European Journal of Education and Pedagogy
Vol 7 | Issue 2 | March 2026
ISSN 2736-4534

RESEARCH ARTICLE

Balancing Feedback Fidelity and Environmental Cost in Digital Clinical Assessment: A Comparative Pedagogical Analysis of AI-Assisted OSCE Feedback

Thomas Kropmans[1,*], Gabia Neverauskaitė[2], Kylee Fort[3], and Edward Melvin[4]

ABSTRACT

Artificial intelligence (AI) is increasingly embedded in digital assessment systems to support the synthesis of written feedback in higher education. While prior research has examined the reliability and structure of AI-generated feedback, considerably less attention has been paid to the pedagogical trade-offs between feedback fidelity, consistency, and environmental cost. This study investigates how different large language models (LLMs) vary in their capacity to generate structured, behaviourally anchored feedback in Objective Structured Clinical Examinations (OSCEs), and how these differences relate to relative environmental burden inferred from output volume.

Using anonymised OSCE performance data from Qpercom, structured written feedback was generated for 51 stratified student profiles, and outputs were analysed as deployable feedback artefacts by quantifying verbosity and stage stability (internal examiner feedback to student-facing feedback drift and within-student variability), generation-time feasibility, and scenario-based emissions.

The findings demonstrate substantial variation between models in verbosity, structural stability, and consistency across performance levels. One model produced substantially longer outputs with strong structural adherence but a markedly higher estimated environmental footprint, while another delivered more concise feedback with comparable pedagogical alignment and lower inferred emissions. These differences were driven primarily by output volume rather than assumed computational efficiency.

AI-assisted feedback can enhance the structural quality and consistency of assessment narratives, but model selection and output governance materially affect both pedagogical coherence and sustainability. Rather than maximising feedback length, responsible educational deployment of AI requires explicit design constraints that balance fidelity, equity, and environmental considerations.

Keywords: AI-generated feedback, digital assessment, environmental sustainability, Objective Structured Clinical Examination (OSCE).

Submitted: February 04, 2026

Published: March 12, 2026

10.24018/ejedu.2026.7.2.70159

[1] *CEO and R&D, Qpercom Limited, Mervue Business Park, Galway Technology Centre, Ireland.*
[2] *Educational Assessment Analyst, Qpercom Ltd, Galway, Ireland.*
[3] *Account Manager and UX Design, Qpercom Ltd, Aran Island, United Kingdom.*
[4] *Chief Strategy Officer, Marino Software, Dublin, Ireland.*

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

1. Introduction

Written feedback is widely recognised as a critical driver of learning in professional education, yet in high-stakes clinical assessments it is frequently criticised for being generic, delayed, or insufficiently actionable (Alsahafi *et al*., 2023; Alsahafi *et al*., 2024). It is important to note that in Objective Structured Clinical Examinations (OSCEs), examiner comments often fail to link observed behaviours to assessment criteria, which limits the developmental value for learners (Alsahafi *et al*., 2023, 2024; Alsahafi *et al*., 2025). Digital assessment platforms have increasingly turned to artificial intelligence (AI), and specifically large language models (LLMs), as shown in

Table I, to synthesise examiner data into coherent narrative feedback at scale (Campbell *et al*., 2025).

Recent studies suggest that AI-generated feedback (ChatGPT 4.0, Sonet Claude 4) can improve structural consistency, behavioural specificity, and alignment with assessment frameworks when compared with traditional free-text examiner comments (Kropmans *et al*., 2025). At the same time, advances (ChatGPT 5.0) in LLM (Table I) capability have led to increasingly verbose outputs, raising new questions about pedagogical efficiency, equity of feedback provision, and environmental sustainability (Mishra *et al*., 2025). As higher-education institutions commit to carbon-reduction targets, the computational cost of AI-supported educational processes has become an emerging concern (Deda *et al*., 2025).

The present study responds directly to critiques that prior work in this area has conflated feedback quality with educational effectiveness, relied on unvalidated proxies, or over-interpreted environmental estimates. Rather than evaluating learning outcomes or student perceptions, this study adopts a narrower and more defensible focus. It examines AI-generated feedback as a pedagogical artefact, analysing how different models vary in structural fidelity, behavioural anchoring, and output volume, and how these variations translate into relative environmental implications under transparent assumptions.

The guiding research question is therefore: How do different large language models differ in their ability to generate structurally consistent, pedagogically aligned OSCE feedback, and what are the relative environmental implications of these differences?

1.1. Conceptual Framework

This study conceptualises written feedback as a designed educational artefact rather than an instructional intervention (Alsahafi *et al*., 2025). From this perspective, feedback quality is not inferred from downstream learning outcomes but from observable properties of the feedback text itself, including structure, specificity, and developmental clarity. Behaviourally anchored feedback is understood as commentary that explicitly links performance judgements to observable actions or omissions (Fig. 1), thereby supporting learner interpretation and self-regulation (Kropmans *et al*., 2025). Fig. 1 illustrates the station- and competency-level summaries derived from Qpercom's digital scoresheet. The figure functions as an input transparency aid, clarifying the underlying performance data that inform the AI-generated feedback. It does not present learner outcomes; rather, it contextualises the stability analyses by making explicit the performance information to which feedback should remain traceable across processing stages.

Environmental impact is treated not as an educational outcome but as a secondary design constraint (Valls-Val & Bovea, 2021). This study does not claim that AI-generated feedback is inherently sustainable, nor does it attempt to calculate absolute carbon footprints. Instead, it examines how differences in model behaviour—specifically output volume under identical prompts—lead to materially different environmental implications when reasonable and clearly stated assumptions are applied. This framing avoids the false dichotomy between educational quality and

TABLE I: Model Identification and Run Conditions

Label used in paper	Provider	Model name	Variant	Date(s) of runs
GPT-4.0	OpenAI	GPT-4o	Auto	27/07/2025
Claude 4	Anthropic	Claude 4	Sonnet	27/07/2025
GPT-5	OpenAI	GPT-5	Auto	25/11/2025

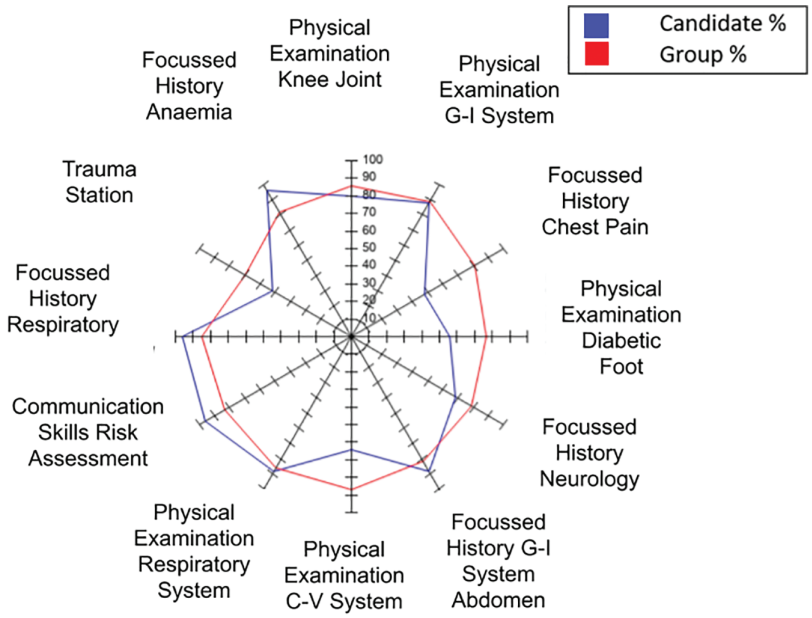

Fig. 1. Station radar plot summaries.

TABLE II: Operational Definitions and Thresholds

Construct	Unit of analysis	Scoring rule	Thresholds
Prompt fidelity ('Rule of 3' completeness)	Feedback text	Required elements present ÷ required elements total	High ≥ 0.90; Moderate 0.70–0.89; Low < 0.70
Hallucination rate	Checkable claims within a text	Unsupported claims ÷ total checkable claims	Minimal 0%–5%; Low 6%–20%; High > 20%
Score alignment	Candidate × stage	Agreement between AI judgement and examiner benchmark	High ≥ 0.90; Moderate 0.70–0.89; Low < 0.70
Actionable targets	Improvement statements	Actionable items ÷ total improvement items	High ≥ 0.80; Moderate 0.50–0.79; Low < 0.50
Stage drift	Candidate × model	Portfolio metric−Preview metric	Report numeric Δ; label only if thresholds pre-set

sustainability, positioning both as considerations within responsible system design.

2. Methods

2.1. Study Design

The study employed a comparative descriptive design combining structured qualitative analysis of feedback texts with quantitative text metrics (Siedlecki, 2020). It does not claim a mixed-methods or outcomes-based evaluation.

2.2. Data Source and Sample

Anonymised OSCE performance data were drawn from Qpercom, a digital assessment platform used in clinical education. From a cohort of 340 candidates, 51 student profiles were purposively selected into three performance strata—high, mid, and low—based on the cut-score and standard error of measurement. Within each stratum, 17 students were randomly selected to form the final study groups. No identifying information was available, and none of the AI-generated feedback analysed in this study was delivered to students.

2.3. Feedback Generation

Structured written feedback was generated for each profile using a fixed prompt, known as the Rule of Three, requiring three strengths, three areas for development, and a global judgement (Faff & Ali, 2023). The same prompt and input data were used across all models to ensure comparability. Feedback was generated in both examiner-facing (preview) and student-facing (portfolio) formats, though the analysis focused on textual properties rather than audience effects (Agrawal & Alvi, 2015).

2.4. Analytical Framework

Analyses focused on measurable output properties (Table II) relevant to implementation governance—word-count distributions, paired examiner-facing feedback and student-facing feedback shifts (including variability), and prompt-format compliance—rather than on educational effectiveness or content validity, which require controller-led validation. Word count was treated strictly as a descriptive characteristic and not as a proxy for educational quality.

2.5. Environmental Estimation

Relative environmental impact was estimated using scenario-based assumptions derived from published ranges for token-level energy use and grid carbon intensity.

Estimates were used solely to compare relative differences between models under identical conditions and are reported as indicative rather than definitive.

Environmental estimates were generated as scenario-based, operational (inference-phase) calculations and are reported transparently as assumptions rather than definitive footprints. Because the LLMs were executed in EU/UK/Ireland cloud regions under an AWS procurement context that includes renewable-energy matching, we report a primary location-based electricity-intensity scenario (reflecting the physical grid mix where computation occurs) and a secondary market-based scenario consistent with renewable matching claims, noting that these approaches address different Scope 2 accounting questions. Grid intensity was parameterised using a central location-based factor (223 gCO_2/kWh) with low/high sensitivity bounds, and paper emissions were modelled using the study's stated A4 sheet emission-factor range. Where token-level logs were not available, output tokens were estimated from word counts using a transparent words-to-tokens conversion with sensitivity bounds. Energy per output token was back-calculated from the study's cohort emissions model under the central location-based grid assumption and then stress-tested across low/high ranges to reflect uncertainty in serving hardware and utilization (Lanka *et al*., 2025). Sensitivity analyses were used to evaluate whether key comparative conclusions (i.e., rank-order differences between models driven by output volume) were robust across plausible parameter sets.

2.6. Ethical Considerations

The study involved secondary analysis of anonymised operational data with no human subjects' interaction. No student-facing decisions or feedback were influenced by the analysis. The role of the platform provider was disclosed, and no claims are made regarding learner benefit or harm.

3. Results

Table III summarises the analytic sample and confirms balanced stratification into high-, mid-, and low-performer

TABLE III: Sample Definition and Feedback Artefacts

Item	Value
Cohort OSCE sitting	340 candidates, 15 stations
Analytic sample	51 profiles (17 top, 17 mid, 17 low performers)
High performers	≥1 SEM above cut-score: 17
Mid performers	within ±1 SEM of cut-score: 17
Low performers	≥1 SEM below cut-score: 17
Preview feedback	Examiner-facing: Generated for all 51 per model
Portfolio feedback	Student-facing: Generated for all 51 per model

TABLE IV: Word Count by Model and Performer Group (Pooled Descriptive Overview)

Performer group	Model	Median	Min–Max
Low	Claude 4	382	304–431
Low	GPT-4.0	302	271–329
Low	GPT-5	2111	969–2701
Mid	Claude 4	389	327–452
Mid	GPT-4.0	300	251–394
Mid	GPT-5	1695	795–2628
Top	Claude 4	401	332–469
Top	GPT-4.0	301	236–354
Top	GPT-5	1067	303–2983

groups (n = 17 per band; N = 51 total), with paired feedback artefacts produced in both examiner-facing (preview) and student-facing (portfolio) feedback formats for each model. This structure enables within-student (paired) stage comparisons and between-model comparisons within the same performance strata.

Table IV presents pooled word-count distributions by model and performer group. GPT-4.0 and Claude-4 outputs remain tightly clustered in the low hundreds of words across all three bands, whereas GPT-5 outputs are substantially longer and more dispersed. Importantly, the magnitude of the GPT-5 ranges indicates strong right-skew and outliers, which limits the interpretability of mean ± SD as a standalone summary and supports the use of robust statistics for GPT-5 in addition to parametric summaries.

To make the examiner-facing and student-facing feedback comparison transparent, Table III reports stage-specific results as paired within-student shifts (examiner-facing and student-facing feedback shift) rather than relying on pooled summaries. In Table V, (mean ± SD), GPT-4.0 and Claude-4 demonstrate small average stage shifts within each performer band, with paired-shift SDs generally in the tens of words, indicating relatively stable output volume across stages at the individual level. GPT-5 shows a different pattern: although the mean stage shift varies by band, the paired-shift SDs are very large (hundreds to >1,000 words) (Table VI), indicating that stage-to-stage changes in output length are highly inconsistent across individuals. In paired testing within bands, there is no evidence of systematic drift for GPT-4.0 or GPT-5, while Claude-4 shows a nominally significant increase in the top-performer band; this effect is small in magnitude and should be interpreted cautiously, particularly in the context of multiple comparisons.

Because GPT-5 output distributions are visibly skewed and outlier-sensitive, Table VII provides robust summaries (Median; Min–Max) for GPT-5 by stage and performer band. These medians show that GPT-5 student-facing feedback tends to be shorter than examiner-facing feedback for high and mid-performers, while low-performers show broadly similar or slightly longer student-facing feedback; however, the wide min–max ranges in both stages and in the paired shifts confirm that extreme contractions and expansions occur within every band. Taken together, the stage results indicate that the primary implementation issue for GPT-5 (Table VIII) is not a consistent direction of drift, but high variance and unpredictability of output volume between examiner-facing and student-facing feedback at the individual level.

Table IX synthesises three governance-relevant output properties—verbosity, consistency, and structural stability—using stage-specific summaries and prompt-compliance checks. Across all performer bands, GPT-5 generated substantially longer feedback than GPT-4.0 and Claude-4, with robust summaries confirming a strongly skewed distribution and wide ranges across both stages.

TABLE V: Stage Drift with Full Mean ± SD (Published Comparator; GPT-4.0 vs. Claude 4)

Model	Preview mean (SD)	Portfolio mean (SD)	Δ words	% change
GPT-4.0	474.3 (55.8)	644.2 (91.4)	+169.9	+35.8%
Claude 4	532.7 (39.3)	630.0 (56.1)	+97.3	+18.3%

TABLE VI: Stage-Specific Word Counts Mean (SD) and Preview→Portfolio Shift by Performer Group

Performer group	Model	Preview mean (SD)	Portfolio mean (SD)
High-performers	GPT-4.0	310 (29)	302 (30)
High-performers	Claude-4	393 (41)	403 (38)
High-performers	GPT-5	1398 (679)	1417 (785)
Mid-performers	GPT-4.0	302 (34)	312 (36)
Mid-performers	Claude-4	394 (32)	386 (29)
Mid-performers	GPT-5	1783 (577)	1457 (472)
Low-performers	GPT-4.0	297 (15)	299 (18)
Low-performers	Claude-4	378 (32)	376 (31)
Low-performers	GPT-5	1890 (589)	1963 (487)

TABLE VII: Robust Summaries for GPT-5 (Median, Min–Max) by Stage and Paired Shift

Performer group	Preview median (Min–Max)	Portfolio median (Min–Max)
High-performers	1070 (756–2983)	920 (303–2898)
Mid-performers	2026 (795–2628)	1305 (903–2176)
Low-performers	2060 (969–2701)	2144 (1176–2534)

TABLE VIII: GPT-5 Paired Shift (Portfolio−Preview), Median (Min–Max)

Performer group	Median shift (Min–Max)
High-performers	−62 (−2680 to +2049)
Mid-performers	−235 (−1670 to +1329)
Low-performers	−161 (−1129 to +1565)

TABLE IX: Verbosity, Structural Stability, and Consistency of AI-Generated Feedback across Stages and Performer Bands

Model	High (Preview/Portfolio)	Mid (Preview/Portfolio)	Low (Preview/Portfolio)
GPT-4.0	310/302	302/312	297/299
Claude-4	393/403	394/386	378/376
GPT-5	1398/1417	1783/1457	1890/1963

Note: *Due to table complexity, data is presented in summary format. See supplementary materials for full detail*. Verbosity (Preview/Portfolio mean word counts by performer band).

TABLE X: Consistency (Paired Shift SD in Words)

Model	High performers	Mid performers	Low performers
GPT-4.0	43.34	53.59	21.65
Claude-4	19.51	35.34	4.48
GPT-5	1140.16	749.91	887.47

TABLE XI: Time-to-Feedback Generation (Per Student; Operational Feasibility)

Model	Estimated generation time per student
GPT-4.0	∼1 minute
Claude 4	∼1.5 minutes
ChatGPT-5	∼3.5 minutes

TABLE XII: Sample-Scale Word Totals and Relative Index (n = 51)

Model	Sample cohort total words (n = 51)	Relative index (GPT-4.0 = 1.00)
GPT-4.0	15,504	1.00
Claude-4	19,669	1.27
GPT-5	87,822	5.66

TABLE XIII: Cohort-Scale AI Emissions Totals (340-Candidate Sitting)

Model	Cohort AI emissions (kg CO_2)
GPT-4.0	5.148
Claude 4	6.468
GPT-5	28.974

Stage-to-stage directional drift (mean examiner-facing and student-facing feedback shift) was generally small, but consistency differed markedly by model: the SD of the paired examiner-facing and student-facing feedback word-count shift (Tables IX and X) remained low for Claude-4 (4.48–35.34 words) and moderate for GPT-4.0 (21.65–53.59 words), whereas GPT-5 exhibited very high within-student variability (749.91–1140.16 words), indicating that individual learners may receive substantially different-length portfolio feedback even under the same prompt structure. Regarding structural stability: the core narrative sections like Intro/Strengths/Development/Conclusion/GRSwere present in exemplar outputs for all models. Prompt-required JSON object with HTML + word-cloud array was not present in retained artefacts.

Table XI reports time-to-feedback generation. Differences in latency align with the observed differences in output volume: GPT-4.0 is fastest, Claude-4 is intermediate, and GPT-5 is slowest. This combination of higher word-count variance and longer generation time is operationally relevant at cohort scale because it influences examiner review burden, platform throughput, and the feasibility of real-time feedback workflows.

Moreover, Table XII links output volume to scaling consequences and summarises relative output totals (normalised index) and Table XIII reports cohort-scale AI emissions estimates under the stated modelling assumptions. The key result is that scaling effects are driven predominantly by output volume: GPT-5's substantially higher word production implies disproportionately higher inference-related emissions compared with GPT-4.0 and Claude-4 under otherwise comparable assumptions. These findings support reporting output governance (e.g., length constraints and structure enforcement) as a practical requirement for responsible deployment, without making claims about differential learning impact.

Table XIV contextualises the AI emissions estimates by presenting paper displacement scenarios for an OSCE sitting. Under an e-scoresheets-only scenario, replacing 5,100 A4 sheets corresponds to an estimated 20.4–30.6 kg CO_2 reduction (min–max using 4–6 g CO_2 per sheet), with a midpoint estimate of 25.5 kg CO_2. Two additional scenarios illustrate that any re-introduction of printing (examiner-facing feedback/student-facing feedback packets) can materially increase paper-associated emissions, emphasising that paper displacement benefits depend on

TABLE XIV: Paper Displacement Emissions (Per OSCE Sitting; Reported Scenarios)

Scenario	Estimated CO_2 (kg) using 4–6 g CO_2 per A4 sheet (min–mid–max)
e-scoresheets only (5,100 sheets replaced)	20.4–25.5–30.6
+ Printed Preview/Portfolio (conservative: 5,780 total sheets)	23.1–28.9–34.7
+ Printed Preview/Portfolio (upper bound: 15,300 total sheets)	61.2–76.5–91.8

TABLE XV: Environmental Parameter Set and Sensitivity Analysis (EU/UK/Ireland; AWS Renewable Matching Applies)

Parameter	Central estimate	Low	High
Grid intensity (gCO_2/kWh), location-based	223	100	452
Electricity accounting, market-based	AWS matched renewables (scenario)	—	—
Tokens per word	1.33	1.20	1.50
Energy per output token (Wh/token)	0.167	0.084	0.251
Serving overhead factor	1.1×	1.0×	1.3×
Paper emissions per A4 sheet (gCO_2e)	5	4	6

Note: The energy-per-token was back-calculated from the study's cohort emissions model under a location-based electricity factor and then stress-tested across low/high sensitivity ranges; conclusions are reported comparatively rather than as absolute footprints.

maintaining digital delivery rather than shifting printing downstream.

Finally, Table XV makes the environmental modelling assumptions explicit and reports the sensitivity bounds used to test robustness of comparative conclusions. Under the central location-based electricity factor (223 gCO_2/kWh) and the stated tokenisation and overhead ranges, the derived energy-per-token estimate (0.167 Wh/token; low–high 0.084–0.251) supports a key interpretation: rank-order differences in estimated AI emissions are driven primarily by output volume, and this conclusion remains stable across plausible parameter sets. Accordingly, the emissions results are presented as comparative scenario estimates, not definitive footprints, with the table providing an auditable basis for interpreting uncertainty.

Overall, the results indicate that between-model differences in output volume and variability are more pronounced than systematic examiner-facing and student-facing feedback drift, with practical implications for feasibility and sustainability at scale.

4. Discussion

This study set out to clarify whether differences between large language models (LLMs) in OSCE feedback generation are primarily a matter of stage drift (Examiner facing (Preview) versus Student facing feedback(Portfolio)), model behaviour (output volume and variability), or both. Taken together, Tables III–XV indicate that the dominant distinction across models is not a consistent Preview→Portfolio contraction or expansion, but rather the scale and predictability of output volume, with downstream consequences for workflow feasibility, fairness, and sustainability.

4.1. Stage Drift is Small on Average, but Variability Matters for Equity

The paired shift analyses show that systematic Preview→Portfolio drift is minimal for GPT-4.0 and Claude-4 across performance bands, with small mean shifts and comparatively modest within-student variability. For GPT-5, mean shifts vary by band, but the most salient feature is the very large paired-shift standard deviation, reflecting substantial individual-level inconsistency in how output length changes between stages. Crucially, this is not merely 'noise' around a stable central tendency: GPT-5's paired shifts (Portfolio−Preview) span from extreme contractions to extreme expansions within every performance band (e.g., high performers: −2680 to +2049 words). Such Min–Max ranges imply that students with comparable performance profiles may receive drastically different feedback volumes under identical prompt conditions and similar performance inputs. In a standardised, high-stakes assessment context, that magnitude of variance constitutes a clear fairness/equity risk, because the 'student experience' becomes inconsistent in ways that may affect perceived legitimacy, usability, and opportunity to act on feedback (Alsahafi *et al.*, 2023; Eva *et al.*, 2004). It strengthens the governance case for explicit length constraints, template enforcement, and human verification at the Preview stage.

4.2. Between-Model Differences in Output Volume are More Consequential than Drift

The pooled word-count distributions and the robust GPT-5 summaries support a clear conclusion: GPT-5 operates in a different output regime from GPT-4.0 and Claude-4. This does not justify claims that GPT-5 produces 'better' feedback—those would require learner outcomes, usability testing, or validated rubric-based content analysis (Campbell *et al.*, 2025). However, it does justify statements about implementation consequences. Longer outputs can increase cognitive load, reduce readability, and shift feedback from actionable guidance toward narrative density; equally, very short outputs risk under-specification (Shakur *et al.*, 2024). In this dataset, GPT-4.0 and Claude-4 appear relatively stable in output volume and stage behaviour, whereas GPT-5 combines verbosity with high dispersion, raising practical questions about standardisation and equity of feedback delivery.

4.3. Operational Feasibility and Governance Implications

Time-to-feedback reinforces the implementation trade-off: increased output volume coincides with increased

generation time. GPT-4.0 is fastest, Claude-4 intermediate, and GPT-5 slowest. In high-stakes OSCE workflows, this affects platform throughput, examiner review time, and the practicality of producing verified, student-facing feedback within expected timeframes. These findings suggest that model selection is not only a technical decision but also a policy decision—requiring explicit articulation of acceptable ranges for output length, acceptable delays, and the degree of permissible variability between students (Alsahafi *et al*., 2024; Kropmans *et al*., 2025).

4.4. Environmental Implications are Driven by Output Volume

The cohort-scale estimates show that sustainability implications in the current modelling approach are driven predominantly by total generated text. Because GPT-5 produces substantially more words, its estimated emissions scale disproportionately under otherwise comparable assumptions. The appropriate interpretation is comparative rather than absolute: model choice and output governance materially influence the environmental burden of AI-supported feedback (Deda *et al*., 2025; Valls-Val & Bovea, 2021). This does not imply that AI feedback is inherently sustainable or unsustainable; rather, the sustainability profile depends on how systems are configured and constrained.

4.5. Reconciling Educational Consequences without Over-Claiming

The results justify cautious educational implications at the level of feedback artefact design, not learning outcomes. Variability in output length and unpredictable stage transitions may influence the usability and perceived fairness of feedback, particularly for borderline or underperforming candidates who are most sensitive to clarity and actionable guidance (Alsahafi *et al*., 2024; Mishra *et al*., 2025). These are plausible consequences that follow from observable artefact properties, motivating the next research step: linking output properties to user-centred endpoints such as learner comprehension, examiner satisfaction, and downstream performance (Dai *et al*., 2023, 2024).

4.6. Limitations

Several limitations shape interpretation. First, the analysis focuses on output characteristics (length, stage shift, variance) rather than content validity or educational impact. Second, GPT-5 distributions are highly skewed; while medians and ranges address representativeness, they cannot replace deeper qualitative evaluation of behavioural anchoring and actionability. Third, carbon estimates depend on modelling assumptions and should be treated as scenario-based indicators rather than definitive footprints (Valls-Val & Bovea, 2021). Finally, results are grounded in one OSCE workflow and may not generalise to other assessment types or disciplines without replication (Siedlecki, 2020).

4.7. Need for Validation Studies and Governance under GDPR

A key implication of the present findings is the need for formal validation studies before AI-generated OSCE feedback can be interpreted as educationally meaningful or safely deployed at scale. The current analysis intentionally focuses on observable artefact properties—output length, stage stability, variability, and modelling implications—because these are measurable within an anonymised processing context (Shakur *et al*., 2024). However, educational validity cannot be inferred from artefact properties alone.

Whether feedback is useful depends on alignment with station intent, fidelity to performance evidence, interpretability by learners, and its capacity to support actionable improvement (Dai *et al*., 2023). Particularly in high-stakes contexts, institutions require evidence that AI-generated narrative feedback is not merely well-formed text but a valid representation of performance and a safe basis for learner interpretation (Dai *et al*., 2024; Shaw & Crisp, 2015). This creates a clear research agenda: moving from descriptive output evaluation to validity arguments grounded in content, response process, internal structure, and consequences (Dai *et al*., 2023).

Critically, the ability to conduct such validation work is constrained by governance. As a data processor under the GDPR, our role is limited to processing assessment data on behalf of the data controller (typically the university or medical school) (Chico, 2018; Data Protection Commission, 2019; Gupta, *et al*., 2024). This limits our capacity to independently conduct studies requiring linkage between AI outputs and identifiable performance evidence, access to protected examiner records, or collection of participant perceptions. In practice, validation requires controller-held permissions for secondary use of assessment data, clarity on lawful basis, ethics review/waiver where appropriate, and agreements specifying purpose limitation, data minimisation, retention, and audit (Rumbold & Pierscionek, 2017). For this reason, rigorous validation is not simply a methodological preference but a collaborative requirement.

A feasible validation programme can be staged and proportionate. First, institutions can run content- and fidelity-validation studies using a sampled subset of OSCE stations and candidates, with independent expert raters evaluating whether AI statements are traceable to underlying station checklists, examiner comments, and scoring rubrics (Shaw & Crisp, 2015). Second, response-process validation can be conducted through structured interviews or surveys with examiners and students, focusing on comprehension, perceived fairness, usability, and whether the feedback prompts appropriate reflection and action planning (Dai *et al*., 2023, 2024). Third, consequential validation can be pursued through low-risk implementation pilots in which feedback is delivered in a controlled manner and outcomes such as learner satisfaction, action-plan quality, and subsequent performance indicators are monitored with appropriate safeguards (Briñol & Petty, 2022). Across all stages, the controller institution maintains governance by approving protocols, managing notice/consent as appropriate, and ensuring AI outputs are treated as

supportive artefacts rather than autonomous judgements (Rumbold & Pierscionek, 2017).

Accordingly, the central implication of this study is not that one model is 'better', but that deployment decisions should be contingent on controller-led validation evidence, with output governance and verification embedded as standard safeguards (Dai *et al*., 2023; Shakur *et al*., 2024).

4.8. Implications for Practice and Research

For practice, the data support four actionable recommendations: (1) retain a Preview verification layer for any high-stakes portfolio release, (2) implement output constraints (length caps and structure checks) to reduce variance and maintain comparability, and (3) include sustainability considerations explicitly in procurement and deployment decisions (Valls-Val & Bovea, 2021; Deda *et al*., 2025). For research, (4) the priority is to move from output metrics to validated quality indicators (behavioural anchoring, alignment with station objectives) and to test whether constrained outputs improve usability without eroding developmental value (Dai *et al*., 2024; Dai *et al*., 2023; Shaw & Crisp, 2015).

The observed Preview→Portfolio divergence highlights a governance risk that is not primarily statistical but procedural: students may receive a different narrative than the one reviewed and approved by the examiner. Even where average stage drift is small, the within-student variability reported implies that portfolio outputs can change materially between stages for some candidates. In high-stakes assessment contexts, this creates a misalignment between accountability (examiner sign-off) and exposure (what the student ultimately reads), undermining the notion of a single authoritative feedback record (Dai *et al*., 2024; Rumbold & Pierscionek, 2017).

These findings support a roadmap change in which AI feedback is generated at the source of assessment—immediately after the examiner completes the scoresheet—so that the examiner edits, verifies, and approves the feedback once, within the same workflow context. Once approved, the feedback should be 'sealed' as the record of assessment, stored with an immutable audit trail (timestamp, model identifier, prompt version, and any examiner edits) (Data Protection Commission, 2019; Rumbold & Pierscionek, 2017). More broadly, the results argue for human-controlled, single-pass generation rather than repeated embedded LLM calls across multiple software layers and time points. Where multi-stage outputs are required, the transformation should be deterministic and governed (template-locked, length-bounded, auditable), or else the portfolio artefact should be derived from the examiner-approved source text rather than regenerated (Dai *et al*., 2023).

5. Conclusions

Across 51 stratified OSCE profiles, between-model differences in output volume and variability were more pronounced than systematic Preview→Portfolio drift. GPT-4.0 and Claude-4 generated relatively stable-length feedback with minimal average stage shift, while GPT-5 produced substantially longer outputs and markedly higher individual-level variability in stage-to-stage length change.

Crucially, the extreme Min–Max variability observed in GPT-5 stage-to-stage shifts—including changes spanning several thousand words—should be interpreted as a fairness and equity risk: variability of that magnitude threatens a standardised student experience even when prompts and performance inputs are held constant (Alsahafi *et al*., 2025; Eva *et al*., 2004). Under the stated scenario assumptions, cohort-scale environmental estimates were driven primarily by total generated text, indicating that model selection and output governance are central levers for responsible deployment (Deda *et al*., 2025; Valls-Val & Bovea, 2021).

Future work should evaluate whether constrained, behaviourally anchored outputs improve learner usability and perceived fairness, and should link these artefact-level properties to educational outcomes before claims of effectiveness are made (Dai *et al*., 2024; Dai *et al*., 2023). Deployment decisions should remain contingent on controller-led validation evidence under GDPR-aligned governance arrangements (Shaw & Crisp, 2015; Data Protection Commission, 2019; Rumbold & Pierscionek, 2017).

Acknowledgment

The authors acknowledge the institutions using Qpercom for their ongoing collaboration and support in improving digital assessment practices.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AI-Assisted Writing Disclosure

OpenAI ChatGPT-5 was used to support drafting, restructuring, and language refinement of this manuscript. The tool did not generate or analyse primary data. The authors remain fully responsible for the study design, analyses, results, and all scientific claims, and they accept full accountability for the accuracy, originality, and integrity of the manuscript.

Conflict of Interest

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

References

Agrawal, P. K., & Alvi, A. S. (2015). Textual feedback analysis: Review. *2015 International Conference on Computing Communication Control and Automation*, pp. 457–460. https://ieeexplore.ieee.org/abstract/document/7155888.

Alsahafi, A., Ling, D. L. X., Newell, M., & Kropmans, T. (2023). A systematic review of effective quality feedback measurement tools used in clinical skills assessment. *MedEdPublish*, *12*, 11.

Alsahafi, A., Newell, M., & Kropmans, T. (2024). A retrospective feedback analysis of objective structured clinical examination performance of undergraduate medical students. *MedEdPublish*, *14*, 251.

Alsahafi, A., Newell, J., Newell, M., & Kropmans, T. (2025). A comparative analysis of objective structured clinical examination (OSCE) observed scores and global rating scores using a novel approach. *European Journal of Education and Pedagogy*, *6*(5), 1–8.

Briñol, P., & Petty, R. E. (2022). Self-validation theory: An integrative framework for understanding when thoughts become consequential. *Psychological Review*, *129*(2), 340–367.

Campbell, K. K., Holcomb, M. J., Vedovato, S., Young, L., Danuser, G., Dalton, T. O., Holcomb, A., Blankenship, R. B., Maloney, C. G., Shakur, A. H., Schwartz, A. (2025). Applying state-of-the-art artificial intelligence to grading in simulation-based education: Assessment, feedback, and ROI. *Discover Artificial Intelligence*, *5*(1), 202.

Chico, V. (2018). The impact of the general data protection regulation on health research. *British Medical Bulletin*, *128*(1), 109–118.

Dai, D. W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pp. 323–325. https://ieeexplore.ieee.org/abstract/document/10260740.

Dai, D. W., Vu, T., Knoch, U., Lim, A. S., Malone, D. T., & Mak, V. (2024). Expanding Kane's argument-based validity framework: What can validation practices in language assessment offer health professions education? *Medical Education*, *58*(12), 1462–1468.

Data Protection Commission. (2019). *Guidance on Anonymisation and Pseudonymisation*. Data Protection Commision Ireland. https://www.dataprotection.ie/en/dpc-guidance/anonymisation-pseudonymisation.

Deda, D., Gervasio, H., & Quina, M. J. (2025). Advancing carbon footprint in higher education: An integrated assessment model. *International Journal of Life Cycle Assessment*, *30*(9), 1930–1943.

Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (2004). An admissions OSCE: The multiple mini-interview. *Medical Education*, *38*(3), 314–326.

Faff, R. W., & Ali, S. (2023). *Rule of 3 … Matters!* Social Science Research Network. https://papers.ssrn.com/abstract=4605763.

Gupta, I., Philip, S. S., & Naithani, P. (2024). Duties and responsibilities of controller and processor. In I. Gupta, S. S. Philip, P. Naithani (Eds.), *Introduction to data protection law: Cases and materials from the EU* (pp. 149–186). Springer Nature. https://doi.org/10.1007/978-981-97-6355-9_4.

Kropmans, T., Bilokrylyi, O., Predchyshyn, D., Cunningham, D., Melvin, E., & Neverauskaitė, G. (2025). Comparing AI-generated preview and portfolio feedback: GPT 4.o vs. Claude 4. *European Journal of Artificial Intelligence and Machine Learning*, *4*(5), 26–32.

Lanka, S., Cabezuelo, A. S., & Vuppalapati, C. (2025). *Trends in Sustainable Computing and Machine Intelligence: Proceedings of ICTSM 2024*. Springer Nature.

Mishra, V., Lurie, Y., & Mark, S. (2025). Accuracy of LLMs in medical education: Evidence from a concordance test with medical teacher. *BMC Medical Education*, *25*(1), 443.

Rumbold, J. M. M., & Pierscionek, B. (2017). The effect of the general data protection regulation on medical research. *Journal of Medical Internet Research*, *19*(2), e7108.

Shakur, A. H., Holcomb, M. J., Hein, D., Kang, S., Dalton, T. O., Campbell, K. K., Blankenship, R. B., Maloney, C. G., Young, L., Schwartz, A., Vedovato, S. (2024). *Large language models for medical OSCE assessment: A novel approach to transcript analysis*. arXiv. http://arxiv.org/abs/2410.12858.

Shaw, S., & Crisp, V. (2015). Reflections on a framework for validation-Five years on. https://www.repository.cam.ac.uk/handle/1810/354544.

Siedlecki, S. L. (2020). Understanding descriptive research designs and methods. *Clinical Nurse Specialist*, *34*(1), 8.

Valls-Val, K., & Bovea, M. D. (2021). Carbon footprint in higher education institutions: A literature review and prospects for future research. *Clean Technologies and Environmental Policy*, *23*(9), 2523–2542.