

ual:

UAL Online: Scalable Assessment Research Project

Interim Report, June 2025

Dave White, Dean of Academic Strategy Online

Ian Truelove, Research and Innovation Coordinator, UAL Online

Contents

Executive summary

Interim Recommendations	2
AI embedded in curriculum	3
Risks in incorporating AI into the assessment workflow	3
Reducing the volume of work submitted by students	4

Main report

Introduction	5
Assessment and technologies of Cultural Production (AI)	6
More than accuracy	6
Avoiding hypocrisy and retaining institutional agency	7
Perceptions of value	7
Principles for incorporating AI into assessment practices	8
Less, but better	10
Scalable assessment: Phase two	12

Appendices – see separate document:

UALO-Scalable-Assessment-Interim-Report-Appendices.pdf

Appendix I: Literature

Appendix II: Assessment workloads from selected UK HE providers

Appendix III: Evidence from assessment tool testing

Appendix IV: Consent, Legal and Ethical approval forms

Executive summary

Interim Recommendations

Assessment practices at UAL are subject appropriate, maintain academic standards, and are developmentally effective for students. They are also of a form which mitigates recent increased risks of plagiarism due to AI.

On the basis that our assessment practices are fundamentally sound, there are two lines of enquiry we have been pursuing to seek efficiencies without eroding quality or engendering inappropriate levels of institutional risk.

1. Incorporate Large Language Model (LLM) AI into the assessment workflow.
2. Reduce the volume of work submitted by students for assessment.

Our enquires to-date have informed the following interim recommendations:

Interim Recommendation 1:

Do not automate student facing aspects of the assessment workflow as the risks outweigh the theoretical efficiencies.

Interim Recommendation 2:

Explore the use of AI in the non-student facing aspects of the assessment such as 'scoping' a body of student work, parity and constructive alignment of feedback.

Interim Recommendation 3:

Reduce the volume of work assessed at each submission point.

AI embedded in curriculum

A consequence of our work on assessment is a better understanding of the guiderails for including AI in the flow of teaching and learning. Whilst there are risks associated with the anthropomorphising of AI¹, when structured into a process of human-centred critique and dialogue, Large Language Model ‘chatbot’ style AI platforms can be useful tools for stimulating reflection and discussion². Insight gained in this research is being fed into UAL thinking and the UAL Online Teaching and Learning model.

Risks in incorporating AI into the assessment workflow

Our research to-date indicates that the technology cannot currently generate outputs which are tailored specifically enough to a given student to be of developmental value³. While outputs from LLMs are often ‘accurate’ we suggest they currently lack the insight required to be meaningful to students.

Beyond considerations of effectiveness, our experiments also indicate that using AI for feedback and marking carries significant reputational risks for UAL. These risks relate to student perceptions of fairness⁴ and the ethics of algorithms being embedded in decisions which have a significant impact on student’s lives⁵.

The ability of LLMs to detect inappropriate or irrelevant content is rapidly improving. However, our assessments routinely demand nuanced critiques of culturally sensitive territories which are challenging to algorithmically interpret. Furthermore, excellent student work may well go ‘beyond’ the brief or interpret it in ways we could not have predicted. In this case strong student work could be misread or not-visible-to AI.

In effect, the only way to maintain academic quality is for academics to check and modify all aspects of the assessment workflow where AI outputs have been incorporated. This then attenuates any presumed student facing efficiencies, while still carrying significant reputational risks.

¹ Manzini, A., Keeling, G., Alberts, L., Vallor, S., Morris, M. R., & Gabriel, I. (2024). The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1), 943-957.

² Waring, M. & Evans, C. (2023). *Facilitating Students' Development of Assessment and Feedback Skills Through Critical Engagement with Generative Artificial Intelligence*. <https://doi.org/10.13140/RG.2.2.19781.83685>

³ See Appendix III: Evidence from assessment tool testing.

⁴ The use of AI by academics in ways which appear similar to those we restrict or ban for students.

⁵ “Key to [human agency] is the right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them. (European Commission, 2019, p. 16)

Exploring the use of AI in managing non-student facing aspects of the assessment process

Gaining an overview of a body of student work to establish the quality-range of a summative assessment is challenging and time consuming. Given this, there is a valid line of enquiry exploring the use of AI in the process of 'scoping': establishing the parameters of, and sequencing, a large body of work to prepare it for assessment.

We also propose to explore the use of AI as an aid to improving parity in marking and feedback, where multiple human assessors are assessing different students who have submitted for the same assignment.

If proved to be of value these processes will inform the assessment process without automating academic judgement.

Reducing the volume of work submitted by students

Smaller unit sizes offer flexibility and improve access but increase the number of assessment points. Smooth continuation through the units that make up a course necessitates a quick turnaround of assessment. To balance this, we should manage assessment load by reducing the volume of work required at each point.

Volume of work and academic quality are not proportionately linked⁶, and national Quality Standards do not specify the amount of work submitted at a given level⁷. The sector is outcomes, not outputs, based. There is also a case to be made that in our specialist subjects a lower volume of written work will rebalance the emphasis towards assessed visual work.

In terms of Quality, where we now face increased risk of sophisticated plagiarism via the inappropriate use of AI, it is better to deeply scrutinise a smaller volume of work rather than be forced to use the time available to 'get through' marking.

⁶ Tomas, C. & Jessop, T. (2019) Struggling and juggling: a comparison of student assessment loads across research and teaching-intensive universities, *Assessment & Evaluation in Higher Education*, 44(1), 1-10.

⁷ QAA (2018) *UK Quality Code for Higher Education. Advice and Guidance: Assessment*. London: QAA.

Main report:

Introduction

This interim report is a summary of UAL Online Academic Strategy's exploration of scalable assessment approaches to date. Interim recommendations are based on initial experiments with LLM AI technologies, ongoing discussions relating to the design of assessment and the experience of the authors in arts education. Recommendations sit within the Quality and regulatory environment of UAL and the OfS⁸.

Summative assessment practices at UAL are intricate, intellectually demanding and time-consuming. They combine the application of frameworks and criteria to ensure constructive alignment with expert, culturally informed, judgement to understand the substance of the work.

Assessment at UAL is rarely about converging on a correct answer. It is more commonly focused on holistically considering individual, or group, interpretation of creative briefs against a set of Learning Outcomes. This type of assessment is central to creative arts education. It is also proving to be a valid method for mitigating the impact of the emergence of AI-type technologies on the authenticity and validity of the education we offer. There is no reason why this would not hold true in fully online provision at UAL.

In short, assessment practices at UAL are academically effective and much of the university sector is now considering a more 'arts' informed approach in response to emerging technologies⁹. Therefore, when considering scaling assessment at UAL and in online provision we are researching methods of reducing cognitive load and time-pressure on academic staff rather than seeking to change the underlying mode of assessment. If assessment practices can be undertaken more efficiently it allows us the possibility of releasing staff time for teaching and/or reducing the cost to students, thereby improving access.

Most of our time to-date has been spent researching the incorporation of AI technologies to improve the efficiency of feedback and marking. It should be noted that feedback and marking are the end point of the process which begins during assessment design. When considering scalable assessment, we therefore also consider the design of assessment.

⁸ At time of publication.

⁹ Kristandl, G. (2024) *Fusion or Confusion? Co-Creating Assessments with GenAI and Students*. Learning and Teaching Festival 2024, University of Greenwich. Manifesto of the essay in the age of AI and A student Manifesto for assessment in the age of AI from LSE – both describe a move to assessment practices which are akin to UAL practices.

Assessment and technologies of Cultural Production (AI)

Artificial Intelligence as a category of technology is vast and vague. The term AI is best thought of as a brand rather than a type of technology¹⁰, something which is applied more for marketing purposes than as a description of how any given technology operates.

Nevertheless, there is a predominant cluster of technologies labelled as AI which can ‘read’ and produce convincing text-based language and various styles of images. This category of AI we can think of as a technology of Cultural Production¹¹.

In broad terms, Higher Education, especially arts education, is a system of Cultural Production. Given this, technologies such as AI are variously seen as a threat or an opportunity in Higher Education because they, in principle, encroach on ‘the work of the academy’.

In a teaching focused university, such as UAL, assessment practices are central to this work and involve critically evaluating huge amounts of text and images. The question then raised is to what extent AI – which purports to have these capabilities – can be, or should be, incorporated into the workflow of assessment.

More than accuracy

As with any emergent technology of Cultural Production the initial focus is on how accurate it is or, to put it another way, how closely can it replicate the work of a human. In the context of this research this means, ‘*how similar is the AI’s output to an academic’s marking and feedback?*’. Under what circumstances does the AI fail to mirror expected outputs?

In our initial tests of automated assessment, we used a leading AI assessment tool, Keath.ai, and custom prompts in Microsoft Copilot¹² to mark simulated and real student submissions¹³. Keath.ai is a commercial AI-powered assessment tool that auto-grades student submissions and generates written feedback. Microsoft Copilot is a general-purpose LLM-powered chatbot that can be used for auto-assessment through custom prompting. Outputs from both approaches were not entirely unexpected and did map onto human grading of the same work¹⁴. Auto-generated feedback from both Keath and Copilot, whilst not up to the standard required, looked human generated, and was not entirely inaccurate.

Although accuracy of these AI assessment tools will improve in time, accuracy is a narrow measure in the context of UAL, and we must also consider reputational risks. Examples include perceptions of value, perceptions of work-security, academic ethics and the possibility of cultural

¹⁰ Goatley, W. (2025) *Nero Fiddled While Rome Burned: Towards Low-Carbon Teaching in AI and Computational Contexts*. UAL Digital Learning and the Environment: Practice Sharing Day, 2 April 2025.

¹¹ Born, G. (2010). The Social and the Aesthetic: For a Post-Bourdieuian Theory of Cultural Production. *Cultural Sociology*, 4(2), 171-208.

¹² <https://keath.ai/> & https://en.wikipedia.org/wiki/Microsoft_Copilot

¹³ Tests of automated assessment using Keath.ai and MS Copilot with simulated work created by UALO colleagues and a small sample of previously submitted student work by MA Performance Writing students at CSM.

¹⁴ See Appendix III: Evidence from assessment tool testing.

bias or cultural ambivalence surfacing. There are also important questions relating to the carbon cost of AI use and ethical issues relating to the use of stolen data to train the underlying foundational AI models, which we acknowledge but have not included in discussion here.

Avoiding hypocrisy and retaining institutional agency

Our students have an acute sense of justice and fairness and rightly expect that the standards they are held to should also apply to the institution as a whole¹⁵. As such, the use of AI in the functioning of the university should not be incorporated in a manner which we would not, in principle, allow of the students. Add to this that assessment is central to our value proposition and therefore an area we do not want to erode our agency.

Perceptions of value

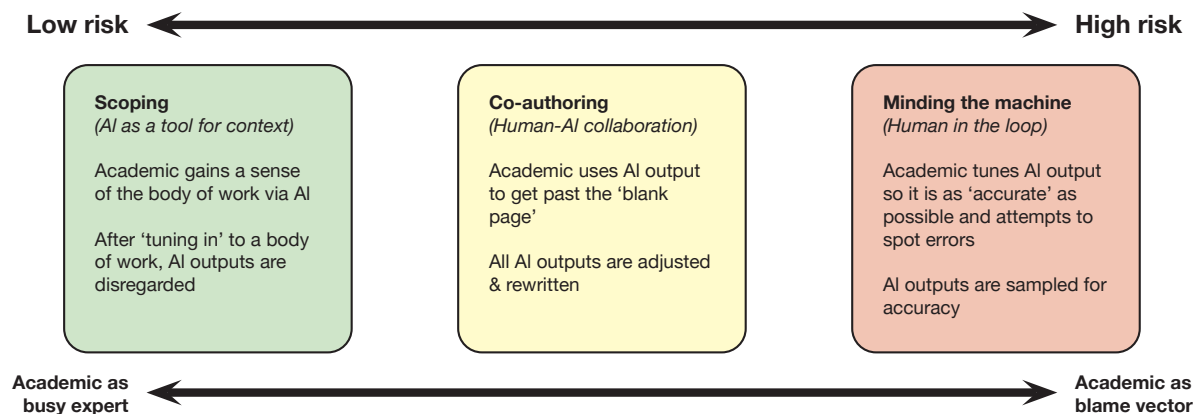
The value of Keath.ai is more in the workflow layer it adds to an LLM than in the specificity of its outputs. The workflow includes batch marking, training the model for given assessment and responding to relevant assessment criteria.

It should be noted that student feedback generated from Keath.ai was not significantly different from that generated by other mainstream and easily accessible chatbots such as Chat GPT and Claude. Students have access to the same technologies as our staff and can use them as they see fit. They can easily generate similar feedback before handing in work and may well question the value of what is institutionally provided if AI is used in the assessment process.

¹⁵ Sherwood, C. (2025) *Art or Algorithm: Exploring UAL Students' Perspectives on the use of AI in Creative Education and the Creative Industries*. London: UAL Students' Union.

Principles for incorporating AI into assessment practices

The following principles for the use of AI in curriculum and assessment take their cue from our current UAL AI Position statement¹⁶ and AI and Assessment guidance¹⁷.



□ Scoping

Assessing project work holistically requires an academic to have a clear understanding of the relevant learning outcomes, the brief, the assessment criteria at the correct level and the overall tone or aims of the unit/course. There is also a process of scoping, or 'tuning into', the work directly, getting a feel for the tone across a group/cohort, any patterns of thinking or themes and the overall quality of the work.

UAL assessment practices include initial sample marking and discussion to improve parity of grading across multiple markers. Everyone is then faced with a pile of work which might contain outliers or atypical work of varying quality. Often the process of scoping the work can be time consuming, involving reviewing then revisiting a several submissions to calibrate thinking and cognitively embed the various components of the process mentioned above.

AI could be used in a low-risk manner to aid 'tuning into assessing' before any grades or feedback are decided on. A sample of work could be inputted among with the assessment criteria, brief and learning outcomes. Agreed and tested prompts could then be used to output a summary of themes, thinking and references used in the work alongside a high-level report on which criteria and learning outcomes appear to have been most successfully met.

Reviewing these outputs would acclimatise an academic to the body of work without providing any individualised information. This overview of the work would be treated as a useful-but-potentially-naïve. Reviewing the AI output is likely to both be a reminder of the pedagogic aims of the assessment and highlight, by absence, the things the AI has missed but the academic knows will be in the work.

¹⁶ <https://www.arts.ac.uk/about-ual/teaching-and-learning-exchange/digital-learning/ai-and-education/ai-position-statement>

¹⁷ <https://www.arts.ac.uk/about-ual/teaching-and-learning-exchange/digital-learning/ai-and-education/ai-and-assessment>

The AI output is then put to one side and the assessment process starts in earnest with the academic equipped with the framework for assessment and an overview of the work. This approach locates the use of AI firmly as ‘assistive’ and not as ‘automated judgement’. The AI is being used to help sense-make across a large body of work, without offering up marks or feedback. Security, data protection and ethical risks associated with uploading student work to commercial cloud-based LLMs could be mitigated in this scenario by using locally hosted bespoke LLMs, such as LM Studio¹⁸.

□ Co-authoring

Work is fed through an agreed and tested AI process to produce suggested marks and feedback. All marks and feedback are reviewed and potentially edited/adjusted by the academic. In theory this is using AI only to get past the ‘blank page’ on the assumption that it’s more efficient to adjust outputs than start from scratch with each student submission.

Interestingly this is what AI marking platforms tend to promote to make the case that the final arbiter of quality must be the expert and not the AI, the implication being that student work must be reviewed by academics to ensure quality. We are wary of the assumed efficiency of this approach as to meaningfully adjust the AI outputs; student work would need to be engaged with in some depth. In practice, this approach is likely to lead to an overreliance on the AI output by time-poor academics with the concomitant institutional risks.

There are forms of assessment where students are likely to submit work within predictable parameters. In these cases, an academic might develop a small menu of set feedback and marks which are lightly adjusted where required. If the design of assessment is geared this way, then co-authoring with AI could be appropriate. The extent to which this method of assessment is desirable or appropriate at UAL is important to consider, as is the reaction of students to receiving ‘set’ or quasi automated feedback.

□ Minding the machine

This approach is often promoted by those providing AI services – the ‘Human-in-the Loop’. In this case the AI produces marks and feedback after an AI training process and the academic checks a sample much like an External Examiner might. This is an industrialisation of knowledge work which relocates expertise away from direct human engagement to maintaining and adjusting the machine.

This might work for assessments where there are agreed correct answers, but it is high risk where student work is creative and individualised. It also risks removing the developmental aspect of

¹⁸ <https://lmstudio.ai>

assessment as academics will often give nuanced feedback to those who have ‘failed’ to find the ‘correct’ answer to put them on the right track.

In short, ‘minding the machine’ could work for forms of assessment which are not common at UAL and the kind of assessment which is an easy target for automated ‘cheating’ by students. In essence, **where AI can be used in assessment for massive efficiency gains is exactly where AI can also be used to mitigate the value of the overall process.** There is an obvious logic to this if we believe that learning is defined by cognitive and creative struggle or, to put it in UAL terms, telling the story of process.

It is of note that one AI assessment platform we reviewed was incorporating a feature which allowed ‘managers’ to see time-on-task of academics – how long they had spent looking at each assessment. This, we believe, is so that the institution can use the platform to automate assessment while insisting that Quality still resides with the academic. The human-in-the-loop then is a euphemism for the human as *blame vector*. It is also a form of institutional surveillance which will be rightly questioned by the unions.

Less, but better.

Whichever form assessment takes, reducing the amount of work that students submit for assessment reduces the amount of time it takes to assess, making assessment easier to scale up, and potentially reducing the time between a submission deadline and a student receiving feedback. However, if students are asked to submit too little work this risks the integrity of the assessment. The challenge is to determine the optimum amount of evidence required for assessors to make robust academic judgments.

According to publicly available data¹⁹, the amount of work students are required to submit for assessment on UK Masters-level programmes varies significantly. As published data tends to focus on word-counts, determining equivalence in submission loads for a UAL course is not straightforward, but some data regarding non-writing-based submissions can be found in [Appendix II](#), with UAL Online coming in at the top end of the volume of work required from students. Anecdotal evidence²⁰ indicates an inflation of submission loads across UAL over the last five years, possibly due to the move to electronic submissions, a lack of confidence in assessment processes, or a natural tendency to add more assessment tasks without subtracting redundant elements.

The literature indicates that research intensive universities, which tend to be perceived as more prestigious, have fewer summative assessment points than teaching intensive universities, with cumulative summative assessment loads significantly higher in post-92 providers²¹. There are

¹⁹ Data sourced from a wide range of UK HE providers’ websites and public-facing documents.

²⁰ UAL AI Network and conversations with UAL colleagues across different colleges.

²¹ Tomas, C. & Jessop, T. (2019) Struggling and juggling: a comparison of student assessment loads across research and teaching-intensive universities, *Assessment & Evaluation in Higher Education*, 44(1), 1-10, DOI: [10.1080/02602938.2018.1463355](https://doi.org/10.1080/02602938.2018.1463355)

many complex factors that drive this disparity, but high prestige research intensive institutions maintain the integrity of their assessments despite assessing less student work overall. Regulators do not specify the amount of evidence that needs to be submitted by students²², so institutions have leeway to determine submission workloads as appropriate to institutional and subject-specific contexts. UAL is in the rare position of being both highly prestigious and teaching intensive, which offers a wide range of options for determining students' assessment workloads. In the context of the specialist subject of art and design, 'less, but better'²³ may be the most appropriate approach. A focus on quality not quantity could also help build students' subject-specific skills and attributes, as skilfully curating a selection of strong creative outcomes is often a 'real-world' task. An additional factor in assessment confidence in the UAL model is that assessors will have normally taught the students they are assessing and will have gained a good sense of their capabilities through extensive formative assessment activities. Whilst not all the evidence that has accumulated over the duration of a unit needs to be submitted at the end, it can legitimately inform the summative assessment of carefully selected outcomes (including selected evidence of developmental processes). This intimate knowledge of a person can enable nuanced judgements that can lead to more accurate and fairer assessments and highlights another potential limitation of AI automation. However, both humans and chatbots are subject to different forms of bias, with risks of bias amplification in hybrid assessment models, so this complex area warrants more research.

The context of UAL Online is different to most of the rest of UAL in that many of our units must be able to act as discrete entities – some students may purchase and complete only one unit. In the traditional residential model, units are more likely to be seen as building blocks towards the completion of an award, and so there is more emphasis on cumulative continuation. Whilst the principle of continuing toward the completion of a full Masters degree is retained in the UAL Online model, the flexibility afforded by standalone units requires a different way of thinking about assessment. Rather than completion at the end of a typical UAL course, there is a more significant completion point at the end of each UAL Online unit. The temptation is to conflate this with end-of-course assessment and cram too much assessment into each unit but, for the reasons stated, this would not lead to better outcomes for students and would most likely lead to assessment workloads that are not sustainable at scale.

Instead, working on the principle that 'less, but better' is good for students, good for assessors and good for scalability, we suggest that the UAL Online Assessment Framework be revisited with the aim of reducing the amount of work students submit at the end of each Unit. Current approved UAL Online documentation supports this approach, with validated documents stating that students should not be overburdened in terms of volume of work submitted, and satisfying learning outcomes should be demonstrated through the quality and not the quantity of work²⁴.

²² QAA (2018) *UK Quality Code for Higher Education. Advice and Guidance: Assessment*. London: QAA.

²³ Rams, D. (1995) *Less, but Better (Weniger, aber besser)*. Berlin: Gestalten.

²⁴ Cover note, MA Graphic Design validation documentation, Section 2, p. 3, Assessment Framework Purpose.

Scalable assessment: Phase two

Interim Recommendation 1:

Do not automate student facing aspects of the assessment workflow as the risks outweigh the theoretical efficiencies.

In phase one, we auto-assessed simulated student work and a very small sample of actual student work, working in collaboration with Ray Grewal, Course Leader of MA Performance: Writing at CSM. The outcomes from these experiments (see Appendix III) combined with our broader research has enabled us to conclude that the risks associated with automating the student facing aspects of assessment outweighed the benefits. Therefore, we do not intend to continue with this line of enquiry in phase two.

Interim Recommendation 2:

Explore the use of AI in the non-student facing aspects of the assessment such as 'scoping' a body of student work, parity and constructive alignment of feedback.

In phase two, rather than using AI to directly assess student work, we will instead expand our research to test whether AI can assist human assessors by initially 'scoping' a set of submissions, with the aim of establishing whether academic cognitive load can be eased in the very early stages of assessment. We will also explore parity and constructive alignment of human feedback with an initial hypothesis that AI may be able to flag up inconsistencies in feedback and may be able to help identify where human feedback could benefit from additional constructive alignment to learning outcomes and assessment criteria. We have recently secured access to student work previously submitted by Level 5 BA Hons Graphic Communication Design students at CSM, along with their ethical consent to use their work in our research. We are collaborating with Jaap De Maat, Stage Leader on this course, to expand our AI experimentation and to connect with the students and staff who would potentially be affected by the deployment of such technologies. We will explore AI technologies beyond Keath.ai, expanding our testing of Microsoft Copilot and locally running Large Language Models via platforms like LM Studio.

Interim Recommendation 3:

Reduce the volume of work assessed at each submission point.

In phase one, the evidence supported a reduction in the volume of work that students should submit at assessment points. In phase two we will revisit the UALO Assessment Framework, with a view to reducing the volume of work required from students in their submissions for assessment.

We anticipate that feedback on, and critique of, this interim report by interested parties will help us refine or refute our interim recommendations and will help identify new lines of enquiry for phase two of this research project. Ideas, suggestions, corrections and objections can be shared with the authors via email at: i.truelove@arts.ac.uk