ASCILITE 2026

The Future is Connection:

Collaboration, Creativity and Compassion in a Digital Society

Classification AI vs Generative AI in grading and feedback

Manjinder Kainth
Inspera AS

Marking is among the most time-consuming activities in education, and artificial intelligence (AI) is widely proposed as a remedy. Attention has centred on generative large language models, yet regulators raise concerns about their bias, consistency, and explainability. This paper contrasts generative grading with per-question classification models trained only on an educator's own marking and deployed in a human-in-the-loop workflow. The architectures are compared against criteria drawn from regulatory guidance: transparency, construct-relevant features, determinism, and specifiable training data. A narrower role for generative models is also identified, in composing and restructuring marker-approved feedback rather than in marking. Deployment analytics from one platform are then reported, covering approximately 166,000 marked questions, 39,500 essays, and six institutions: 66.2% of questions were marked in under one minute, 77.6% to 89.5% of AI suggestions were accepted, and essay models reached a quadratic weighted kappa of 0.93 for high-confidence predictions. Questionnaire responses from earlier pilots (33 markers, 30 students) are also reported: markers described marking as faster and consistent feedback as easier to provide; most students rated feedback more consistent and higher in quality. Implications for institutions selecting or building such tools are discussed, alongside the limitations of vendor-collected data.

Keywords: automated marking, feedback, classification AI, generative AI, large language models, human-in-the-loop, user experience

Introduction

Marking and feedback sit at the centre of teaching, and they are expensive. Educators routinely describe assessment as one of the most repetitive parts of their work, and feedback loses much of its value to students when it arrives weeks after submission. It is therefore unsurprising that artificial intelligence (AI) has been proposed as a way to reduce marking workload while protecting, or even improving, the quality and consistency of feedback.

Since the release of ChatGPT in late 2022, most of this conversation has focused on generative large language models (LLMs). The evidence on generative grading is mixed. Flodén (2025) compared ChatGPT with human graders on 463 higher education exam responses and reported exact agreement of around 30%, alongside differences in grade distributions and limited replicability across repeated runs. Warr et al. (2025) found that an LLM assigned different scores to identical student writing when demographic descriptors were varied, providing experimental evidence of implicit bias. Other studies are more encouraging: Yavuz et al. (2025) reported high inter-rater reliability between LLMs and experienced instructors on rubric-based essay grading, and systematic reviews suggest performance is strongest on closed and short-answer tasks and weakest on open, subjective work (Yan et al., 2024). The picture that emerges is of a technology whose grading behaviour depends heavily on the task, the rubric, and the prompt, and whose failure modes (bias, inconsistency, and opacity) are difficult to audit.

Regulators have responded with caution. In England, Ofqual (2024) has adopted a precautionary principle and has been explicit that using AI as the sole marker of student work does not comply with its regulations, citing the requirement for human judgement and known risks around bias, inaccuracy, and transparency. The European Union's Artificial Intelligence Act classifies AI systems used to evaluate learning outcomes as high-risk, attracting obligations around transparency, data governance, and human oversight (European Union, 2024). UNESCO's guidance similarly calls for a human-centred approach and for institutions to validate generative tools before use (Miao & Holmes, 2023).

"AI" is not, however, a single architecture. Supervised classification models trained on human-marked responses predate LLMs by many years (Bai & Stede, 2022) and have quite different technical properties: their training data can be fully specified, their outputs are deterministic, and their decision features can be constrained to the construct being assessed. What has been missing from the debate is evidence at deployment scale, since most published evaluations of AI grading are small laboratory studies. This paper addresses that gap in four steps. First, it compares classification and generative architectures against criteria derived from regulatory guidance, and identifies where generative models can contribute without taking marking decisions. Second, it reports usage analytics from a commercial classification-based marking platform spanning approximately 166,000 marked questions, 39,500 essays, and six institutions. Third, it reports marker and student questionnaire responses from earlier institutional pilots of the same platform in order to set the efficiency metrics alongside the experience of the people using them. Fourth, it discusses what these findings mean for institutions selecting or building AI marking tools under emerging regulation. Four research questions guide the paper: (a) How do the architectures differ against regulatory criteria, and to what extent are those differences properties of model family rather than of design? (b) What do deployment data show about marking efficiency, feedback reuse, and accuracy? (c) Is model confidence a valid signal for routing human review? And, (d) how do markers and students describe the feedback produced through such a workflow?

Background

Approaches to automated marking

Automated marking systems fall into several families: rule-based scoring; supervised machine learning (classification) models trained on human-marked scripts; deep neural network scoring; fine-tuned LLMs; and generative AI systems prompted with a mark scheme (Bai & Stede, 2022). Each involves a different trade-off between data requirements, transparency, and coverage. Rule-based systems are fully transparent but brittle. Neural approaches generalise well but typically require large training corpora and resist inspection. Fine-tuned LLMs and prompted generative systems require little or no local training data, which explains their appeal, but their behaviour is shaped by pre-training corpora that cannot be fully specified or audited by the deploying institution.

The classification approach examined in this paper builds a separate, bespoke model for each assessment question, trained only on how the educator marks responses to that question. No external corpus is involved. This narrows what the model can learn to the marking behaviour of the responsible educators, which has two consequences. First, the sources of potential bias are confined to the human marking itself, which institutions already moderate and quality-assure. Second, the model's decision features can be limited to construct-relevant characteristics of the response, rather than whatever correlations exist in web-scale text.

Comparing the architectures

Table 1 below summarises the properties of the main architectures against criteria that recur in regulatory guidance: transparency to subject experts, reliance on construct-relevant features, determinism (the same output for the same response), specifiable training data, data volume required, and sources of bias (European Union, 2024; Ofqual, 2024). The final column describes generative systems constrained by design, whose properties depend more on implementation choices than on model family; the boundaries it draws reflect present capability rather than a permanent settlement, since quality assurance frameworks, model documentation standards, and regulatory specifications are all developing alongside the technology. Two points deserve emphasis. Determinism matters for appeals and moderation: a grade that changes between runs is difficult to defend to a student or an external examiner, and lack of replicability has been observed empirically in LLM grading (Flodén, 2025). Specifiable training data matters for bias auditing: a model trained only on local marking can inherit local biases, but those biases are visible and correctable through existing moderation processes, whereas biases inherited from pre-training corpora are not (Warr et al., 2025).

Table 1.
Properties of automated marking architectures

Property	Per-question classification	Neural network scoring	Fine-tuned LLM	Generative AI with mark scheme
Fully transparent to subject experts	Yes	No	No	No
Decisions based on construct-relevant features	Controlled by designers	No	No	No
Same output for the same response	Yes	Yes	No	No
Fully specifiable training data	Yes	Yes	No	No
Training data volume required	Low	Large	Low	None
Main sources of bias	Local human marking	Human marking corpus	Pre-training data and human marking	Pre-training data

This comparison should not be read as one-sided. Classification models cannot grade a question they have never seen marked; they require an initial period of human marking for every new question, and they do not remove bias so much as constrain where it can come from. Generative systems, by contrast, can produce a plausible grade for almost any prompt immediately, which is precisely why their weaker guarantees around consistency and explainability matter in regulated assessment. Nor does it follow that generative models have no place. A more promising division of labour applies to them after the marking decision rather than to it: once a marker has confirmed a grade and selected the feedback that applies, a language model can compose and restructure that already-approved content for the individual student, adjusting register, ordering, and level of detail without altering the judgement or the substance of what was said. On this reading the generative component is a presentation layer over human-authored feedback, and the properties in Table 1 that matter for marking are not put at risk, because no marking decision is delegated to it. The remainder of this paper reports deployment evidence for one classification-based implementation.

A classification-based, human-in-the-loop platform

The platform studied here is Inspera Graide, a commercial marking and feedback system developed by the author's organisation, Inspera AS. The author is employed by that organisation and therefore has a commercial interest in the system evaluated; the implications of this are addressed in the limitations. It supports mathematics, short-text, and essay responses, including handwritten work via optical character recognition. For each question, a classification model is trained on the educator's marking decisions for that question. Once trained, the model suggests a grade and feedback for each new response, together with a confidence score.

Figure 1 below shows the workflow. Every suggestion passes through human review before anything is released to a student. Suggestions at or above a confidence threshold are routed to a fast review, in which the marker confirms or amends the suggestion; suggestions below the threshold receive a detailed review. Accepted, edited, and overridden decisions feed back into training, so the model updates as marking proceeds. Feedback items written by markers are stored in a reusable bank, which is the mechanism behind the feedback reuse results reported in the next section.

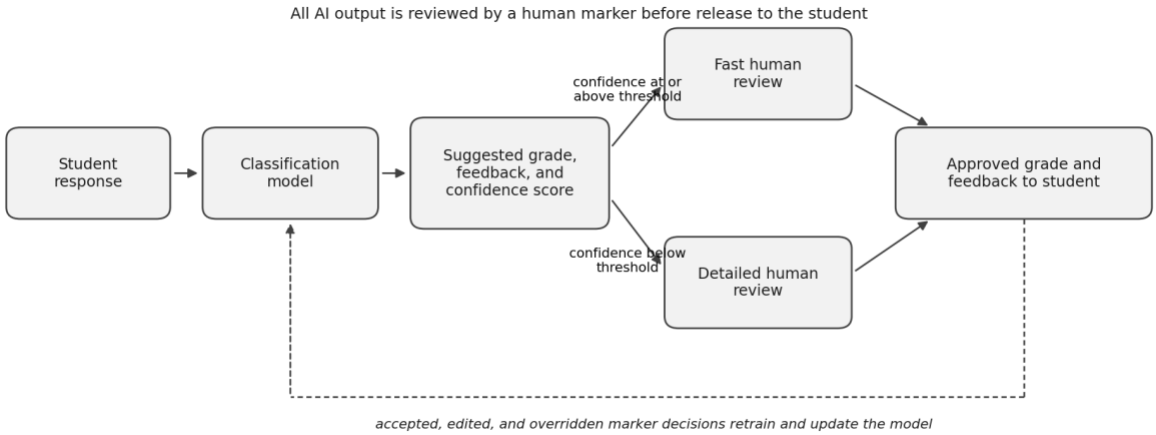


Figure 1. Human-in-the-loop marking workflow. All AI suggestions are reviewed by a marker before release, and marker decisions retrain the model.

Method

The study is a retrospective analysis of anonymised usage analytics from six customer institutions: two universities, an apprenticeship delivery provider, two vocational training providers, and an online college (labelled University A, University B, Apprenticeship Provider A, Vocational Provider A, Vocational Provider B, and Online College A). The dataset covers approximately 166,000 marked questions across more than 15,000 assignments, 196,000 unique feedback items, and a wide range of disciplines, question types, and difficulty levels. Essay model accuracy was evaluated separately on 39,500 essays across 144 rubrics. Confidence calibration was analysed on 15,400 AI suggestions from four of the six institutions.

A second, earlier dataset provides the marker and student perspective. During pilot deployments at three UK universities in 2022 and 2023, participating markers and students were invited to complete a questionnaire about their experience of marking, or of receiving feedback, through the platform. Responses were obtained from 33 markers (31 at one university in early 2022, one each at two further universities in early 2023) and from 30 students at one of those universities. The instrument combined comparative items rating the platform against the systems participants normally used, five-point agreement items on usability, and free-text comments. These pilots preceded the deployment period covered above and do not draw on the same institutions or cohorts, so the responses are reported as indicative context on user experience rather than as a matched evaluation.

Five groups of measures are reported, each defined in non-technical terms at first use:

- marking time: the distribution of time taken to mark individual questions and whole assignments;
- feedback reuse: the mean number of times each unique feedback item is applied to student work;
- suggestion acceptance: the proportion of AI suggestions accepted by markers, overall and by confidence band, where a confidence score is the model's own estimate, expressed as a percentage, of how likely its suggestion is to be correct, and a confidence band groups suggestions by that score;
- essay accuracy: exact accuracy, meaning the percentage of essays to which the model assigned the same rubric band as the human marker; adjacent accuracy, meaning the percentage placed either in the same band or in one immediately either side of it; and quadratic weighted kappa (QWK), a measure of agreement running from 0 to 1 that gives partial credit for near misses and discounts the agreement that would be expected by chance alone. All three are calculated against human grades and reported by number of rubric bands, following standard practice in automated scoring evaluation (Williamson et al., 2012). Results are reported for the full model and for the high-confidence subset of predictions;
- marker and student perceptions: the distribution of questionnaire responses on feedback consistency, feedback quality, marking or turnaround speed, usability, and intention to continue using the platform.

Two design limitations should be stated at the outset. These are operational data from live deployments, not a controlled experiment: institutions self-selected into using the platform, and no counterfactual condition exists. Second, suggestion acceptance is a behavioural proxy for accuracy; markers can, in principle, accept an incorrect suggestion. Both limitations, and the vendor's conflict of interest, are returned to in the discussion.

Ethics

The two datasets rest on different ethical bases. The platform analytics are aggregate operational data generated in the ordinary course of each institution's use of the system. They contain no student or staff identifiers, no response content, and no field permitting re-identification, and were processed under the data processing agreements already in place between each institution and the supplier. No separate ethical approval was sought for this retrospective analysis, which involves no human participants, no intervention, and no personal data.

The pilot questionnaires involved human participants and are different in kind. Each pilot was approved by the host institution as part of the contracted pilot process, and markers and students opted in both to the pilot and to the questionnaire, having been informed of the purpose of the study and how their responses would be used. No participant received payment or any other incentive. Responses are reported only in aggregate: institutions are not named, no free-text comment is attributed to an identifiable individual, and the name and email fields collected by the questionnaire tool were excluded from analysis. One of the three universities is nonetheless identifiable from the title of the earlier efficacy study cited below, which was conducted there; no questionnaire response is attributed to that university or to any other. The response rate cannot be established retrospectively, which is noted as a limitation.

Results

Marking time

Table 2 below shows the distribution of marking times. Across more than 100,000 questions, 66.2% were marked in under one minute, and 79.4% in under two minutes. At assignment level, 48.8% of more than 15,000 assignments were marked in under five minutes. Earlier efficacy research on the same platform, reported in a company white paper rather than in a peer-reviewed venue, compared paper-based marking with the platform across 10 questions and 172 submissions per question at one university, and found that median per-script marking time fell from 11.2 minutes to 2.8 minutes, a 74% reduction, while the volume of written feedback increased by a factor of 7.2 (Kainth, 2022). These figures cannot be attributed to the AI component alone; the marking interface, feedback bank, and automatic mark collation all contribute.

Table 2.
Distribution of marking times for questions (n > 100,000) and assignments (n > 15,000)

Question marking time	Share of questions (%)	Assignment marking time	Share of assignments (%)
Under 1 minute	66.2	Under 5 minutes	48.8
1–2 minutes	13.2	5–10 minutes	17.4
2–5 minutes	11.4	10–20 minutes	21.6
5–10 minutes	5.3	Over 20 minutes	11.7
Over 10 minutes	4.0		

Feedback reuse

Markers created 196,000 unique feedback items across the six institutions. The mean number of times each item was reused varied considerably: 8.5 at University A, 7.9 at University B, 3.8 at Apprenticeship Provider A, 2.6 at Vocational Provider A, 1.9 at Vocational Provider B, and 1.41 at Online College A. Reuse was highest where cohorts are large and submissions repetitive, which is consistent with the mechanism: a comment written carefully once can be applied to every subsequent response that warrants it. Reuse changes the economics of detailed feedback, since the cost of writing a thorough comment is amortised across many

students, although the analytics cannot show whether reused comments were always the most appropriate ones.

Suggestion acceptance and confidence calibration

Across the four institutions with suggestion-level data, markers accepted between 77.6% and 89.5% of AI suggestions (Table 3 below). The share of submissions in which the AI contributed varied more widely, from 27.0% at Online College A to 69.3% at Apprenticeship Provider A, reflecting differences in question types and in how long each deployment had been running. Notably, the institution with the highest acceptance rate was not the one with the highest AI involvement, which suggests markers were exercising judgement rather than accepting by default.

Table 3.
Acceptance of AI suggestions and AI involvement in marking, by institution

Institution	Suggestions accepted (%)	Submissions partly AI-assisted (%)	Submissions fully from accepted AI suggestions (%)
University A	82.3	33.3	9.7
Apprenticeship Provider A	87.2	13.8	55.5
Vocational Provider A	89.5	8.8	28.0
Online College A	77.6	7.9	19.1

For risk-based human review to be defensible, the model's confidence score must actually predict whether markers will agree with it. Figure 2 below plots acceptance against confidence band for 15,400 suggestions. Acceptance rose from approximately 58% in the lowest band (0–20% confidence) to approximately 91% in the highest (80–100%). The relationship was not perfectly monotonic: the 60–80% band, which contained only 479 suggestions, dipped below the 40–60% band. Overall, however, the pattern supports using confidence to decide how much human scrutiny a suggestion receives, with low-confidence suggestions receiving detailed review.

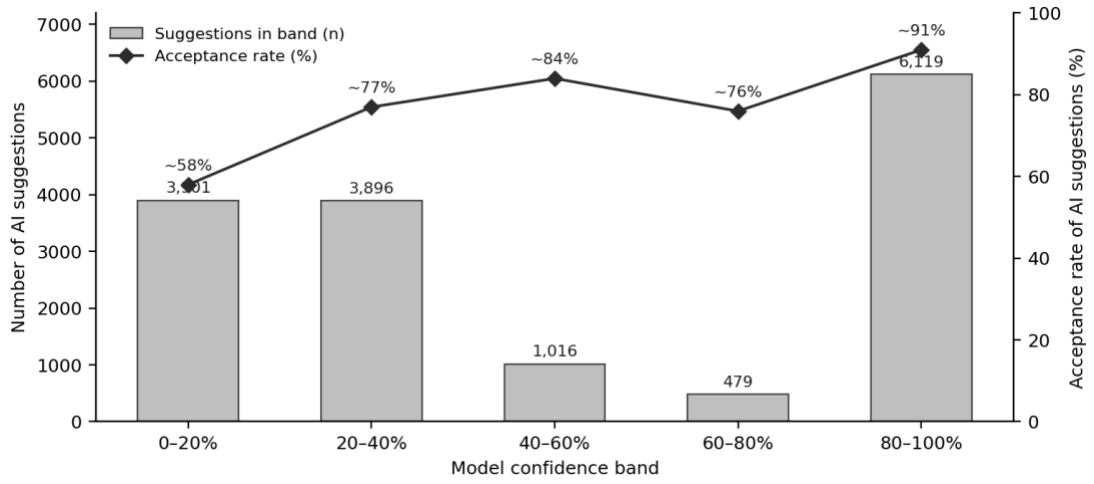


Figure 2. Acceptance of AI suggestions by model confidence band (15,400 suggestions, four institutions). Acceptance rates are approximate; suggestion counts are exact.

Essay model accuracy

Essay grading was evaluated on 39,500 essays across 144 rubrics (Table 4 below). Following Williamson et al. (2012), exact accuracy is most informative for rubrics with few bands, adjacent accuracy for medium rubrics, and QWK for larger ones. On two-band rubrics the full model achieved 89% exact accuracy. As the number of bands increases, exact accuracy falls, as expected, but restricting to high-confidence predictions recovers much of it: 80% on four-band and 83% on five-band rubrics. Adjacent accuracy remained between 84% and

94% for the full model. On the largest rubrics, QWK reached 0.76 to 0.81 for the full model and 0.85 to 0.93 for high-confidence predictions. Two cells run against the general pattern: on two-band rubrics and on adjacent accuracy for six- to seven-band rubrics, the full model scored higher than the high-confidence subset. The data are reported as observed.

Table 4.
Essay model accuracy by number of rubric bands (39,500 essays, 144 rubrics)

Rubric bands	Exact accuracy, full model (%)	Exact accuracy, high confidence (%)	Adjacent accuracy, full model (%)	Adjacent accuracy, high confidence (%)	QWK, full model	QWK, high confidence
2	89	80	–	–	–	–
3	63	73	–	–	–	–
4	54	80	94	95	–	–
5	46	83	88	93	–	–
6–7	–	–	84	80	0.76	0.93
8–9	–	–	–	–	0.81	0.85

Note. QWK (quadratic weighted kappa) indicates the metric was not the primary measure for that rubric size and was not reported.

Marker and student perspectives

The questionnaire data address a different question: not how quickly marking was completed, but how the people involved described it. Among the 33 markers, the strongest results concerned speed and consistency. In the largest group (*n* = 31), 90.3% reported marking faster than with the systems they normally used, 51.6% significantly faster, and none reported it slower. Consistent feedback was rated easier to provide by 80.6% and high-quality feedback easier by 71.0%, although 9.7% and 19.4% respectively found those tasks slightly harder. Every respondent rated becoming familiar with the marking interface at four or five on a five-point scale, and 93.5% wanted to continue using the platform. The two markers at the other universities responded in the same direction.

Students (*n* = 30) were more measured but broadly positive. Twenty of the 30 rated their feedback more consistent than that from the systems they normally used, and the same number rated it higher in quality; 18 found it easier to receive and attend to, 28 agreed it was easy to view, and 25 wanted to continue.

Discussion

Interpreting the findings

Taken together, the analytics describe a system in which efficiency gains come from three reinforcing mechanisms rather than from removing the human: a marking interface that reduces friction, a feedback bank that amortises the cost of writing good comments, and AI suggestions whose reliability the marker can gauge from a calibrated confidence score. The acceptance results are worth dwelling on. If markers accepted everything, acceptance rates would tell us little; the facts that acceptance varies by institution, tracks confidence, and sits well below 100% even in the highest band all indicate active human judgement. That is the behaviour a human-in-the-loop design is supposed to produce.

The essay results illustrate both the promise and the boundary of the classification approach. Where rubrics are coarse, or where the model is confident, agreement with human grading is high. Where rubrics are fine-grained and confidence is low, exact agreement drops to 46%, and the correct system response is not to trust the model but to route the work to detailed human review. This is a different philosophy from generative grading in its unconstrained form, which produces an answer regardless and offers no comparable, validated signal of when it should not be trusted. The gap is not necessarily permanent: a constrained implementation that exposed calibrated uncertainty and routed low-confidence cases to human review would narrow it

considerably, and that is one reason for treating confidence calibration, rather than model family, as the property worth specifying in procurement and in institutional policy.

The earlier paper-versus-platform comparison points the same way on a dimension the analytics cannot reach. In that study the volume of written feedback students received rose by a factor of 7.2 as median marking time fell (Kainth, 2022): a claim about what students receive rather than what markers save. Efficiency and feedback quality are often treated as competing objectives; here the released time is reinvested in the feedback itself, and it is that reinvestment which carries the educational value.

The questionnaire responses matter in a way the analytics cannot. Efficiency metrics establish that the workflow is fast; they say nothing about whether the people inside it experience marking as improved, or feedback as more useful. On the marker side the two sources converge: markers described marking as faster and consistent feedback as easier to provide, which is what the timing and reuse data would predict. On the student side they diverge, because students did not consistently perceive faster turnaround even where marking time fell. Time released from marking reaches students only through release schedules, moderation windows, and internal deadlines that no marking tool controls.

This is also where the argument connects to the conference theme. A human-in-the-loop design is a claim about connection rather than automation: it keeps the educator as author of the judgement and of the comment the student reads, and uses the machine to absorb the repetition that erodes time for that relationship. The feedback bank is the clearest illustration, since a comment written carefully once, by someone who knows the cohort, can reach every student whose work warrants it. Compassion in assessment is hard to sustain at scale when marking is exhausting; the case for this class of tool is that it makes attentive feedback affordable, not that it makes the marker unnecessary.

Regulatory alignment

The architecture maps closely onto current regulatory expectations. Ofqual (2024) requires human judgement in marking decisions; here, no grade reaches a student without marker review. The EU AI Act's obligations for high-risk systems include human oversight, transparency, and data governance (European Union, 2024); a deterministic model trained on fully specifiable local data is straightforwardly auditable in a way that a system built on undisclosed pre-training corpora is not. Ofqual's guidance also distinguishes adaptive, autonomous AI from technical systems whose decision criteria are set by human experts, and per-question classification sits far closer to the latter category than generative grading does. None of this makes the approach automatically compliant; it makes compliance demonstrable, which for awarding organisations is the practical requirement. Nor does it exclude generative models from the workflow altogether. A generative component confined to composing and restructuring feedback a marker has already approved does not engage the marking obligations at all, because no grade or judgement is delegated to it. What matters for compliance is where in the workflow the model sits, and how much of the necessary evidence the deploying institution is able to obtain from its supplier.

Limitations and future work

The limitations are substantial and should temper the conclusions. The data are vendor-collected, from institutions that chose to adopt the platform, and the author has a commercial interest in the system studied. There is no randomised comparison, either against unassisted marking or against a generative system on the same responses; the contrast with generative grading in this paper is drawn from the published literature rather than from a head-to-head evaluation. Acceptance is a proxy that could be inflated by automation bias, in which markers accept suggestions they should have checked, and the acceptance rates in Figure 2 are approximate. Constraining bias sources to local human marking does not eliminate bias, and no demographic fairness audit was performed on this dataset. The questionnaire data carry their own constraints. They were collected in 2022 and 2023, before the deployment period covered by the analytics, at three universities that do not correspond to the six in the analytics dataset, so the two sources describe overlapping but distinct populations. Marker samples are very small outside the largest group, with one respondent at each of the two remaining universities, and the student data come from a single cohort. Participants were volunteers in supplier-supported pilots completing a questionnaire administered by the supplier, which invites both self-

selection and social desirability effects, and response rates cannot now be reconstructed. No inferential statistics are reported, and no claim is made that these responses generalise; what they establish is narrower, that in these pilots the direction of marker and student opinion was consistent with the efficiency findings on every dimension except turnaround time.

Future work should include independent evaluation on shared datasets, controlled comparisons between classification, constrained generative, and hybrid grading on the same responses, demographic audits of both the human marking and the models trained on it, contemporaneous and independently administered studies of marker and student experience, and studies of how confidence displays affect marker behaviour.

Conclusion

The question facing institutions is not whether to use AI in marking but which kind, under what controls. This paper has argued that the distinction between generative and classification architectures is not an implementation detail: it determines whether training data can be specified, whether outputs are repeatable, and whether the system's limits are visible. Deployment analytics from six institutions show that a per-question classification approach, embedded in a human-in-the-loop workflow, can mark most questions in under a minute, multiply the reach of well-written feedback, and signal reliably when its own suggestions deserve scrutiny. Markers and students in earlier pilots described the resulting feedback as more consistent and, for markers, considerably faster to produce, although students did not consistently experience faster turnaround. Within the constraints of vendor-collected, observational data, the findings suggest that regulators' preference for transparent, human-supervised systems and the practical interests of educators point in the same direction. None of this establishes that classification is the only viable route. Generative models have a role in composing and restructuring marker-approved feedback for individual students, and institutions with the technical capacity can build and govern their own tools against their own academic regulations; the criteria set out in this paper are intended to help evaluate any of these rather than to settle the question in advance. Independent, comparative evaluation is the obvious next step, and the author would welcome it.

References

Bai, X., & Stede, M. (2022). A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. *International Journal of Artificial Intelligence in Education, 33*, 992–1030. https://doi.org/10.1007/s40593-022-00323-0

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*. https://data.europa.eu/eli/reg/2024/1689/oj

Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal, 51*(1), 201–224. https://doi.org/10.1002/berj.4069

Kainth, M. (2022). University of Birmingham – Graide efficacy research. 6 Bit Education. https://www.graide.co.uk/blog/university-of-birmingham-research

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

Ofqual. (2024). *Ofqual's approach to regulating the use of artificial intelligence in the qualifications sector*. GOV.UK. https://www.gov.uk/government/publications/ofquals-approach-to-regulating-the-use-of-artificial-intelligence-in-the-qualifications-sector

Warr, M., Oster, N. J., & Isaac, R. (2025). Implicit bias in large language models: Experimental proof and implications for education. *Journal of Research on Technology in Education, 57*(6), 1324–1349. https://doi.org/10.1080/15391523.2024.2395295

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

ASCILITE 2026

The Future is Connection:

Collaboration, Creativity and Compassion in a Digital Society

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology, 56*(1), 150–166. https://doi.org/10.1111/bjet.13494

Kainth, M. (2026). Classification AI vs Generative AI in grading and feedback. In Cochrane, T., Narayan, V., Poole, T., Lai, P., Lake, P., Moore, O., & Li, L. (Eds.), *The Future is Connection: Collaboration, Creativity and Compassion in a Digital Society*. Proceedings ASCILITE 2026. Brisbane (pp. 251–260). https://doi.org/10.65106/apubs.2026.3409

Note: All published papers are refereed, having undergone a double-blind peer-review process.
The author(s) assign a Creative Commons by attribution license enabling others to distribute, remix, tweak, and build upon their work, even commercially, as long as credit is given to the author(s) for the original creation.

© Kainth, M. 2026