[1]Mochammad Ariel Sulton, [2]Tita Karlita, [3]Nyoman Bayu Surapati, [4]Firnanda Pristiana Nurmaida, [5]Faros Alaudin Althaf , [6]Yesta Medya Mahardhika, [7]Paramita Eka Wahyu Lestari, [8]Aris Bahari Rizki

[1] Applied Data Science, Electronic Engineering Polytechnic Institute of Surabaya, Indonesia
E-mail: arielsulton@ds.student.pens.ac.id
[2] Applied Data Science, Electronic Engineering Polytechnic Institute of Surabaya, Indonesia
E-mail: tita@pens.ac.id
[3] Game Technology, Electronic Engineering Polytechnic Institute of Surabaya, Indonesia
E-mail: nyomanbayu@gt.student.pens.ac.id
[4] Game Technology, Electronic Engineering Polytechnic Institute of Surabaya, Indonesia
E-mail: firnandapn@pens.ac.id
[5] Power Plant Engineering, Electronic Engineering Polytechnic Institute of Surabaya, Indonesia
E-mail: farosalaudinalthaf@pg.student.pens.ac.id
[6] Informatics Engineering, Electronic Engineering Polytechnic Institute of Surabaya, Indonesia
E-mail: yesta@pens.ac.id
[7] Telecommunication Engineering, Electronic Engineering Polytechnic Institute of Surabaya, Indonesia
E-mail: mita@pens.ac.id
[8] Telecommunication Engineering, Electronic Engineering Polytechnic Institute of Surabaya, Indonesia
E-mail: arisbr@pens.ac.id

Article History; Submitted: *12-05-2026;* ***Accepted:*** *21-05-2026;* ***Published:*** *22-05-2026*

Abstract. *Internal Quality Audit (AMI) is critical for Indonesian higher-education accreditation, yet study programs still score 60+ indicators manually by cross-referencing a 200-page Self-Evaluation Report (LED) and a 50-sheet Study Program Performance Report (LKPS), a process that takes 3–5 days per program and varies between assessors. This study develops an automated quality-assessment agent that ingests both documents and benchmarks three open-weight Large Language Models, Qwen 3.5 35B, DeepSeek-R1 32B, and Mistral-Small 3.2 24B, on 19 LAM-Teknik indicators sampled across all nine accreditation criteria (9 qualitative, 6 quantitative, 4 composite). The reference scores are taken from the certified human assessor's official record produced during the program's 2024 post-audit cycle. The results show that Qwen 3.5 attains the lowest MAE (0.605) and RMSE (0.769) with 100% within ±1 accuracy and 32.5 s per indicator; DeepSeek-R1 is the most cautious but the slowest (MAE 1.395; 79.9 s) and collapses on quantitative items (MAE 2.17); Mistral is the fastest (13.6 s).*

Keywords: *Tri Dharma; Academic Accreditation; Automated Quality Assessment; Large Language Models (LLMs).*

Abstrak. *Audit Mutu Internal (AMI) sangat penting untuk akreditasi perguruan tinggi di Indonesia, namun program studi masih menilai lebih dari 60 indikator secara manual dengan membandingkan laporan Evaluasi Diri (LED) sepanjang 200 halaman dan Laporan Kinerja Program Studi (LKPS) sebanyak 50 lembar, sebuah proses yang memakan waktu 3–5 hari per program dan berbeda antara penilai. Penelitian ini mengembangkan agen penilaian mutu otomatis yang memproses kedua dokumen tersebut dan membandingkan tiga Large Language Model berbobot terbuka, Qwen 3.5 35B, DeepSeek-R1 32B, dan Mistral-Small 3.2 24B, pada 19 indikator LAM-Teknik yang diambil sampelnya dari seluruh sembilan kriteria akreditasi (9 kualitatif, 6 kuantitatif, 4 komposit).*

Skor referensi diambil dari catatan resmi penilai manusia bersertifikat yang dibuat selama siklus pasca-audit program tahun 2024. Hasil menunjukkan bahwa Qwen 3.5 memperoleh MAE terendah (0,605) dan RMSE (0,769) dengan 100% akurasi dalam ±1 dan 32,5 detik per indikator; DeepSeek-R1 adalah yang paling berhati-hati tetapi paling lambat (MAE 1,395; 79,9 detik) dan gagal pada item kuantitatif (MAE 2,17); Mistral adalah yang tercepat (13,6 detik).

Kata kunci: *Tri Dharma; Akreditasi Akademik; Penilaian Kualitas Otomatis; Model Bahasa Besar (LLMs).*

INTRODUCTION

Higher-education accreditation in Indonesia operates under increasingly stringent quality-standardization demands, in which the Internal Quality Assurance System (SPMI) through its operational arm, the Internal Quality Audit (AMI) must ensure that every study program meets the National Standards for Higher Education (SN-Dikti) and the criteria set by BAN-PT or LAM (Efendi, Hapsari, & Yusup, 2025; Sakuroh, Marlina, Krismayanti, & Adiningsih, 2026; Aljarallah & Dutta, 2022). The primary obstacle is the reliance on manual assessment that produces fragmented "data silos" and delayed reporting (Efendi et al., 2025), particularly for technical study programs that require rigorous and continuous quality monitoring (Subijanto et al., 2021; Sugiarti, 2022); earlier research therefore explored data-driven alternatives such as K-Means clustering and OLAP for accreditation analytics (Sakinah, Syaputra, Rahman, Fajri, & Rachmansyah, 2025), AI-assisted self-evaluation frameworks (Ticolau, 2026), and data-lakehouse architectures for academic metrics (Isnaeni, Putranto, Andriyani, & Khomsah, 2025).

Recent studies converge on the conclusion that conventional internal-audit processes cannot capture the dynamics of large and fast-paced academic data (Isnaeni et al., 2025); a comprehensive automated architecture that directly processes academic metrics is needed (Ticolau, 2026; Başaran, 2026), and the integration of automated data-analysis technology is expected to minimize human error and improve institutional transparency (Sakuroh et al., 2026). Large Language Models (LLMs) open new opportunities for automated processing of complex academic data (Naveed, Khan, & Qui, 2025; Demir & Yavuz, 2026), but model behavior diverges sharply across families: Qwen 3.5 is an Alibaba dense decoder with strong multilingual instruction-following and native tool-use through bind_tools (Bai et al., 2023); it demonstrates high precision in error minimization (Aydin, Karaaslan, Erenay, & Bacanin, 2026). DeepSeek-R1 is a reasoning-distilled model that emits an explicit *"<think>...</think>"* chain before answering and therefore cannot use native function calling on Ollama, requiring prompt-based JSON dispatch instead (DeepSeek-AI, 2025; Jegham, Abdelatti, & Hendawi, 2025). Mistral-Small 3.2 is a 24 B mixture-of-experts variant emphasizing throughput; it offers the lowest latency among the three but is prone to hallucination on Indonesian-language tasks (Mistral AI, 2024; Nurohim, Setyadi, & Fauzi, 2025).

This study aims to investigate whether a cross-modal Retrieval-Augmented Generation (RAG) agent, integrating LED PDF and LKPS Excel data, can effectively reduce the Akreditasi Mandiri Indonesia (AMI) cycle time for LAM-Teknik study programs while strictly preserving an auditable evidence trace. To achieve this, it evaluates which large language model among Qwen 3.5, DeepSeek-R1, and Mistral-Small offers the optimal trade-off between accuracy, calibration, and runtime when tested on a multi-criteria indicator sample. Furthermore, the research explores how scoring performance varies across different indicator types—specifically qualitative, quantitative, and composite indicators—and analyzes what these variations imply for the overarching design of prompts and tools within the system.

This research aims to (1) design a cross-modal RAG agent that synchronously cross-references qualitative LED PDF evidence with quantitative LKPS Excel data through four specialized tools; (2) benchmark three open-weight LLMs against certified human-assessor reference scores on 19 indicators spanning all nine LAM-Teknik criteria; and (3) characterize per-type error patterns, calibration, and runtime to derive deployment guidance for Indonesian quality-assurance offices. The novelty of this study lies in four complementary contributions not addressed by the prior LLM-

assessment literature, which has largely focused on English-language student-work grading (Sun et al., 2025) and single-modality RAG-in-education surveys (Gao et al., 2025): a domain contribution as the first open-weight LLM-agent study on Indonesian LAM-style accreditation; a cross-modal RAG contribution that joins narrative PDF and structured Excel evidence in a single agent state; a tool-calling paradigm contribution that contrasts native function calling (Qwen, Mistral) with prompt-based JSON dispatch (DeepSeek-R1) on the same task; and a calibration contribution that documents confidence-error inversion in DeepSeek-R1 and false-refusal-with-overconfidence in Mistral on Indonesian institutional documents, extending the calibration literature (Lin, Hilton, & Evans, 2022; Geng, Yang, Cao, Zhang, & Tang, 2024) to a new domain.

METHODS

a. Approach and Scope

This study uses a quantitative experimental approach with single-run benchmarking against three open-weight LLMs. To address the limitation of an earlier Tri-Dharma-only sample, the scope is broadened to 19 indicators (≈32% of the 60-indicator LAM-Teknik instrument), purposively stratified across all nine criteria (VMTS, Tata Pamong, Mahasiswa, SDM, Keuangan-Sarpras, Pendidikan, Penelitian, PkM, Luaran) and across all three score-type families: 9 qualitative, 6 quantitative, and 4 composite indicators. The minimum cell size of $n \geq 4$ per type permits a stable per-type breakdown.

b. Operational Variable Definitions

The primary dependent variable is the model score (ordinal 0–4) for each indicator, compared against the human-assessor reference score. Supporting variables are the justification text, the model-stated confidence (0–1), the runtime per indicator (seconds), and the indicator type (qualitative / quantitative / composite). Independent variables are the model (Qwen 3.5 35B / DeepSeek-R1 32B / Mistral-Small 3.2 24B) and the indicator type.

c. Location and Sample

The study was conducted on one Applied Bachelor in Telecommunication Engineering Distance Education program at an Indonesian polytechnic. The LED and LKPS documents are the post-audit version from the 2024 cycle. Single-program scope is intentional as a tractable initial study; multi-program generalization is acknowledged as a limitation.

d. Materials and Main Tools

The system is built with:

- Embedding model paraphrase-multilingual-MiniLM-L12-v2 (384-d).
- Vector database ChromaDB with HNSW cosine similarity.
- LangGraph as the agent state-machine framework.
- Ollama as the inference backend on a 24 GB VRAM GPU server accessed via a Cloudflare-tunnel HTTPS endpoint.

The technical components described above are integrated to transform the Manual Flow (Baseline) into the Automated Flow (Proposed) as illustrated in Figure 1. The synergy between LangGraph and ChromaDB enables the system to transition seamlessly from data ingestion (Self-Evaluation Report PDF and Study Program Performance Report Excel) to RAG-based automated scoring. Supported by the Ollama inference backend, the system processes 60+ indicators simultaneously to generate justifications and confidence scores, which are then streamed into a CSV format for final validation by the assessor through a human-in-the-loop mechanism.

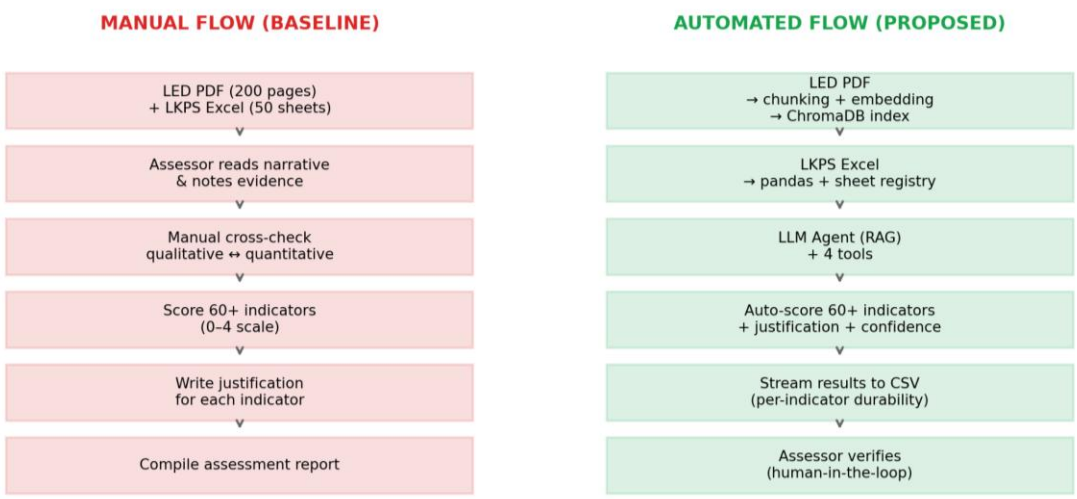

Figure 1. Comparison of Manual vs Automated Workflow

e. Four Agent Tools

The agent has access to four LangChain tools: (a) get_indicator_rubric(indicator_no) returns the score criteria, formula, and LKPS table references; (b) search_led_evidence(query, n_results) performs vector search over LED chunks; (c) get_lkps_table(table_reference) returns the LKPS table as markdown; and (d) compute_formula_score(formula_expression, variables) is a sandboxed Python evaluator for formula-based indicators.

f. Two Agent Paradigms

Two paradigms are applied automatically based on the model-name pattern. The native paradigm (Qwen 3.5 and Mistral-Small) uses structured tool_calls directly executed by LangGraph's ToolNode. The prompt-based paradigm (DeepSeek-R1) instructs the model to emit single-line JSON of the form *"{"action": ..., "args": ...}"* for tool calls or *"{"action": "final", "score": ..., "justification": ..., "confidence": ...}"* for the final answer, with the *"<think>...</think>"* block stripped before JSON extraction.

g. Data Collection

Each indicator is evaluated once per model with identical sampling parameters (temperature = 0.1, max_tokens = 8192), producing 57 invocations (3 models × 19 indicators). Results are streamed to benchmark_results.csv per row for crash-durability.

h. Ground-Truth Construction (Validity Test)

Ground-Truth Construction (Validity Test). The reference scores used in this benchmark are the official post-audit scores recorded by the certified human assessor during the program's 2024 internal-audit cycle, after the assessor reviewed the LED narrative and the LKPS tables and produced a justification per indicator. Two properties support the validity of this reference set: (a) the scores reflect the actual decision-making of a trained accreditation assessor familiar with the LAM-Teknik rubric, and (b) each score is anchored to the assessor's own written justification, which itself underwent the post-audit verification loop and is therefore traceable to the same evidence base (LED chunks + LKPS rows) that the benchmarked LLMs subsequently access. The 19 reference scores span 2.0 to 4.0 (mean 3.39, SD 0.65) and cover all nine LAM-Teknik criteria.

i. Statistical Analysis Techniques

The quantitative analysis uses six metrics:

1. Mean Absolute Error (MAE): MAE = $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
2. Root Mean Squared Error (RMSE): RMSE = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
3. exact-match accuracy as the fraction of indicators with $|y_i - \hat{y}_i| = 0$
4. within ±1 accuracy as the fraction with $|y_i - \hat{y}_i| \leq 1$, the commonly accepted inter-assessor agreement threshold
5. Pearson r as the linear correlation between model and reference scores
6. the score-zero rate as a proxy for the false-refusal failure mode.

The analysis is performed both in aggregate and disaggregated by indicator type (qualitative / quantitative / composite) to surface per-type failure profiles.

RESULTS AND DISCUSSION

a. Aggregate Performance

Table 1 summarizes the aggregate performance of the three models on the 19-indicator sample. Two findings stand out. First, Qwen 3.5 dominates accuracy across every metric: lowest MAE (0.605), lowest RMSE (0.769), highest exact match (36.8%), and the only model with 100% within ±1 accuracy. Second, the runtime ordering inverts the prior 7-indicator finding: Mistral is now the fastest (13.6 s/indicator), DeepSeek-R1 the slowest (79.9 s/indicator), the increased DeepSeek runtime is attributable to its extensive "*<think>*" deliberation on a broader indicator set.

Table 1. Aggregate Performance on 19 Indicators (n = 19 per model)

Model	MAE	RMSE	Exact	Within ±1	Pearson r	Conf.	RT/ind
qwen3.5:35b	0.605	0.769	36.8%	100.0%	0.108	0.78	32.5 s
deepseek-r1:32b	1.395	1.825	21.1%	63.2%	0.148	0.83	79.9 s
mistral-small3.2:24b	2.184	2.598	21.1%	31.6%	0.204	0.53	13.6 s

DeepSeek-R1 has the highest average confidence (0.83) despite the second-largest error band, a clear overconfidence pattern previously documented for reasoning-distilled models (Geng et al., 2024). Mistral combines low average confidence (0.53) with the lowest within ±1 accuracy, indicating that even its low-confidence outputs are often destructive (score 0). Pearson correlations are uniformly weak (0.108–0.204), so LLM scores should be treated as candidate evidence for human review rather than as a stand-alone ranking.

b. Break Down Analysis by Scoring Type

To gain a deeper understanding of the error patterns of each model, an analysis was conducted based on the type of indicators. Table 2 and Figure 2 disaggregate the performance by indicator type. The three models reveal sharply different per-type failure profiles.

Table 2. Breakdown by Scoring Type

Model	Type	n	MAE	Exact	±1	Score Zero Count
qwen3.5:35b	Qualitative	9	0.44	55.6%	100.0%	0
qwen3.5:35b	Quantitative	6	0.67	33.3%	100.0%	0
qwen3.5:35b	Composite	4	0.88	0.0%	100.0%	0
deepseek-r1:32b	Qualitative	9	1.11	33.3%	77.8%	1
deepseek-r1:32b	Quantitative	6	2.17	0.0%	33.3%	2
deepseek-r1:32b	Composite	4	0.88	25.0%	75.0%	0
mistral-small3.2:24b	Qualitative	9	2.89	0.0%	11.1%	7
mistral-small3.2:24b	Quantitative	6	0.50	66.7%	83.3%	0
mistral-small3.2:24b	Composite	4	3.12	0.0%	0.0%	3

Qwen 3.5 maintains 100% within ±1 across all three types, it is the only model with a balanced per-type profile. DeepSeek-R1 behaves opposite to the conventional expectation that reasoning models excel at numbers: its quantitative MAE (2.17) is twice its qualitative MAE (1.11), driven by failures on RBK (indicator 30) and RIPK (indicator 42), where the model retrieved the correct LKPS rows but its <think> block reasoned itself into a "data is insufficient" refusal. Mistral shows the most dramatic inversion: 7 of 9 qualitative indicators receive score 0 (false-refusal rate ≈ 78%) while its quantitative MAE is the lowest of all three models at 0.50. The per-type MAE pattern is visualized in Figure 2.

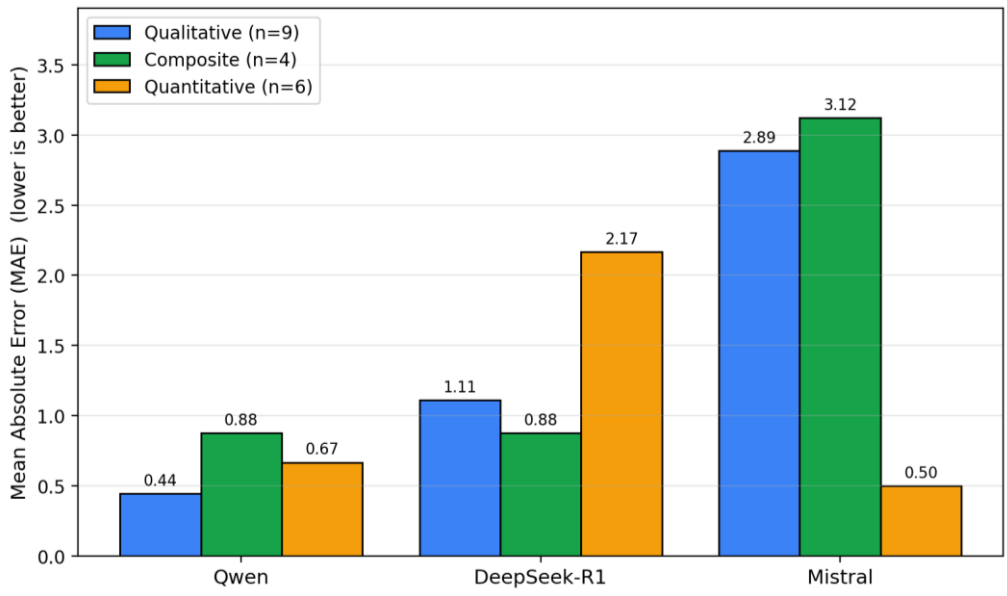

Figure 2. Comparison of MAE per Indicator Type

c. Analysis of Indicator Scoring per Model

Table 3 reports the per-indicator scoring for all three models against the human-assessor reference (marker: ✓ = exact, ~ = within-±1, ✗ = differs by > 1).

Table 3. Per-Indicator Scores Across the Three Models

No	Type	Ref	deepseek-r1	mistral-small	qwen3.5
1	Qualitative	3.0	4.0 ~	0.0 ✗	3.0 ✓
2	Qualitative	3.0	4.0 ~	0.0 ✗	3.0 ✓
4	Composite	2.0	1.0 ~	4.0 ✗	3.0 ~
7	Qualitative	3.0	3.0 ✓	0.0 ✗	3.0 ✓
8	Qualitative	4.0	4.0 ✓	0.0 ✗	4.0 ✓
13	Composite	3.0	2.0 ~	0.0 ✗	4.0 ~
17	Qualitative	2.0	1.0 ~	0.0 ✗	3.0 ~
22	Qualitative	3.0	2.0 ~	0.0 ✗	4.0 ~
28	Quantitative	4.0	2.0 ✗	4.0 ✓	4.0 ✓
30	Quantitative	3.0	0.0 ✗	3.0 ✓	3.0 ✓
32	Quantitative	4.0	2.0 ✗	3.0 ~	3.0 ~
35	Quantitative	3.0	4.0 ~	1.0 ✗	4.0 ~
39	Qualitative	4.0	2.0 ✗	1.0 ✗	3.0 ~
42	Quantitative	4.0	0.0 ✗	4.0 ✓	3.0 ~
48	Qualitative	4.0	0.0 ✗	0.0 ✗	4.0 ✓
50	Quantitative	4.0	3.0 ~	4.0 ✓	3.0 ~
53	Composite	3.5	2.0 ✗	0.0 ✗	3.0 ~
58	Composite	4.0	4.0 ✓	0.0 ✗	3.0 ~
60	Qualitative	4.0	4.0 ✓	3.0 ~	3.0 ~

Qwen 3.5 is the only model that never breaches the ±1 band on any of the 19 indicators and never produces a score of 0; it matches the reference exactly on seven items (ind. 1, 2, 7, 8, 28, 30, 48) and stays within one point on the remaining twelve. Its conservative tendency is visible on high-score indicators (39, 42, 50, 53, 58, 60) where the reference is 4 but Qwen issues 3, a one-point under-estimation that is easy for a human assessor to verify and adjust. DeepSeek-R1 displays a bimodal pattern: on qualitative items (ind. 7, 8, 60) and the high-evidence composite item 58 it matches exactly, but it accumulates five large miss-by-more-than-one errors (ind. 28, 30, 32, 42, 48, 53) clustered on quantitative items. Chain-of-thought inspection of the <think> block for ind. 30 (Beban Kerja DTPS) and ind. 42 (IPK Lulusan) shows the model correctly retrieving the LKPS rows but rejecting them with phrases like "the LKPS table is partial", an over-cautious refusal that the model rationalizes inside its own reasoning chain.

Mistral-Small 3.2 exhibits a near-total collapse on qualitative and composite items: ten of thirteen non-quantitative items receive score 0, including straightforward indicators such as Pengelolaan Keuangan (ind. 8) where the LED p.52 explicitly states all five required aspects. However, Mistral matches the reference exactly on three of six quantitative items (ind. 28, 30, 42, 50), every formula-driven item.

Failure Mode: False Refusal

Across the 19 indicators, the dominant failure mode is false refusal, the model emits a score of 0 with the justification "no evidence" even when LED chunks containing the evidence are retrieved. Table 4 summarizes the false-refusal rate per model.

Table 4. False-Refusal Rate per Model

Model	Total score = 0	Affected indicators
qwen3.5:35b	0	none
deepseek-r1:32b	3	30, 42, 48
mistral-small3.2:24b	10	1, 2, 4, 7, 8, 13, 17, 22, 39, 48, 53, 60

Qwen 3.5 shows perfect reliability, it never refuses to score. DeepSeek-R1 has three false refusals, with chain-of-thought inspection revealing over-cautious phrasing such as "evidence is too generic to verify all four aspects". Mistral exhibits severe false refusal on 10 of 19 indicators, consistent with the over-cautious instruction-tuning pattern previously reported on Indonesian-language tasks (Nurohim et al., 2025).

Discussion

Comparison with Prior Work

Compared to Aljarallah and Dutta (2022), who built a rule-based "Quality Automation Framework" for Saudi accreditation specifications, the present agent removes the maintenance burden of rule encoding by delegating evidence retrieval to a learned retriever and scoring to an LLM. Compared to Sakinah et al. (2025), who used K-Means + OLAP for Indonesian accreditation analytics, our system operates on the source-of-truth documents (LED + LKPS) rather than on aggregated metrics, eliminating an ETL hop that loses fine-grained evidence. Compared to Demir and Yavuz (2026), a generic comparative LLM study on multidimensional data, this work narrows scope to a regulated audit task with verifiable score-rubric grounding, where the generative-model strengths they report translate into both opportunities (Qwen) and risks (Mistral hallucination). Compared to the AI-assisted self-evaluation frameworks proposed at a conceptual level by Başaran (2026) and Ticolau (2026), this study contributes an end-to-end implementation with a reproducible benchmark and an open per-indicator error log. The trade-off identified here, Qwen's reliability against DeepSeek's overconfidence and Mistral's qualitative collapse is consistent with the calibration literature on reasoning-tuned and instruction-tuned mid-sized LLMs (Lin et al., 2022; Geng et al., 2024) but extends it to Indonesian accreditation documents, a domain not previously covered.

Implication for Practice

For deployment as an AI draft scorer to be verified by a human assessor, Qwen 3.5 35B is recommended because its 100% within ±1 accuracy means the assessor only needs to correct at most one scale point. DeepSeek-R1 may be used for composite-heavy batches with awareness of occasional false refusals on quantitative items. Mistral-Small 3.2 should be avoided for qualitative-heavy batches because its 78% false-refusal rate would require extensive manual correction; it remains useful for quantitative-only fast scoring (13.6 s per indicator, MAE 0.50). This recommendation aligns with the human-in-the-loop assistive tool framing: the final decision remains with the certified assessor, and the LLM agent serves only as an accelerator that produces verifiable draft scores and justifications.

Limitations

Four limitations bound this study. First, the reference scores come from a single certified assessor on a single post-audit cycle; inter-rater reliability across multiple human assessors is not yet quantified and would strengthen the gold-standard claim. Second, n = 19 is purposively stratified for diagnostic per-type analysis, so the reported MAE/RMSE figures are descriptive of behavior on a stratified diagnostic set rather than a population point estimate. Third, the benchmark uses a single

deterministic-leaning sampling configuration; sensitivity to temperature and "*top_p"* is not explored. Fourth, the study is single-site, so cross-institution generalization is unverified.

CONCLUSION

This study developed and benchmarked a cross-modal RAG agent for Internal Quality Audit in Indonesian LAM-Teknik accreditation, evaluating three open-weight LLMs (Qwen 3.5 35B, DeepSeek-R1 32B, Mistral-Small 3.2 24B) on 19 indicators sampled across all nine criteria. Qwen 3.5 is the only model that satisfies operational accuracy across qualitative, quantitative, and composite types simultaneously (MAE 0.605, 100% within ±1, 32.5 s per indicator). DeepSeek-R1 is the most cautious in qualitative items but is the slowest and collapses on quantitative indicators (MAE 2.17). Mistral-Small is the fastest (13.6 s) and surprisingly strong on quantitative items (MAE 0.50) yet fails 78% of qualitative items by emitting a score-zero refusal. End-to-end, the agent reduces AMI cycle time from 3–5 days to under one hour per program while preserving a fully auditable evidence trace for human verification. For practical deployment as an AI draft scorer the study recommends Qwen 3.5 35B because of its consistent reliability. Future work will (1) extend the benchmark to the full 60-indicator instrument across multiple study programs; (2) measure inter-rater reliability across multiple certified human assessors to strengthen the gold-standard reference; and (3) introduce a refusal-suppression filter and a confidence-recalibration head for production deployment.

REFERENCES

Aljarallah, N. A., & Dutta, A. K. (2022). Developing a Quality Automation Framework to Assess Specifications for Academic Accreditation in Saudi Arabian Universities. TEM Journal, 11(2), 667–674. https://doi.org/10.18421/TEM112-22

Aydin, O., Karaaslan, E., Erenay, F. S., & Bacanin, N. (2026). Generative AI in Academic Writing: A Comparison of DeepSeek, Qwen, ChatGPT, Gemini, Llama, Mistral, and Gemma. Turkish Journal of Engineering, 10(1), 3–18. https://doi.org/10.31127/tuje.1545876

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., … Zhu, T. (2023). Qwen Technical Report. arXiv preprint arXiv:2309.16609. https://doi.org/10.48550/arXiv.2309.16609

Başaran, B. (2026). Evaluating Large Language Models for Educational Measurement: Insights from Automated and Human Scoring of Language Exams. Journal of Artificial Intelligence and Technology, 6(2), 349–355. https://doi.org/10.37965/jait.2026.0349

DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv preprint arXiv:2501.12948. https://doi.org/10.48550/arXiv.2501.12948

Demir, F., & Yavuz, U. (2026). Evaluating the Performance of Generative Artificial Intelligence Models in Multidimensional Data Analysis Tasks: A Comparative Study with Large Language Models. Discover Computing, 29(1), 1–28. https://doi.org/10.1007/s10791-025-09584-9

Efendi, Y., Hapsari, R. F., & Yusup, R. R. (2025). Thematic Analysis of Accreditation and ISO 21001 at Higher Education Institution in Batam. International Journal of Management, Innovation, and Education, 4(1), 1–8. https://doi.org/10.56873/ijmie.v4i1.1023

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2025). Retrieval-Augmented Generation for Large Language Models: A Survey. ACM Computing Surveys, 57(7), Article 178. https://doi.org/10.1145/3701228

Geng, J., Yang, X., Cao, B., Zhang, Y., & Tang, J. (2024). A Survey of Confidence Estimation and Calibration in Large Language Models. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024), 6577–6595. https://doi.org/10.18653/v1/2024.naacl-long.366

Isnaeni, N., Putranto, B. P., Andriyani, W., & Khomsah, S. (2025). Comprehensive Lakehouse Data Architecture Model for College. Journal of Dinda Data Science, Information Technology, and Data Analytics, 5(1), 47–58. https://doi.org/10.31544/jdda.v5i1.470

Jegham, N., Abdelatti, M., & Hendawi, A. (2025). Visual Reasoning Evaluation of Grok, DeepSeek's Janus, Gemini, Qwen, Mistral, and ChatGPT. arXiv preprint arXiv:2502.16428. https://doi.org/10.48550/arXiv.2502.16428

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 1, 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

Mistral AI. (2024). Mistral Small 3 Model Card [Technical report]. Mistral AI. Retrieved from https://mistral.ai/news/mistral-small-3

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A Comprehensive Overview of Large Language Models. ACM Transactions on Intelligent Systems and Technology, 16(2), Article 27. https://doi.org/10.1145/3744746

Nurohim, G. S., Setyadi, H. A., & Fauzi, A. (2025). Benchmarking Deepseek-LLM-7B-Chat and Qwen1.5-7B-Chat for Indonesian Product Review Emotion Classification. Journal of Applied Informatics and Computing (JAIC), 9(5), 3068–3078. https://doi.org/10.30871/jaic.v9i5.8893

Sakinah, P., Syaputra, A. E., Rahman, Z., Fajri, M., & Rachmansyah, H. F. (2025). Optimalisasi Akreditasi Perguruan Tinggi dengan Orkestrasi Business Intelligence Berbasis K-Means dan OLAP. Jurnal Pusat Akses Kajian Teknologi Artificial Intelligence, 2(2), 550–561. https://doi.org/10.62411/pakaitai.v2i2.550

Sakuroh, L., Marlina, R., Krismayanti, Y., & Adiningsih, N. U. (2026). Quality Assurance Transformation in Higher Education Institutions: Linking Accreditation, Digitalization, and Organizational Performance. Socius: Jurnal Penelitian Ilmu-Ilmu Sosial, 13(1), 3–14. https://doi.org/10.24252/socius.v13i1.0304

Subijanto, Kadaryanto, B., Ali, N. B., Sulistiono, A. A., Widiputera, F., & Martini, I. A. (2021). Quality Assurance System of Distance Education. Jurnal Penelitian Kebijakan Pendidikan, 14(2), 135–150. https://doi.org/10.24832/jpkp.v14i2.521

Sugiarti, E. (2022). The Impact of Tri Dharma Performance on Higher Education. Jurnal Mahasiswa Humanis, 2(2), 120–123. https://doi.org/10.37481/jmh.v2i2.524

Sun, Y., Chen, T., Liu, Y., Wang, Z., Wang, R., & Liu, K. (2025). A Systematic Review on LLM-Powered Automated Grading and Assessment in Education. arXiv preprint arXiv:2508.02442. https://doi.org/10.48550/arXiv.2508.02442

Ticolau, V. E. (2026). Designing an AI-Assisted Self-Evaluation Framework to Support Higher-Education Accreditation. Jurnal Teknologi Informasi dan Pendidikan, 18(1), 1279–1290. https://doi.org/10.24036/jtip.v18i1.1279