UNIVERSITY OF CAMBRIDGE
Manchester Metropolitan University
The University of Nottingham
University of Lancashire
ai@cam

AI in University Assessment: Evaluating the Opportunities and Risks of Automated Marking

OPRAISE

Executive Summary

Generative Artificial Intelligence (AI) has disrupted the entire education sector. Our work concerns the ability of AI, particularly Large Language Models (LLMs), to evaluate students' work, particularly their long-form responses to open-ended questions. Many students report using LLMs to seek feedback on essays, including in high-stakes situations. Our research question was whether AI-generated numerical feedback is sufficiently robust to support students and educators. We contextualised our evidence by considering the views of stakeholders on the broader opportunities and risks of integrating AI systems into University assessment practices.

Since assessment is a core academic activity, introducing AI into assessment practices may well have substantial, broad, and long-lasting implications for higher education. Nevertheless, there are strong drivers for AI adoption, which make it likely that Universities will soon need to agree on their stance. Our research provides evidence on the current capabilities of AI systems in assessment to illuminate the key considerations for decision-makers.

We tested how well three frontier AI systems assess examination submissions in Psychology undergraduate degrees across three UK Universities, representing diverse locations and student cohorts. We compared AI marks to the formal marks these submissions received by human assessors during routine examination and investigated the factors that influenced AI marks. Alongside this quantitative comparison, we gathered the views of staff and students through focus groups.

We found that the AI systems we tested exhibited intriguing, often highly impressive capabilities, yet none is currently "good enough".

Accuracy

For some institutions in our sample, the agreement between the best-performing AI system and human markers was at the level expected across the sector between human markers. Nevertheless, accuracy was not uniformly high and depended on the choice of AI system, prompting strategy and assessment context. The agreement between human and AI systems on which degree band to award, an important accuracy metric, was only moderate, ranging between 35%-65%.

Validity

The evidence challenged the validity of AI marking. The ability of AI systems to predict human marks was sensitive to the levels of student attainment, such that the discrepancy between AI systems and human markers was greatest for the best and the poorest submissions. AI systems were oversensitive to linguistic features, assigning higher marks to longer essays, those with complex sentences and a greater range of words, and where words indicated connection between ideas.

Reliability

The AI systems we evaluated provided marks that were consistent when queried at several different time points. The marks provided by one system were more consistent with the marks provided by another system than with human-provided marks; AI-to-AI agreement was often excellent. However, unanimous agreement across all tested AI systems on which degree band to award was more limited (56%). This is important because focus group data suggests that while variability of human assessors is tolerated, and sometimes welcomed, it is less acceptable in AI systems.

Feedback

AI systems readily offer more extensive feedback compared to human markers. In our sample, human feedback averaged between 100-200 words, while AI systems provided feedback that was 3-8 times longer. Participants in our focus groups indicated that lengthy feedback, while welcome by some, may not always be educationally useful or universally liked. When constrained to be the same length, they found it difficult to tell whether a piece of feedback was provided by a human or by AI.

The staff and students who participated in our focus groups, as well as our advisory board members, emphasized that current assessment practices and feedback are core to the 'social contract' between academics and students. While current practices are not perfect, our research suggests that adopting AI in assessment poses a risk of breaking this contract. All stakeholders, including both students and academics, fundamentally value human contact, and view human assessment and feedback as core to this engagement.

> ***"If both the students and the teachers are using it, then like, who's learning? What are we doing here?"*** - Manchester Met student

RECOMMENDATIONS

1 Exert caution in adopting current capabilities — we recommend caution in adopting AI systems for assessment.

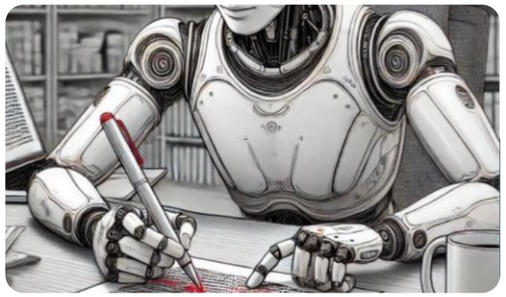
We found that the accuracy and validity of AI systems is not yet sufficiently robust to support students and educators. Furthermore, our findings indicate that accuracy and validity vary across contexts, with particularly concerning discrepancies at grade boundaries and for the highest - and lowest-performing submissions. Deploying AI systems for assessment should be conditional on evidence of stability, alignment with human judgement, and the absence of systematic bias, to ensure fairness and maintain academic standards.

2 Evaluate AI systems locally — performance does not generalise across institutions

While our results showed that frontier AI systems were not yet "good enough", the technology is developing quickly and performance could reach desired levels soon. In considering the adoption of future AI systems in assessment, institutions should not rely on performance evidence generated in other contexts. Our results show substantial variation in accuracy between institutions, likely reflecting differences in assessment design, student cohorts, and mark distributions. Since AI performance was context-dependent, institutions should conduct local validation using the specific AI system that is being considered, and their own assessment materials and marking practices, before any deployment. The system's performance should be continuously monitored. Our project provides a recipe for institutions on what evidence to collect prior to considering adopting AI for assessment.

3 Protect the human relationship at the core of assessment

The adoption of AI in assessment must preserve the central role of human relationships in higher education. Evidence from focus groups indicates that assessment is not only a technical process but part of a broader "social contract" between students and staff, underpinning trust, motivation, and engagement. Reliance on AI for assessment risks weakening this relationship by reducing opportunities for students to feel seen and for academics to engage with students. Institutions should therefore ensure that human authority, visibility, and dialogue remain integral to assessment practices, and that they are enhanced, rather than weakened, where AI tools are introduced.



Why adopt AI in assessment practices?

Pedagogy

AI systems could enhance student learning outcomes. For example, they can provide more frequent formative feedback to students who have limited contact time with academics, and tailor textual feedback to suit students' individual preferences or desired learning outcomes, hence supporting active engagement with feedback.

Efficiency

AI systems could save resources and channel resources more optimally. For example, using AI systems for marking could free academics to carry out tasks that have greater value for them and their students, such as small-group teaching, 1:1 mentorship or oral examination (viva).

Fairness and transparency

AI systems offer opportunities to monitor human marking processes both within and across degree courses and institutions. Used well, this can increase the quality of marking. For example, if AI systems are used to moderate human marks, they may be able to decrease the variability between markers by pointing to submissions where marking certainty is lower.

AI adoption scenarios

The opportunities and risk associated with adopting AI in assessment depend on the scenario in which AI is deployed. The degree of alignment required prior to adoption may also be scenario-dependent.

Quality assurance of human marking
AI independently marks submissions alongside a human marker. This could include producing a parallel mark where any significant differences trigger a review.

Marking Assistant
AI ranks or categorises submissions by predicted quality or uncertainty. This could include using AI marking to triage and prioritise complex cases; ranking submissions according to the preference of human markers; or expanding brief human comments to feedback that is optimised to support learning.

AI as a Primary Marker
AI marks all students' work with some human review. This could include where a human marks a sample of the submissions or reviews mark distributions, or when AI is used for formative feedback. The evidence we gathered suggest that this scenario is only acceptable when the performance of AI systems improves. Stakeholders did not endorse adopting AI as a sole marker.

Quantitative evaluation of three frontier AI systems

The question concerns the alignment between the marks that AI systems award a student's submission, and the mark that a human expert would give the same submission.

We acknowledge that the mark a submission received in the process of routine examination is not a perfect measure of its true quality. Furthermore, assessors of complex, open-ended, long-form essays, never perfectly align with each other, often due to legitimate differences in academic judgement. Nevertheless, since academic judgement is a socially accepted gold standard for assessments in higher education, it was used here as a benchmark to evaluate the performance of AI systems.

We evaluated the alignment of three frontier models (Gemini 3 Flash, GPT-5.4, Claude Opus 4.6), individually and as an ensemble, against human marks assigned during routine Psychology undergraduate examination at three UK Universities. We used a structured approach to identify the optimal prompting strategy and report the results from the best-performing strategy.

The following sections consider several characteristics of alignment, informed by the scholarly literature, the views of stakeholders, and our advisory board.

Accuracy:
How often does AI give the exact same mark as a human, or a reasonably close mark?

Reliability:
How close are the marks that the same model or different models give for the same submission?

Validity:
Are there systematic deviations between human and AI marks?

ARE AI MARKS ACCURATE?

(FIGURE 1)

Institution	Band accuracy	Mark differences	QWK	Spearman ρ
University of Cambridge	63%	5.2	0.61	0.65
University of Nottingham	53%	5.2	0.42	0.50
Manchester Metropolitan University	35%	8.3	0.33	0.36

- **Band accuracy:** *The proportion of times AI and human award the same UK degree classification band (First, Upper Second, Lower Second, Third, Fail).*
- **Mark differences:** *Average difference between AI and human marks.*
- **Quadratic Weighted Kappa (QWK):** *The degree of correspondence between AI marks and human marks. It is corrected for chance and penalises larger disagreements.*
- **Spearman** *ρ: The degree of similarity between the way that humans and AI order the essays.*

The table depicts four complementary measures of the accuracy of ensemble AI systems. The ensemble represents an aggregate across the three systems we tested, to improve performance. Overall, the accuracy of ensemble AI systems was moderate at best.

Figure 2 presents accuracy data separately for each institution and AI system. To evaluate the magnitude of the correlation between human and AI marks (QWK) it is important to consider that the correlation between two human markers of essay submissions is also often moderate.

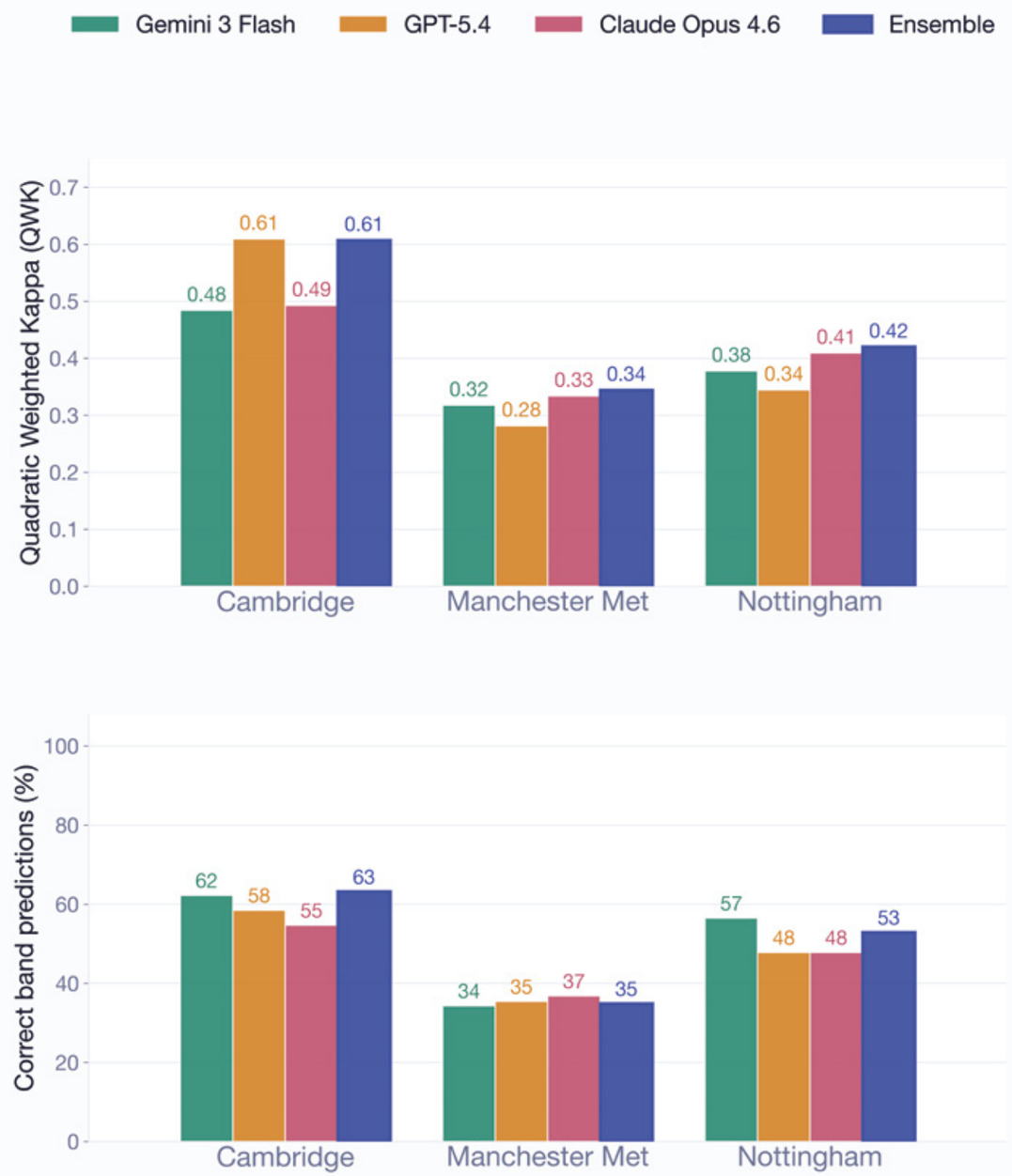
> *Interested readers will find a useful exposition here: https://www.gov.uk/government/publications/principles-of-ai-use-in-marking/principles-of-ai-use-in-marking#executive-summary*

This puts a logical limit to the degree of correlation we should expect between human and AI marks. While at Cambridge human-AI agreement, expressed as a correlation (QWK), was close to that expected from human-to-human agreement, it was lower than that in the other two Universities.

Stakeholders suggested that agreement on the degree classification is a more meaningful measure of alignment between human and AI systems than the correlation between them, while acknowledging that this measure might hide important within-band variation this information is displayed in Figure 2. Across all three Universities, band accuracy fell short of a threshold that would typically be considered appropriate for adopting AI marking in any assessment scenario.

Taken together, the results show that accuracy was not equal across Universities. This means that headline accuracy figures from one context (e.g. at one University or for one type of assessment) cannot be taken as evidence that AI marking is "ready" for deployment outside of the context in which it was tested. The variability we observed could be due to a range of factors associated with assessment design, the module, the course, the student body, or the institutional culture. Therefore, decision-makers should evaluate the AI systems they are considering in their own assessment context before adopting AI marking.

(FIGURE 2)



Agreement within the same AI model over time. LLM users often notice that AI systems provide different answers to the same prompt. It may therefore be surprising that each AI system provided highly consistent marks when asked to re-mark the same essays. All models performed excellently, with GPT-5.4 slightly noisier.

Agreement between AI models. The three AI systems also showed strong agreement with one another. To verify this, we compared the marks that each of the three systems provided for each submission. Considering each possible pair of models at a time, we found that their agreement was consistently high (QWK > 0.74 for all three pairs). Considering all three models together, the Intra-Class Correlation (ICC), a measure of reliability, was excellent (ICC = 0.91). In over 99% of cases, at least two of the three models assigned the same degree band, and almost all disagreements were confined to adjacent bands. However, inter-model reliability was lower when we considered the strictest criterion, whether all three models assigned the same degree band. This occurred for only 56% of submissions.

(FIGURE 3)

	ICC	% identical marks
Gemini-3 Flash	1.00	99
Claude Opus 4.6	1.00	98
GPT-5.4	.97	84

Data from a re-scoring of a sample of 100 essays over five days, stratified across the three Universities and all grade bands. The table depicts Intra-Class Correlations (ICC).

Are AI marks valid?

Academic judgement is based on reasoning, while AI marks are based on statistical predictions. Alignment between human and AI marks is not due to shared underlying generation mechanism. Additionally, alignment may reflect appropriate and valid processes, as well as weakness and problem. In this section, we discuss systematic differences between human and AI marks. However imperfect human marks may be, this section considers them to provide the "ground truth".

When compared to human markers, AI systems displayed two core biases. They assigned middling marks to all submissions, resulting in particularly inaccurate marking of the best and worst essays, and they were more sensitive to the linguistic characteristics of submissions compared to human markers.

AI marks exhibit a central tendency bias

Stakeholders agreed that it is important to reflect on the magnitude and direction of differences between human and AI marks. They felt that disagreement between AI and human markers is not a fault to be corrected, but a signal to be understood.

We observed that the gap between human and AI marks was systematically larger at the tails of the mark distribution and smaller in the middle for every AI system we investigated and when considered together as an ensemble. The top row of the Figure 4 plots the difference between AI and human marks against the human-assigned grade for each institution.

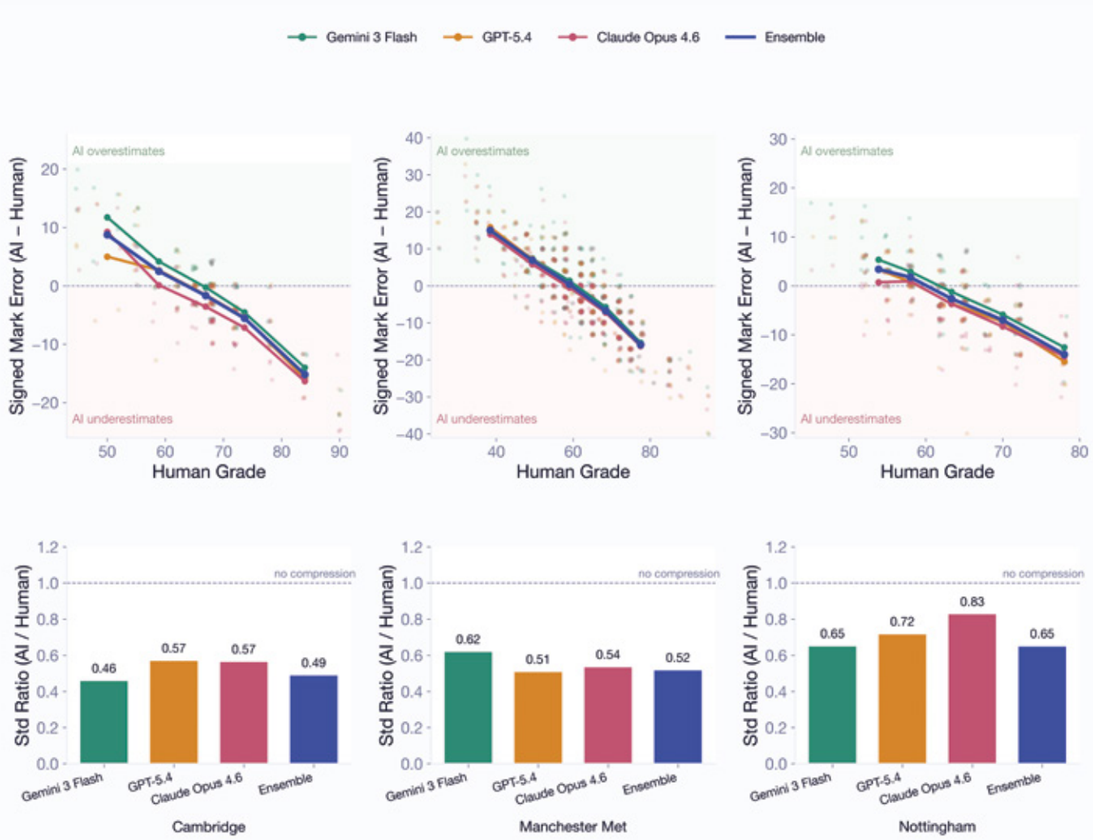
We devised a range of prompting strategies to overcome this bias, but as Figure 4 shows, even the best-performing model exhibited this bias. Considering all models together as an ensemble does not correct this bias because it reflects a systematic property of current-generation LLM scoring.

Due to this bias,

- An essay marked 75 by a human is, on average, scored several points lower by every AI system.
- An essay marked 50 by a human is, on average, scored several points higher by every AI system.
- The crossover - the point at which AI and human agree on average - sits in the upper-50s to low-60s across institutions, i.e. near the centre of each cohort's grade distribution.

This bias meant that the range of AI marks were compressed relative to the human marks they were trying to approximate, as quantified in the bottom row of the Figure 4. A compression score of 1.0 would indicate that the AI preserves the spread of the human marks, but all AI systems fall below this line. The scores range from 0.47 (Cambridge, Gemini) to 0.82 (Nottingham, Claude). The practical consequence of this bias is that the AI is least accurate precisely where assessment decisions matter most, that is, at the boundaries that distinguish Firsts from Upper Seconds, and passes from fails.

(FIGURE 4)



AI marks are oversensitive to linguistic features

A range of language features were extracted from each submission and grouped into categories. Figure 5 lists and describes some of the categories we considered and their constituent features.

(FIGURE 5)

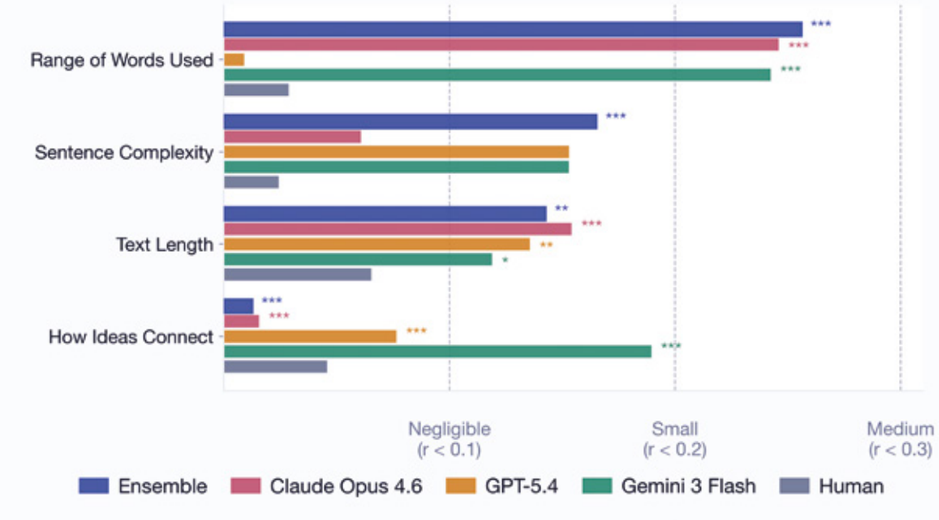
Linguistic Category	Features included
Range of Words: the variety and richness of vocabulary	Type-Token Ratio (TTR); Guiraud's Index; Herdan's C; Maas's a^2; Dugast's U
How ideas connect: the cohesion and coherence of the discourse.	Average content overlap; Connective count; connective density; Number of entities; Entity Continuity
Sentence Complexity: Syntactic complexity and structural elaboration in sentences	Average parse depth; Average word length; Clause ratio; Average sentence length; Passive ratio
Text Length	Word count

The figure shows how these linguistic categories relate to marks given by human and AI systems, separately and as an ensemble. The length of each bar shows how strongly each feature predicts AI and human marks and asterisks indicate if effects were statistically significant.

We found small but statistically significant relationships between all linguistic categories and AI marks. In contrast, the relationships between linguistic categories and human marks were broadly negligible.

Different AI systems varied in the strength of their responses to specific linguistic features. For example, Range of Words was the feature that predicted ensemble AI marks most robustly, with a small-to-medium effect, AI sensitivity was due more to the sensitivity of Gemini and Claude than GPT.

(FIGURE 6)



The impact of adopting AI in assessment on stakeholders

This section draws on the three themes identified in the qualitative study — (1) human–machine ambiguity, (2) overreliance weakening cognition and the value of education, and (3) the social contract — to describe the impact of AI adoption on stakeholders. We discuss impact on stakeholders overall, while acknowledging distinct implications for stakeholder groups. Students are primarily affected by issues of fairness, transparency, and the right to explanation, with potential additional risks for some student groups. Academic staff face challenges related to professional identity; while AI may reduce marking time, reduced marking experience could affect skill development and job satisfaction. Institutional decision-makers must balance the decreased workload associated with marking with risks such as financial irreversibility and model instability, and with legal, governance and regulatory risk.

The discussions in focus groups characterised an implicit "social contract" between academics and students at Universities, which sets shared expectations of respect, goal of learning and commitment to fairness and mutual support.

Assessment is a core element of this social contract. The engagement with students' thinking and struggle through marking underpins core academic functions, including teaching, tutoring, curriculum and assessment design. These functions are essential for staff development and skill maintenance. Moreover, assessment shapes students' sense of recognition, fairness, and belonging, with some student participants reporting they would feel "cheated" if AI marked their work. Our participants perceived the adoption of AI systems in assessment as a potential threat to learning at University.

> *"Assessment is not simply a system for distributing marks. It is also part of how educational meaning is made: how students feel seen, how standards are enacted, how trust is maintained, and how institutions reproduce their own values"* - Dr Steve Watson - Advisory Board Member

Human interactions are fundamental to this social contract. Students need to be seen and valued by their teachers just as much as academics need to see and evaluate their students through assessment. Irreplaceable moments include comments prompting new inquiries, discussions following mark disputes, and the reassurance of being respected. Participants agreed that lack of human touch could stop students attending and cause staff to leave.

The term "AI" clearly raised ambivalent emotional reactions for participants. The discussions in focus groups suggest that consultations about AI adoption would be more productive if it is focused on concrete scenarios and examples, such as those illustrated in this report. Furthermore, these discussions suggest that AI marking magnified the low trust that students and staff were already feeling towards each other and their institutions. Trust, alongside technical accuracy, is vital; a system must be fair and transparent to be effective. There was a broad agreement that AI should be introduced carefully and gradually,

Taken together, adopting AI into assessment in higher education will require wide, open and transparent discussions. These are essential to build trust and maintain the elements of assessment that are crucial to the fundamental meaning of higher education.

The evidence we gathered raises legal, governance, and regulatory implications that should be considered carefully.

While our research primarily focused on the ability of AI systems to provide numerical marks, under GDPR Article 22 students have a right to explanation, which has greater implications for some AI adoption scenarios. Furthermore, the evidence we collected regarding the impact of student attainment levels and language use on AI marks, and regarding the variation of AI performance across Universities – potentially due to characteristics of the student cohorts – suggests that some groups may be affected by AI adoption more than others. Therefore, decision-makers should reassure themselves that adopting AI systems conforms to their duties under the UK Equality Act. By their very nature, AI systems can make mistakes, which may not be caught easily. Therefore, any of the AI adoption scenarios may require revision to current appeal processes.

Adopting AI systems in assessment will require implementing change across many professional services. Clearly, adoption is easier where Universities already utilise digital examination processes, but it nevertheless crucially relies on support from IT, examination, quality assurance, data protection, compliance and contract services. The workload associated with the required adaptation should be balanced against the resources that may be freed by using AI systems. Consultation with operational and administrative staff is also needed regarding practical constraints.

Additionally, it is important to acknowledge the risk associated with these organisational changes. Even when AI systems perform well at one point in time, the evidence we collected suggests that alignment is model and institution-specific, and therefore, that continued review is essential. Once staffing and investment have shifted, institutions may find it challenging to continue collecting high-quality human marking and return to human marking, even if they wish to do so.

To discharge their duties responsibly, AI literacy training is required for all staff and students, especially those involved in institutional decision making.

> *"But if you put that into AI … it doesn't take into account any of that relationship, the conversations, the discussions, the intellectual sort of thoughts that the student has had and sharing with you. None of that…"* - (Manchester Met staff)

The advisory board agreed that final authority over marking must remain with human markers, regardless of how extensively AI is used in assessment.

While AI was widely seen as potentially valuable for quality assurance, moderation, triage, feedback support, consistency checks, and error detection, members emphasised that AI lacks the judgement, contextual understanding, and professional accountability that underpins academic assessment. Disagreements between AI and human marks were framed not as errors to be resolved algorithmically but as signals requiring human interpretation, with large discrepancies triggering further human review rather than AI override. The view that AI should never determine the final mark was endorsed by staff, students, and board members alike, reflecting concerns about validity, fairness, explainability, trust, educational value, and the social contract between Universities and students. Taken together, AI was accepted only insofar as it supports, informs, or scrutinises human judgement, not replaces it, and any scenario in which AI would hold final marking authority was regarded as the riskiest and least defensible option.

Selected questions for further research:

What are the best ways to use AI as a marking assistant, and what are some key pitfalls? There is a risk that work pressures and natural cognitive biases may encourage human markers to defer to AI marks inappropriately. How Universities best deploy AI to increase the quality and consistency of human assessment, while avoiding introducing new biases?

What is the most useful format of AI-generated feedback, and how can AI be used to improve the usefulness of human-generated feedback? AI offers students and educators the ability to tailor textual feedback according to specifications, and can be deployed as a dialogic tool to help students engage with feedback. How might Universities deploy this ability to confer genuine educational benefits?

What implications does the evidence presented here have for student guidance? The evidence we presented here on the success and limitations of these systems could help universities offer better guidance to their students on how to use AI systems more wisely.

Project methodology

Why Psychology?

The project focused on a single academic discipline, Psychology. An initial focus on a single discipline was deemed beneficial. We selected Psychology because essays are central to Psychology degree results, increasing the impact of this work. The nature of writing in academic Psychology presents an ideal testing ground for AI assessment capabilities because unlike many other fields, Psychology does not primarily assess outcomes or results, but rather context-dependent 'research judgement'. Psychology essays require the evaluation of multiple competencies, increasing the generalisability of our results. The focus on Psychology was particularly beneficial for the qualitative strand of the research, because the discipline requires both scientific and reflective skills. As such, participants could provide multifaceted insights into the role of AI in a human academic environment.

QUANTITATIVE METHODOLOGY

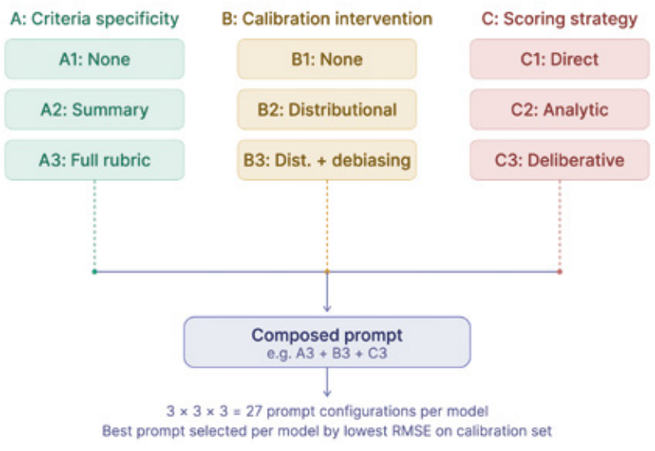
The selection of AI systems was limited to those that could meet the strictest privacy protections required to protect this sensitive dataset. The project received ethical approval from Cambridge Higher Education Studies Research Ethics Committee and Manchester Metropolitan EthOS.

The dataset - 125 students in 3 UK Universities volunteered 761 authentic long-form undergraduate Psychology essays (University of Cambridge: 133, University of Nottingham: 172, Manchester Metropolitan University: 456). All essays were submissions to formal assessments between 2022-2025. They spanned 50 modules and 87 distinct assignments across all years of study. Assessments spanned coursework, open book at-home examinations and invigilated examination. Essay marks, on a 0-100 scale, were moderated formal marks provided by expert human assessors who followed routine institutional processes under routine examination conditions.

AI platforms and models - We tested the performance of three frontier large-language-models, Claude Opus 4.6 (Anthropic), GPT-5.4 (OpenAI), and Gemini 3 Flash (Google). Models were selected from different providers on the principle that models from different families are likely to exhibit different biases and failure modes, were limited to those that could meet the strictest privacy protections required to protect this sensitive dataset. This diversity is desirable in its own right for ensemble-based marking and guards against the risk of correlated errors that can arise when closely related systems share training data or architecture. Within each family, we used the latest generation of generally available commercial model at the time of study. All three were accessed through their native APIs at temperature = 0 (the standard deterministic setting). We note that model updates pose instability risks, such that the data presented here is only accurate for these models.

Prompt design - Rather than committing to a single prompt, varied the prompt along multiple dimensions to isolate distinct sources of prompt influence on scoring. Figure 5 describes the selected dimensions. At the most basic level, models were prompted by the following statement: *"You are an experienced <University name> (terms in quotes were replaced by the appropriate names) examiner marking <degree name> undergraduate assignment."* At the other end of the dichotomy, models were given the marking rubric, information about the expected mark distribution, and asked to justify aspects of the evaluation prior to providing a mark. Best-performing prompts per model were selected on a 20% calibration subset (n = 153); the same prompt configurations were then applied to the full corpus for the analyses reported here.

(FIGURE 7)



- **(A) Criteria specificity** - what rubric specification the model receives. Three levels: no rubric; summary grade boundaries only; or the full analytic rubric with per-dimension performance descriptors.
- **(B) Calibration intervention** - how the model is distributionally anchored, addressing the known tendency of LLMs to compress scores toward the middle of the scale. Three levels: no calibration; distributional base-rate information; or distributional information plus an explicit debiasing instruction.
- **(C) Scoring strategy** - how the model is asked to reason. Three levels: holistic direct scoring; analytic-decomposed scoring across rubric dimensions followed by synthesis; or holistic-deliberative scoring, in which the model first identifies a grade band and then assigns a specific mark within it.

Crossing the three dimensions fully yields 27 prompt configurations per model.

Aggregation - Since no single AI marker is reliably better than the others, predictions were combined across the three models to test whether combining models can improve their performance. We compared four aggregation rules and report the winning method - the inverse-RMSE weighted mean, in which each model's contribution is scaled by its accuracy on the calibration set.

QUALITATIVE METHODOLOGY

We conducted 9 semi-structured focus groups with 25 participants, which were each approximately 1 hour long. Discussions focused on general uses of generative AI in academia, and more specifically, its uses in assessment and marking. Participants were provided with examples of textual feedback produced by human and LLM. Thematic analysis identified three themes.

Theme 1: Human/Machine ambiguity

Focus group participants often referred to AI as one's 'Frienemy', and only on occasions recognised that these are multiple tools. This Frienemy is both 'intimidating' and a 'collaborator', illustrating the complex emotional reactions humans have to this technology. As identified in previous research, trust in AI is low, and participants seek transparency in its use.

Our focus group discussions revealed that trust between staff, students, and Universities is low. For example, academics worry that Universities are trying to exploit staff and students for profit, and students are not sure that Universities offer value for money, or that all processes are transparent and fair. We observed that AI magnified these feelings. For example, staff suspect that Universities will implement AI systems to cut jobs rather than improve education. Some participants in focus groups agreed that avoiding the term 'AI marking' and instead using "a specialised marking tool, developed for marking Psychology essays, with high reliability and extensive feedback" may instil more confidence.

Last in this theme of blurred boundaries was the response to the quality of Human and machine feedback on essays. While some participants suggested that AI can be the solution to the perceived "subjectivity" of human marks, many welcomed human subjectivity, arguing it is equitable and important. By contrast, the tolerance to the difference between human and AI was generally lower.

Participants struggled to identify the source of example feedback. When the source was revealed, there was disagreement about which is better, with some arguing that longer and more detailed feedback is better, while others suggesting it might put them off.

Theme 2: Overreliance, Weakening Cognition & Value of Education

The identified concerns participants have on the impact of the use of AI on education, learning and human cognition:

"I think it [AI] kind of makes them [people who use it] more illiterate, not illiterate in the literal sense, but like illiterate in the sense of not being able to read a paper and understand what's going on"
— (Cambridge student)

In this whirl of concerns, participants saw positive potential for their academic discipline of Psychology, which offers insights on changes to cognition and skills, as well as offering participants themselves skills that they considered that AI cannot do well, such as qualitative research and detailed critical evaluation.

Overall, participants in the discussions were concerned about the value

(FIGURE 8)

Participant	Mean age	N Male	N Female	N Other	N Total
Staff	37.48	4	10	0	14
Students	21.5	3	6	2	11

of education in the age of AI:

"it's really easy to complete it with the AI- you just you don't really need to think that much. I'm kind of worried, like, what did I actually learn from the essay?" — (Nottingham student)

Theme 3: The social contract: personal and interpersonal values.

There was a unanimous agreement that a key value in education is the interaction between and among staff and students:

"there's nothing like that real personal human contact"
— (Nottingham staff)

The social contract involves discussions and interaction and affects motivation:

"if someone's worked really hard on an essay and then they get terrible feedback by AI, it may feel as if... their work hasn't been received by anyone of importance...no interaction there"
— (Manchester Metropolitan student)

"... [when a lecturer talks to you about your work] that makes you feel better about your work... it almost inspires you to try better ..." — (Cambridge student)

The social contract is also the thing that facilitates the magic of feeling as though one is being seen and valued, which was core in both students and staff discussions.

ACKNOWLEDGEMENTS

The project was funded by ai@cam. The project benefitted from the advice of Isla Fay and Ruth Walker and the support of Inbar Bobrovsky, as initial OpRaise project coordinator. We warmly thank our Advisory Board colleagues: Cathy Cheung (Google Research), Dr Gladson Chikwa (Manchester Metropolitan University), Prof. Keeley Crockett (Manchester Metropolitan University), Hayley Spain (University of Nottingham), Dr Carolina Kuepper-Tetzel (University of Glasgow), Dr Steve Watson (University of Cambridge), and Dr Gabriela Zapata (University of Nottingham) and students: Nikolett Franyo (Manchester Metropolitan University), Yashrag Garg (University of Cambridge), and Rowan Meijer (University of Nottingham) for their time, attention, and support. We are very grateful to colleagues and students who participated in our advisory board for their ideas, passion, and friendly but vigorous challenge. We especially want to thank the students who contributed their essays and thereby made this project possible.

PROJECT TEAM

OpRaise combined two fundamental areas of expertise practitioner experience, with team members possessing decade-long experience in teaching and assessment in higher education, and AI experts' ability to develop and interrogate frontier AI models. Our interdisciplinary project team included experimental, cognitive, and social Psychology, decision science, computer science and AI.

Human Assessment & Stakeholder Perspectives

Principal Investigator:
Dr Deborah Talmi
University of Cambridge

Dr Yael Benn
Manchester Metropolitan University

Lyba Razzaq
Manchester Metropolitan University

Dr Roni Tibon
University of Nottingham

Machine Assessment

Dr Maryam Abo-Tabik
University of Central Lancashire

Dr Giulio Corsi
University of Cambridge

Dr Alexandru Marcoci
University of Cambridge

Impact

Georgiana Thorpe Apreutesei
University of Cambridge

Dr Sarah Mattey
Cambridge University Press & Assessment

Dr Hilla Tal
LOGOS, University of Cambridge

To find out more, please scan the QR code or visit below:

www.emotional-cognition.psychol.cam.ac.uk/opraise