European Journal of Artificial Intelligence and Machine Learning
Vol 4 | Issue 5 | October 2025
ISSN 2796-0072

RESEARCH ARTICLE

Comparing AI-Generated Preview and Portfolio Feedback: Gpt 4.o vs. Claude 4

Thomas Kropmans[1,*], Oleh Bilokrylyi[1], Dmytro Predchyshyn[2], David Cunningham[1], Edward Melvin[3], and Gabia Neverauskaité[1]

ABSTRACT

This report provides an in-depth comparative analysis of AI-generated portfolio feedback delivered through two leading Large Language Model platforms (LLM): Gpt4.o OpenAI and Claude-sonnet-4 (Anthropic) via Amazon Bedrock. The feedback was analyzed in two distinct stages: preview feedback, which serves as a safety and verification layer for examiners/administrators, and portfolio feedback, which is delivered directly to the students. These systems are integral to Qpercom's digital assessment tools and support high-stakes clinical assessments such as Objective Structured Clinical Examinations (OSCEs), high-stake recruitment using Multiple Mini Interviews (MMIs), and Video Interviewing and Digital Scoring (VIDS). This evaluation examines how accurately each model reflects students' actual high-, mid-, and underperformance and whether its feedback provides safe, constructive, and educationally valuable input.

Keywords: Clinical education, feedback, LLM, OSCE.

Submitted: August 28, 2025

Published: October 14, 2025

10.24018/ejai.2025.4.5.82

[1] *Qpercom Ltd Platform 94 Mervue Business Park Galway, Ireland.*
[2] *InventorSoft, Ukraine.*
[3] *Marino_Software, DCU Alpha, Dublin.*

Corresponding Author:
e-mail: thomas.kropmans@qpercom.ie

1. Introduction

AI offers scalable solutions for formative and summative feedback in the context of increasing digitization in medical education and recruitment. Gpt4.o and Claude 4 represent two of the most sophisticated AI platforms available, each with distinct language-generation models and prompt design capabilities. Gpt4.o is known for its creative and flexible natural language outputs and is often praised for its empathetic and student-friendly tone [1]. Claude 4, on the other hand, is designed with enterprise-grade reliability and greater customizability, often resulting in structured, detailed, and technically accurate outputs [2].

Recent MSc (Kamilla Mahrlamova, 2024) and PhD theses (2025), including that by Akram Alsahafi and Kropmans, emphasize the necessity for actionable, balanced, and behaviorally anchored feedback in OSCEs [3]. Their work advocates avoiding overly general praise and ensuring that feedback is specific to clinical actions and observable behaviors. This study explored the degree to which these principles are upheld by each AI system.

A broader educational challenge also emerges in the context of Multiple Mini Interviews (MMIs), which are internationally used as recruitment tools for healthcare education programs. Despite their widespread use, applicants often receive no feedback on their performance, leaving them without developmental insight if they consider re-applying. Furthermore, interviewers lack access to data-driven feedback, limiting their ability to improve consistency and fairness in feedback delivery. Addressing this shortcoming with AI-generated feedback holds substantial promise for improving transparency and fairness in educational assessment and recruitment, a key concern in the business-process and organizational development literature [4], [5].

2. Methodology

We examined feedback for 51 (15%) anonymized OSCE candidates from a random data file containing anonymous data of approximately 340 students using a mixed-method approach. Seventeen students had high performance scores and 17 were underperformers. Seventeen students were randomly selected as mid-performers. High performance scores were above the cut-off score plus one Standard Error of Measurement (SEM). Mid-performers

had a score between cut-score plus or minus 1 SEM. In contrast, underperformance means a result lower than the cut-off score minus one SEM. For each student, we collected the preview and portfolio feedback generated by the Gpt4.o and Claude systems. Despite offering five methods of feedback (FEEDME, Boost, Rule of 3, AID, and Calgary Cambridge), for this research, we used the most commonly used method (Rule of 3), providing three aspects that went well and three aspects that needed improvement. These AI-generated texts were then cross-referenced with detailed performance data, including station item scores, examiner notes, and expected behavioral observations from station notes.

In this study, we distinguish between two stages of AI-generated feedback: *preview* and *portfolio*. *Preview feedback* refers to the initial AI-generated narrative presented to examiners/administrators for verification prior to being finalized. This feedback was designed to be reviewed, amended if necessary, and validated for factual accuracy, tone, and educational value. In contrast, *portfolio feedback* is the finalized version delivered directly to students as part of their official performance record within the Qpercom platform. Both feedback types are generated from the same underlying assessment data, prompt, and API (Application Programming Endpoint (API) endpoint, yet were found to vary in length, specificity, and in some cases, factual consistency, raising questions about model behavior between stages of deployment.

To ensure robust evaluation, we employed both quantitative and qualitative analyses. Word counts and standard deviations were calculated to examine verbosity and variation. We also flagged hallucinations, defined as AI-generated content that is not supported by the underlying performance data. Thematic coding was guided by Alsahafi and Kropmans's taxonomy of specificity, developmental value, balance (strengths and areas for improvement), and behavioral observation. This methodology aligns with the recent literature comparing LLM feedback with expert-written educational responses [6], [7]. For broader organizational relevance, comparisons were also considered through a business-process lens, consistent with performance feedback systems used in professional recruitment contexts.

3. Results

The cut-off score calculated for this random and anonymized cohort within Qpercom's advanced assessment module was 63.26% with a Standard Error of Measurement (SEM), providing a 68% confidence interval of +/− 6.88% (lower bound 56.38%; upper bound 70.14%). Overall Cronbach's Alpha is 0.68 varying between 0.55 and 0.79 (poor-moderate), the sample size standard deviation is 12.43% (score range 0%–100%). Furthermore, a random group of an additional 17 students was selected out of 254, the results of which varied between the upper and lower bounds of the cut-off score (56.38%–70.14%) of the sample (254 out of 340). Additional feedback analysis was performed to verify the midgroup results.

3.1. Word Count Comparison

We refer to Table I for the word count comparison between OpenAI and Claude 4.

3.2. Thematic Summary

- *Specificity:* Portfolio feedback was more narrative, often losing bullet-point clarity from previews.
- *Developmental Value:* Preview feedback contained clearer developmental goals. Portfolios used supportive language, sometimes diluting precision.
- *Balance:* Portfolio feedback leaned more positive, even for underperformers.
- *Tone:* The portfolio was more student-friendly but occasionally at the expense of accuracy.
- *Hallucination:* Detected in Gpt4.o portfolio versions for underperformers.
- *Factual Alignment:* Claude 4 feedback remained accurate and consistent, whereas Gpt4.o's portfolio often introduced inaccuracies.

3.3. Key Insight

Gpt4.o's portfolio feedback was generally longer and exhibited a greater variation in length. This may increase the risk of content hallucinations or over-generalized encouragement. Claude 4 feedback maintained a more consistent word count across students, suggesting greater stability and adherence to the structured templates.

3.4. Overall Insights

OpenAI hallucinations occurs not only exclusively in underperformer portfolio feedback. OpenAI hallucinations occurs also in mid-performers portfolio feedback (Table II). Claude 4 exhibits 0% hallucination and perfect score alignment across all cases. Actionable improvement suggestions are consistently present across both models.

Gpt4.o portfolio feedback frequently employs a reassuring and emotionally supportive tone, which may benefit student morale but at times compromises accuracy. Claude 4, while slightly less warm in phrasing, was consistently reliable and constructive, offering tangible steps for improving and maintaining a professional standard that supports learning (Table III). These observations align with findings from recent LLM feedback studies that show that large models may underperform in identifying errors while maintaining high linguistic fluency (Zhou Z, 2024).

3.5. Comparison with Actual Performance Data

1. Claude 4 (Preview and Portfolio)

TABLE I: Word Count Analysis

Model	Feedback type	Average word count	Standard deviation
OpenAI	Preview	474.3	55.8*
OpenAI	Portfolio	644.2	91.4*
Claude 4	Preview	532.7	39.3
Claude 4	Portfolio	630.0	56.1

Note: Word Count Analysis between OpenAI Preview vs. Portfolio: $t = -5.407$, $p = 0.0004$ (significant).Claude Preview vs. Portfolio: $t = -1.191$, $p = 0.2581$ (not significant); OpenAI Preview vs. Claude Preview: $t = -2.717$, $p = 0.0147$ (Significant). OpenAI Portfolio vs. Claude Portfolio: $t = 0.304$, $p = 0.7641$ (not significant).

TABLE II: Accuracy Summary by Percentage

Model	Performance	Feedback type	Hallucination (%)	Score alignment (%)	Actionable targets (%)
OpenAI	High Performer	Preview	0%	100%	100%
OpenAI	High Performer	Portfolio	0%	100%	100%
OpenAI	Underperformer	Preview	0%	100%	100%
OpenAI	Underperformer	Portfolio	80%	20%	100%*
OpenAI	Mid-performer	Preview	0%	100%	100%
OpenAI	Mid-performer	Portfolio	100%	0%	60%*
Claude	Mid-performer	Preview	0%	100%	100%
Claude	Mid-performer	Portfolio	0%	100%	100%
Claude	High Performer	Preview	0%	100%	100%
Claude	High Performer	Portfolio	0%	100%	100%
Claude	Underperformer	Preview	0%	100%	100%
Claude	Underperformer	Portfolio	0%	100%	100%

Note: *Key Insight: OpenAI's GPT-4.o preview feedback aligns well with scores but is significantly altered—introducing hallucinations and losing specificity—in the portfolio stage. Claude 4 maintains both score alignment and fidelity throughout.

TABLE III: Quality Preference Evaluation

Criterion	Gpt4.o portfolio	Claude 4 portfolio	Preferred source
Accuracy to scores	Often exaggerated	Consistently aligned	Claude 4
Detail and Specificity	Inconsistent per student	Detailed per station	Claude 4
Transparency on failures	Frequently softened	Direct and supportive	Claude 4
Language and Tone	Friendly, supportive	Supportive and factual	Tie

- Score Alignment: 100%
- Hallucinations: 0%
- Accurate differentiation between high and underperforming students.
- Faithful representation of performance without exaggeration or omission.

2. Comparison with OpenAI:

- OpenAI Preview Feedback: generally accurate, especially for high performers.
- OpenAI Portfolio Feedback for Under performers:
    - Often hallucinates or softens critical failures.
    - Only 20% aligned with actual performance data.

3.6. Key Patterns for High, Mid-Performers and under Performers

- *Claude 4:* Consistently reflects the data sheets across both the feedback and performance categories. Preservation structure, rubric alignment, and factual integrity. Softens tone for portfolio, but retains performance specificity.
- *OpenAI:* shows reliable performance only for high-performing students in both stages. Tends to excessively generalize or soften portfolio feedback, resulting in hallucinations and reduced actionability.

This suggests that Claude 4 is currently the most dependable model for generating accurate, safe, and educationally valuable feedback.

3.7. GRS Interpretation Summary

Across the full sample of 51 students, significant differences emerged in how OpenAI and Claude 4 applied Global Rating Scores (GRS) in comparison to examiner judgments. Notably, Claude 4 demonstrated strong alignment with the examiner's GRS across all performance levels, accurately distinguishing between fail (0), borderline fail (1), borderline pass (2), pass (3), good (4), and excellent (5). In contrast, OpenAI frequently deviated from examiner benchmarks, underestimating high performers by assigning pass instead of excellent, and overestimating underperformers by upgrading borderline fail-level scores to borderline pass or pass. These patterns suggest a tendency in OpenAI's portfolio feedback to soften extremes, likely contributing to a drift away from true summative judgment. Claude 4, however, maintained score integrity and showed a better understanding of the rubric and performance thresholds (Table IV).

4. Discussion

Although we found statistically significant outcomes, including differences in word count, hallucination frequency, and model fidelity across performance categories, caution is warranted when generalizing these findings. The study was limited to a subset of students (n = 51), and variations in prompt interpretation or model updates could have affected replicability. Nonetheless, the observed trends, particularly the consistent factual alignment of Claude 4 and the vulnerability of OpenAI feedback in low- and mid-performance scenarios, highlight meaningful differences in model reliability (Table V). These insights can guide informed decision-making for institutions integrating AI feedback tools in high-stakes educational

TABLE IV: Comparative Analysis: AI Feedback for 17 Mid-Performers–OpenAI and Claude 4

Student ID	AI model	Preview feedback	Portfolio feedback	Key differences	Hallucination	Score alignment
200152075	OpenAI	Balanced and diagnostic	More narrative, softened tone	Less specific action points	Yes	Partial
200152075	Claude 4	Structured and accurate	Consistent, minor rewording	Tone softened slightly	No	Full
200153164	OpenAI	Strong rubric references	Generalised praise	Loss of detail	Yes	Partial
200153164	Claude 4	Organized, rubric-aligned	More structured tone	Minor clarity edits	No	Full
200360647	OpenAI	Station-focused	More generalised, longer	Reduced clinical depth	Yes	Partial
200360647	Claude 4	Technical and clear	Maintained feedback themes	Minor tone shift	No	Full
200360795	OpenAI	Diagnostic and concise	Broader phrasing	Less actionable content	Yes	Partial
200360795	Claude 4	Highly aligned	Slightly simplified	Maintains clarity	No	Full
200360865	OpenAI	Concise, targeted	General and positive	Loss of detail	Yes	Partial
200360865	Claude 4	Structured and consistent	Minor tone polishing	Feedback consistency maintained	No	Full
200361437	OpenAI	Focused item critique	Generalised tone	Key gaps obscured	Yes	Partial
200361437	Claude 4	Critical but fair	Encouraging tone added	Lowered emphasis on gaps	No	Full
200361529	OpenAI	Focused on interpersonal skill	Overly broad praise	Missed concrete feedback	Yes	Partial
200361529	Claude 4	Clinically detailed	Polished language	Same specificity	No	Full
200361622	OpenAI	Insightful and sharp	Wordy and less direct	Weaker call to action	Yes	Partial
200361622	Claude 4	Clinically aligned	Consistent phrasing	Minor syntax changes	No	Full
200361987	OpenAI	Specific and item-based	More polished tone	Slightly diluted analysis	Yes	Partial
200361987	Claude 4	Structured and detailed	Balanced summarising	Maintained alignment	No	Full
200362010	OpenAI	Diagnostic clarity	Simplified expressions	Fewer clinical examples	Yes	Partial
200362010	Claude 4	Balanced clinical framing	Polished summary	Aligned tone	No	Full
200362021	OpenAI	Pointed and actionable	Narrative and safe	General comments increased	Yes	Partial
200362021	Claude 4	Rubric matched	Slightly softer tone	Consistency maintained	No	Full
200362180	OpenAI	Data-driven focus	Redundant praise added	Less behavior-focused	Yes	Partial
200362180	Claude 4	Objective feedback	Maintained clinical references	Structure unchanged	No	Full
200362375	OpenAI	Accurate gaps identified	Softened phrasing	Actionability reduced	Yes	Partial
200362375	Claude 4	Balanced station feedback	Minor linguistic edits	Clarity preserved	No	Full
200362582	OpenAI	Feedback-specific on technique	Abstracted praise	Missed actionable feedback	Yes	Partial
200362582	Claude 4	Diagnostic tone	Minor grammar edits	Clear and consistent	No	Full
200362892	OpenAI	Exam-specific feedback	General guidance tone	Details diluted	Yes	Partial
200362892	Claude 4	Behavioral references included	Tone adjusted for clarity	Structure unchanged	No	Full
200363349	OpenAI	Evidence-based pointers	Motivational reframing	Technical depth reduced	Yes	Partial
200363349	Claude 4	Specific about consultation	Tone softened, same points	Consistent clinical insight	No	Full
220177201	OpenAI	Score-calibrated critique	Motivational focus added	More generic tone	Yes	Partial
220177201	Claude 4	Criterion-linked detail	Slight restructuring	Maintains alignment	No	Full

Note: Detailed Comparison per mid-performing StudentID (Each ID Listed Twice for both OpenAI and Claude 4) for AI Model used; Preview feedback and Portfolio feedback; the Key differences between the two; Hallucination (Yes/No) and Score alignment.

TABLE V: Global Rating Score (GRS) Comparison–All 51 Students (Scale: 0 = Fail, 1 = Borderline Fail, 2 = Borderline Pass, 3 = Pass, 4 = Good, 5 = Excellent)

GRS level	Examiner	OpenAI	Claude 4
Excellent (5)	17	11	17
Good (4)	12	16	12
Pass (3)	11	14	11
Borderline Pass (2)	7	7	7
Borderline Fail (1)	4	3	4
Fail (0)	0	0	0

environments. AI-generated feedback produces more educationally relevant, accurate, and data-driven actionable feedback for all participants involved. Alsahafi observed that examiner feedback in OSCEs is often overly general, vague, and lacks actionable guidance [3], [8]. Common issues include the use of generic praise without referencing specific, observable behaviors and a failure to align comments with assessment criteria or station objectives. He emphasizes the need for behaviorally anchored feedback that supports student development and reflects actual performance. His work highlights the value of structured tools in enhancing the clarity, specificity, and educational impact of clinical feedback.

In line with Alsahafi's recommendations, our study shows that AI-generated feedback, particularly from Claude 4, can address many of the shortcomings typical of examiner feedback. The model consistently produced behaviorally specific, structured, and actionable responses that aligned with the performance data and station objectives. This suggests that when properly prompted and validated, AI can enhance the clarity and developmental value of feedback, offering a scalable solution to improve transparency and fairness in clinical assessment [9]. However, traditional paper-based OSCE processes often lead to delayed, generic feedback, usually reserved only for students who fail. This limits broader learning opportunities and reduces the perceived fairness of assessments. Because of our role as data processors, we are not able to verify the perceived fairness and educational value of this AI-generated feedback for participants, including examiners.

We conducted and reported a focused analysis of AI-generated feedback for mid-performing students, specifically comparing the outputs from the two AI models. This group was defined by scores within one Standard Error of Measurement (SEM) above or below the cut-off score, a zone often associated with borderline outcomes and increased ambiguity. Since the use of SEM in pass/fail decision-making is still uncommon, feedback for these candidates warrants more critical scrutiny. Consequently, variations in accuracy, alignment, and language between previews and portfolio feedback were more pronounced than those observed in the high- and underperforming groups.

Based on this evidence-based comparative analysis, Qpercom recommends prioritizing *Claude 4 Portfolio Feedback* for student-facing communication in high-stakes OSCE settings. The system's greater alignment with actual performance data, ability to produce precise and targeted developmental advice, and consistent avoidance of hallucinated content make it a safer and more educationally effective option. This conclusion has broader implications for integrating AI-based feedback not only in high-stake clinical formative and summative assessments but also in recruitment interviews such as MMIs, where structured feedback for applicants and interviewers remains lacking. AI models such as Claude 4 can be leveraged to enhance data-driven transparency, support process accountability, and improve standardization in high-stakes educational decision-making and selection contexts.

This study has several important limitations, primarily stemming from its sample size and generalizability.

4.1. *Drawbacks and Considerations*

1. *Accuracy Assessment Relies on Interpretation:*
   - Determining hallucinations and score alignment requires **subjective comparison** with human examiner notes, introducing possible bias or inconsistency.
2. *Single Context (OSCE):*
   - The findings are grounded in clinical OSCE data and may not generalize to other academic or practical contexts (e.g., MMIs, VIDS, written exams, and peer feedback).
3. *Model Updates & Prompt Dependency:*
   - As AI models evolve rapidly, the results may not remain stable over time or with prompt iterations.

4.2. *What Can We Improve Next Time?*

Several improvements are recommended in future studies. First, the sample size should be expanded to include 20%–30% of the student cohort, ensuring representation across all performance levels. Each performance band should be analyzed in detail using at least one full station per student. To better understand how well AI feedback aligns with human judgment, short interviews with examiners should be conducted to clarify their intent and to help identify any discrepancies or hallucinations in the AI-generated responses. Additionally, feedback sources should be triangulated by comparing AI feedback with examiner comments and station notes to offer a richer and more accurate analysis. The use of mixed-method scoring rubrics, evaluated by multiple raters, would help assess key dimensions, such as hallucination frequency, alignment with scores, tone, and developmental value. Importantly, future research should track student outcomes, such as learning gains, behavioral changes, and satisfaction levels to evaluate the real-world impact of AI feedback. Finally, reproducibility trials by academic institutions and other users of Qpercom's advanced assessment platform should be performed by re-running identical prompts across different model versions or sessions to assess the consistency and reliability over time.

5. Conclusion and Recommendations for Practice

This study is novel in its kind and uses an anonymized dataset of 340 students to explore qualitative differences in AI-generated feedback using both OpenAI's Gpt-4.o and Anthropic's Claude 4 models within Qpercom's advanced

assessment platform. Considering the descriptive statistics of 17 underperforming students with a score lower than the cut-off score minus 1 SEM, the average score was 56.30% and askew. A median score of 58.0% with a minimum score of 49.0% and a maximum score of 66.0% provides better insight.

Qpercom offers clients access to both models and a diverse set of feedback frameworks, including FeedMe, Boost, Rule of 3, AID, and Calgary-Cambridge. Despite the fact that we used Rule of 3 (three of a kind) for this research, we enable flexible adaptation to institutional needs. Based on our findings, we recommend that clients prioritize Claude 4 for high-stakes feedback (Rule of 3) because of its superior factual alignment and consistent tone, especially in underperforming cases. Institutions are encouraged to trial multiple feedback models across performance tiers and integrate examiner validation checkpoints. For optimal use, prompt frameworks should be tailored not only by the model but also by the assessment context, and feedback should be reviewed before dissemination when used in formative or summative evaluations. Moving forward, embedding AI feedback into structured reflection workflows and aligning it with educational outcomes (station notes) will be key to maximizing its pedagogical value.

5.1. Unexplained Variation in AI Output

Despite careful study design and strict control over prompting - using a single, consistent prompt structure for both preview and portfolio feedback (see below); we observed subtle yet consistent differences in output between the two. These variations were most notable in tone, specificity, and, in some cases, factual interpretation, particularly in OpenAI's portfolio feedback. One hypothesis is that the model may be influenced by prior content generated in the preview phase, effectively using earlier outputs as context or memory to refine subsequent responses. However, given that AI models such as Gpt-4.o and Claude 4 operate statelessly by default and do not retain conversational memory between sessions unless explicitly designed, this remains speculative. This raises important questions about potential internal caching, prompt sensitivity, or latent model behaviors that warrant further investigation. Clients, in their role as data controllers, should be aware of these nuances when comparing outputs from different stages of the feedback process, even when using standardized prompts.

The prompt we used is addressed below:

*<article> **Role definition:** You are a world-class clinical educator and medical feedback expert. </article><article> You specialize in delivering high-quality, evidence-based, actionable feedback for clinical skills trainees in healthcare education for ALL stations. </article><article> You use AI tools to translate assessment data and examiner comments into constructive, encouraging feedback that supports the development of knowledge, skills, and attitudes. </article><article> Generate detailed and constructive feedback based on all stations and the following inputs, in order of priority: 1. **Station notes* – frame and contextualise all feedback 2. **Examiner comments** – reference and paraphrase constructively 3. **Item descriptors** – link to observed behaviours 4. **Percent scores** – interpret using performance thresholds 5. **Global Rating Score (GRS)** – calibrate final judgment </article><article> You must incorporate **station instructions** wherever available to ensure that feedback is aligned with the clinical context and learning objectives. </article><article> **Examiner comments** must be paraphrased and integrated constructively to illustrate observed performance, avoiding direct quotations unless otherwise instructed. </article><article> Use the **Three of a Kind** structure: - **Three strengths** (what went well) - **Three areas for development** (with clear, specific actions) </article><article> Structure the written feedback as follows: 1. **Introduction**: Contextual summary of student performance in relation to station objectives 2. **Bullet Points**: - Three strengths (what went well) - Three areas for development (with actionable guidance) 3. **Conclusion**: Wrap-up remarks, motivational guidance, and practical action points </article><article> All feedback must be: - **SMART**: Specific, Measurable, Achievable, Relevant, Timely - **Evidence-based**: Grounded in station instructions, examiner comments, scores, and descriptors - **Positive and adaptive**: Support learner development and resilience </article><article> Avoid vague praise such as "excellent", "keep up the good work", or "can be improved." Use precise language tied to performance evidence. </article><article> Use percent scores and item descriptors to identify performance tiers: - <50% = significant concern - 50%–69% = developing - ≥70% = satisfactory/excellent </article><article> Where applicable, provide brief examples of good and poor practice that illustrate observed behaviours. </article><article> Use the student's first name where available to personalise feedback. </article><article> Only generate feedback for stations with available performance or examiner data. Do not speculate or invent content. </article><article> End the feedback with a **Global Rating Score**, selecting one of the following: - FAIL - BORDERLINE FAIL - BORDERLINE PASS - GOOD - EXCELLENT </article><article> Generate a **word cloud** that visually summarises the core themes in the feedback, based on the following input sources, in order of priority: 1. **Examiner comments** 2. **Item descriptors** 3. **Item scores and performance terms** (e.g. "hesitation", "structured approach", "rapport") 4. **AI-generated feedback text** </article><article> **Word cloud logic:** - Exclude stopwords and generic filler terms (e.g. "the", "can", "your", "may") - Focus on clinically and educationally meaningful keywords or short phrases (1–3 words) - Assign a 'frequency' score from 1–7 based on repetition or prominence in the feedback - Assign a 'color' to each word: - '"green"' = strength or positive theme - '"orange"' =*

*area for improvement - '"grey"' = technical or neutral. Return at least 30 words </article><article> **Output Format (JSON):** Provide a valid, parseable JSON object with two fields: - 'text': A string of HTML content, with each paragraph wrapped in ' ' tags - 'words': An array of word cloud entries in the format: '{ word: string; frequency: number; color: string; }[]' </article><article> Do **not** use Markdown syntax (no "'json or "'html). Output only the JSON object. Ensure structure is valid and safe for ingestion by external systems. </article>.*

In this study, a carefully stratified sample of 51 students (15% of the total cohort) was analyzed, with equal representation of high-, mid-, and underperformers. This approach ensured balanced insights across performance levels and allowed meaningful comparisons between AI feedback models. While the sample size limits the generalizability of the findings, it is well suited for the exploratory and comparative aims of this study, particularly in identifying model-specific differences in feedback quality, hallucination patterns, and score alignment. Structured sampling and focused methodology provide a solid foundation for initial conclusions, paving the way for future studies with larger cohorts and broader station coverage.

5.2. Data Governance and Client Responsibility

As a data processor, Qpercom facilitates the secure delivery of AI-generated feedback based on structured assessment data provided by our clients. Qpercom does not act as a data controller and, therefore, does not determine the purpose or means of processing personal data, including controlling various stages of feedback. The processed AI-feedback is based on evidence, whereas examiners' feedback often does not correspond directly with the assessment data [8]. Access to real-time or identifiable student data is restricted by design, ensuring compliance with privacy and confidentiality standards. Clients remain responsible for ethical deployment, interpretation, and educational integration of AI-generated evidence-based feedback. The findings of this study should inform client decision-making regarding model selection, prompt configuration, and human oversight protocols. We encourage institutions to conduct local evaluations of the impact of AI feedback and implement validation mechanisms to ensure that the feedback aligns with pedagogical goals and institutional assessment policies.

Conflict of Interest

The authors declare that they do not have any conflict of interest.

References

[1] Carlson M, Pack A, Escalante J. Utilizing OpenAI's Gpt-4 for written feedback. *Tesol J*. 2023;759:e759.

[2] Richard-Ojo O, Wimmer H, Redman CM Jr. RAG chatbot for healthcare related prompt susing Amazon Bedrock. *J Inform Syst Appl Res*. 2025 Oct;18(3):18–29. doi: https://doi.org/10.62273/RQAT8911.

[3] Alsahafi A, Ling DL, Newell M, Kropmans T. A systematic review of effective quality feedback measurement tools used in clinical skills assessment. *MedEdPublish*. 2023 Jun 19;12:11.

[4] Jo H, Shin D. The impact of recognition, fairness, and leadership on employee outcomes: a large-scale multi-group analysis. *PLoS One*. 2025;20(1):e0312951. doi: 10.1371/journal.pone.0312951.

[5] Anseel F, Sherf E. A 25-year review of research on feedback in the workplace. *Annu Rev Organ Psychol Organ Behav*. 2025 Jan 21;12(1):19–43. doi: 10.1146/annurev-orgpsych-110622-031927.

[6] Zhou Z, Liu H, Chen S. On assessing the faithfulness of LLM-generated Feedback on student project reports. *Proceedings of the 17th International Conference on Educational Data Mining (EDM 2024); 2024 Jul*, pp. 386–90, Buenos Aires, Argentina.

[7] Wambsganss T, Kueng T, Söllner M, Leimeister JM. ArgueTutor: an adaptive dialog-based learning system for argumentation skills. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24); 2024 May 11–16*. Honolulu, HI, USA.

[8] Alsahafi A, Newell M, Kropmans T. A retrospective feedback analysis of objective structured clinical examination performance of undergraduate medical students. *MedEdPublish*. 2024 Oct 24;14:251.

[9] Ngim CF, Lee SS, Chan YH. Comparison of face-to-face and enhanced written feedback in OSCEs: student perceptions and preferences. *BMC Med Educ*. 2021;21:322. doi: 10.1186/s12909-021-02585-z.