RESEARCH AND ANALYSIS

Principles of AI use in marking

Authors

- Joanna Williamson

With thanks to

- Tom Bramley, Jo Handford, Fiona Leahy, Tim Stratton and Frances Wilson
- Awarding organisation colleagues and other subject matter experts who participated in technical discussions

Contents

Executive summary

This working paper explores the principles and implications of using artificial intelligence (AI) in high stakes marking, specifically within the context of qualifications regulated by Ofqual in England. It considers the nature and capabilities of current AI technology, characteristics of the high stakes marking context, different models for AI integration, and frameworks for evaluating AI use in marking processes. It emphasises the importance of assessment validity, transparency, fairness and accountability for maintaining the integrity of qualifications and public confidence in them, while considering the potential benefits and risks of AI use in marking.

The motivation for publishing this working paper is to support thinking and foster constructive discussion about the potential applications of AI in marking, acknowledging that use of AI as the sole mechanism for determining a student's mark does not comply with Ofqual's regulations – as previously clarified in Ofqual's policy approach to regulating AI use in qualifications (Ofqual, 2024). This paper underlines the need to go beyond technological perspectives on AI and to integrate ethical, social and assessment validity concerns to inform future policy decisions.

Current AI technology

Current leading AI models, including large language models (LLMs) such as GPT models and BERT[1], are deep neural networks (DNNs) with advanced capabilities. Most notably, LLMs can generate human-like text in response to prompts but lack true semantic understanding and the capacity of human-like judgment. These current AI models function as 'black boxes,' making it difficult (even for experts) to explain how specific outputs are generated. This poses a challenge for transparency, and for reliably predicting the performance of AI on new tasks. These current AI systems also exhibit variability and unpredictability in outputs, with confidence measures that do not necessarily reflect real-world accuracy, which can complicate their use in high stakes contexts.

AI and regulation

AI regulation in the UK is based on 5 principles:

- safety, security and robustness
- appropriate transparency and explainability

[1] BERT (Bidirectional Encoder Representations from Transformers) is an open-source deep learning language model developed by Google for language processing tasks.

- fairness
- accountability and governance
- contestability and redress.

Achieving appropriate transparency and explainability, especially for complex AI models, is essential. However, it is also a significant challenge due to differing stakeholder needs and technical limitations.

In the context of qualifications and assessment, Ofqual has emphasised the need to ensure fairness for students, maintain the validity of qualifications, protect security, maintain public confidence and enable innovation. These are Ofqual's priorities in shaping its regulatory approach to AI, and they connect to the overarching UK government principles.

Current high stakes marking

Current marking for high stakes qualifications depends on academic judgment by trained experts, who apply mark schemes to assess candidate responses against assessment criteria. Markers draw upon strategies including matching, scanning, scrutinising and evaluation. Different marking tasks demand different types of judgment, and features of tasks and mark schemes influence the easiness or difficulty of marking. For example, tasks which are less constrained and offer students greater freedom can result in a very wide range of student responses and tend to be more difficult to mark consistently.

It is important to understand high stakes marking in its social context. Marking takes place within structured social systems and processes, including the wider education sector, policy and regulatory frameworks, and operational processes designed by awarding organisations. Qualifications are part of the social contract of education and are also social contracts in themselves.

Marking quality

Marking quality encompasses accuracy and reliability in the application of assessment criteria. In particular, fairness requires consistent application of criteria regardless of marker or candidate identity. Transparency and fairness in marking are not only ethical obligations but also functionally needed to uphold trust in qualifications.

AI and automated mark generation

AI and automated systems for generating marks vary widely. They range from rule-based deterministic engines to deep learning models using LLMs, which may be fine-tuned on human-marked responses or operate via prompting without specific training. These systems differ in explainability, consistency, reliance on training data, and susceptibility to bias, and Table 1 summarises some key differences.

The terms “AI marking”, “traditional AI marking”, and “automated marking” are used differently by different authors and organisations. This working report uses “AI and automated marking systems” to include all approaches shown in Table 1 collectively. In addition, hybrid marking systems may combine different AI and automated approaches, or combine them with human marking for approaches such as similarity-based marking.

The integration of an AI or automated marking system into an overall marking process can also follow multiple designs. For example, rather than serving as the primary marker, an AI or automated system may be used to quality assure human marking. Different designs have different implications for efficiency, oversight, and fairness. A significant attraction of using an LLM-based generative AI system is the possibility of a single system for different tasks and populations, rather than building or fine-tuning a dedicated marking system for each task.

Table 1: Some properties that vary by AI or automated marking system.

Type of AI or automated marking system	‘Classical’ feature-based engine trained on human-marked responses	DNN trained on human-marked responses	LLM fine-tuned on human-marked responses	Generative AI based on pre-trained LLM
Mark generation fully interpretable, for experts	Yes	No	No	No
Mark generation based on construct-relevant features	Under the control of system designers	Unknown – must be investigated	Unknown – must be investigated	Unknown – must be investigated
Consistency: same candidate response leads to same output every time	Yes	Yes	Not necessarily	Not necessarily
Reliability of marks generated	Empirical question	Empirical question	Empirical question	Empirical question
Training data fully specifiable	Yes	Yes	No	No
Training data under control of those building the scoring system	Yes - fully	Yes - fully	Data for fine-tuning - yes; data to train LLM - no.	No
Potential bias learned from human marking	Yes	Yes	Yes	No
Potential bias from other training data	No	No	Yes	Yes

Evaluating AI use in high stakes marking

Validity frameworks

Evaluation of AI use in high stakes marking must be grounded in assessment validity. Established frameworks for evaluating validity in automarking remain relevant, and offer a constructive route forward. This involves systematically gathering and weighing evidence that marks reflect the intended assessment construct. Recent research has demonstrated use of these frameworks with current AI technologies (for example, Casabianca et al., 2024).

Both research and industry guidance emphasise that agreement with human marks alone is insufficient for assuring validity. Important topics to consider include:

- **Relationship to human marking**
  An AI or automated marking system may have one of several different relationships to human marking. The intention may be to predict the results of human marking, complement human judgment, replicate human marking (including its flaws), or improve upon human marking. The intended relationship impacts the logic of the validity argument, and therefore the validity evidence needed.
- **Stakes and implementation**
  The level of validity evidence required increases with the stakes of the assessment and the extent to which AI influences final assessment outcomes. Use of AI as a sole marker requires more robust evidence than AI used for quality assurance.
- **Transparency and explainability challenges**
  Both responsible AI principles and assessment validity frameworks stress the need to understand how and why AI systems generate particular marks. AI systems based on deep learning and LLMs are difficult to interpret, which complicates the collection of validity evidence. Explainable AI (xAI) methods like saliency analyses offer some insights but have some limitations for the assessment context. In particular, they can produce misleading explanations.
- **Bias and fairness**
  AI models can perpetuate biases present in or arising from training data. In marking specifically, empirical studies have identified variable AI marking performance across demographic groups, emphasising the importance of fairness evaluations. For some current AI technologies, limited transparency about the sources of training data poses additional challenges for bias mitigation.

- **Accountability and oversight**
  Maintaining meaningful human oversight and responsibility for AI use in high stakes marking is critical. However, integrating human review in this context without negating potential efficiency gains is challenging. If an AI system were implemented as a primary marker, possible models could include human second marking of all or a subset of responses. Alternatively, it could involve human experts only at a higher level of oversight, for example in monitoring the statistical properties of autoscored items. For any implementation of AI in high stakes marking, monitoring should consider the overall marking process, including both human and AI inputs and their effectiveness in combination. The type and degree of human involvement considered sufficient for accountability remains a key regulatory question.

Consequences

Motivations for using AI in marking can include increasing efficiency, speeding up marking, and improving the quality of marking. These potential benefits are significant but require empirical validation within specific qualification contexts and taking account of the entire marking process.

Use of AI in marking also has the potential for negative consequences. It is important these are investigated, monitored and, if necessary, mitigated for. Potential risks include loss of assessment expertise due to reduced human marking, negative washback effects such as gaming, and a narrowing or stagnation of assessment constructs. Public perception and trust may also change where AI use reduces – or is perceived to reduce – transparency or accountability.

The environmental impact of AI use is another type of consequence to consider. AI systems require significant energy and water consumption, and it is important to weigh the overall resource implications of marking processes with and without AI use.

Finally, data security and candidate privacy, especially regarding the use of student work for AI training, are also critical concerns. These will be a factor in weighing potential AI implementations, and particularly choices between in-house AI development and use of third-party AI models.

Conclusions

This working paper explores the potential role of AI in high stakes marking, emphasising the importance of upholding assessment integrity and public trust in qualifications.

Principled arguments exist for accepting AI and automated marking in high stakes assessments, in the right conditions and with the right safeguards in place. Current evidence does not support an overall or general case for the validity of AI marking in high stakes qualifications, and at present points to the need for context-specific evidence (that is, specific to the particular qualification, in terms of the relevant candidate population, constructs and item types involved). Established assessment validity frameworks can support the design, documentation and evaluation of AI and automated marking. AI use may continue for now to be more viable in some qualifications than others, particularly where repeated use on similar tasks and candidate populations supports appropriate investment in evidence gathering.

While the use of AI as a sole marker is currently non-compliant with Ofqual regulations, there are promising applications of AI in marking, including in quality assurance. Ofqual is particularly supportive of work to explore such use of AI. To support responsible innovation in this area, it is important to continue research that addresses the challenges and unanswered questions identified in this working paper. In particular, these relate to the transparency and explainability of current AI, appropriate methods for assuring human oversight and accountability, and the relationship between AI use, public trust and qualification value.

Introduction

Purpose

This working paper explores the principles of using AI in marking and the implications for the context of high stakes qualifications in England. In particular, these include GCSEs, AS and A levels, and vocational and technical qualifications included in performance tables. Ofqual has previously published its policy approach to regulating AI use in qualifications (Ofqual, 2024). This makes clear that use of AI as the sole mechanism for determining a student’s mark would not comply with Ofqual’s regulations. However, we consider it very important to continue our thinking on AI and potential applications.

The paper looks at the nature and capabilities of current AI, the key characteristics of the marking context, different models for the integration or application of AI to high stakes marking, and the questions that should be used to evaluate potential AI use in this context. The paper is not a review of AI models, nor does it review the performance and capabilities of different models and systems with regard to different marking tasks. Rather, the purpose of this working paper is to consider the principles involved, to signpost informative areas of research (including some from different disciplines), and to try to synthesise insights from these areas into constructive and principled thinking about the use of AI in marking.

Understanding the implications of AI use in high stakes marking is necessary to inform our future policy in this area. While research on technical developments is very important, it is also necessary to go beyond thinking about AI as a technology question, not least because the details of technical developments are changing rapidly. Our thinking on this important topic is ongoing, and the work presented here includes questions that are not yet resolved. The aim in sharing this working paper is to be open and transparent about thinking on this topic and continue to foster constructive discussion.

Background

This paper focuses on marking for high stakes qualifications in England, particularly those qualifications regulated by Ofqual. Ofqual’s General Conditions specify that students taking a regulated qualification should be differentiated by criteria that are understood by assessors, accurately applied, and consistently applied (Ofqual, 2017, Section H).

Currently, marking for high stakes qualifications is underpinned by the application of academic judgement by a trained expert. The marking process involves assessors or markers applying their academic judgement to determine what marks to award a candidate response against the assessment criteria in the mark scheme. Marker judgement is informed by the marker's knowledge, expertise, experience, and their training, especially standardisation activities training them in application of the mark scheme.

As noted, the use of AI as a sole or primary marker in our context would not comply with Ofqual's current regulations. The published policy approach explains that this is due to the requirement for human judgment in marking decisions, as well as awareness of AI risks relating to bias, inaccuracy and transparency (Ofqual, 2024). In this working paper, however, we consider it important to address the hypothetical case of AI as sole or primary marker in thinking through the principles of AI use in marking.

Use of AI in marking

The motivations for using AI in a marking process can include improving the consistency of marking (for example, Foltz et al., 2020), increasing the perceived fairness of marking (for example, Figueras et al., 2025), and – most commonly – improving efficiency by reducing marking time, marking cost and the number of human experts required (Clauser et al., 2024; Rotou & Rupp, 2020).

AI systems may be used in various ways as part of a marking process. AI can be used to generate marks, but AI systems can also perform other roles in a marking process – for example, identifying responses likely to be atypical. Where an AI system is used to generate marks, that might be as the primary marker, but equally, the AI could be generating marks in a support or monitoring role, as part of a quality assurance process that still has people as primary markers. Besides performing different roles, AI systems used in a marking process may be based on diverse models, trained on varying datasets, and designed to produce different outputs. For example, a rule-based system differs significantly from a fine-tuned classifier or a generative AI model, and each has distinct properties and risks (see Section 3: AI and automated mark generation).

The immediate background to this work includes the rapid development of AI technologies over the past few years, most notably large language models (LLMs) and generative AI systems. LLMs and generative AI systems represent genuinely new possibilities for marking, but the use of AI in marking is far from new.

There is a long history of research and development in automated marking technologies, generally subdivided according to the type of assessment task. Research dedicated to automated marking of short-answer items is well established and known as ASAG (automated short answer grading). Short-answer questions can be thought of as a middle ground between closed questions (such as multiple-choice items) and open-ended questions (such as essay questions) and will generally require up to 2 or 3 sentences in response to a specific question. To mark such questions, the human marker or automated marking system must be able to recognise paraphrasing and identify different sentences with equivalent meaning, but does not need to assess complex arguments or quality of writing (Kortemeyer, 2024). Useful surveys of ASAG methods are given by Haller et al (2022) and Clauser et al (2024). Kortemeyer (2024) notes that most models are still fine-tuned or explicitly trained for marking particular tasks in ASAG, but research is now exploring the use of general-purpose pre-trained LLMs without fine-tuning (for example, Henkel et al., 2024). The difficulty of automarking short-answer items relates to the variability (both in terms of type and extent) and predictability of the possible student responses (Zesch et al., 2023). The most difficult items to mark are those where students produce a wide range of different responses, not all of which can be predicted in advance, including a wide range of potential correct responses that may be formulated differently but essentially convey the same meaning.

While ASAG aims to award marks for correct answers and not writing ability, the field known as AES (automated essay scoring) deals with the inverse: the ‘essays’ that are the focus of AES research and development are most commonly fairly short pieces of writing (about 300 words) scored for quality of written communication rather than their content (in many cases, the assessment context is English as an additional language). A helpful overview of AES systems is provided by Ifenthaler (2022), and there is a great deal of recent research investigating the use of deep learning and LLMs for AES (Flodén, 2025; Lottridge et al., 2024; Mizumoto & Eguchi, 2023; Pack et al., 2024).

Current AI

Currently leading AI models, including large language models (LLMs), are forms of deep artificial neural network models (DNNs) that demonstrate capabilities far in advance of earlier generations of AI. Most notably, LLMs, trained on text prediction tasks, respond to natural language inputs or prompts and are able to produce human-like text outputs. Whereas earlier AI models were constructed and trained using specific data to achieve specific ends, foundation models such as LLMs are pre-trained using very large quantities of text data to produce a general-purpose model, which can then be fine-tuned or adapted to carry out different specific tasks. The process of fine-tuning involves taking the original pre-trained model and training it further on new data (typically, focused on a particular task or domain). This additional training refines the model’s original learned representations of language to be more attuned to the nuances of the language used in the new task or domain – the model is not re-trained from scratch (IBM, 2024). For example, the LLMs BERT and GPT-4 are foundation models produced by Google and OpenAI respectively.

BERT has been adapted for many scientific purposes by taking the pre-trained BERT model and carrying out additional training and fine-tuning on domain-specific texts (for example, the BioBERT model for biomedical tasks, Lee et al., 2020). Meanwhile, the much larger model GPT-4 is particularly good at generating text, and underpins user-facing applications such as ChatGPT, Duolingo Max and Khanmigo (Jones, 2023)[2].

Current AI models also differ from earlier AI and machine learning technologies in the extent to which we can explain their working. In particular, the structure of an LLM is sufficiently complex that it functions as a 'black box' and neither users nor expert model designers are able to straightforwardly explain why particular inputs result in particular outputs. The field of 'explainable AI' (xAI) is concerned with improving model explainability, and includes various techniques to help provide explanations of 'black box' model functioning (see Section 1: Current AI, and discussion in the context of assessment validity in Section 4: Evaluating AI use in high stakes marking).

AI and regulation

In the UK, regulation of AI is underpinned by a principles-based framework which sector-specific regulators interpret and apply to their contexts (UK Government, 2023). The framework has 5 core principles, which build on the OECD principles for responsible use of AI (OECD, 2019). The 5 principles are:

- safety, security and robustness
- appropriate transparency and explainability
- fairness
- accountability and governance
- contestability and redress

(UK Government, 2023, page 26)

Many frameworks for responsible and ethical AI use exist. They can differ in terminology and emphasis, and may define key terms with different scope, but there is significant overlap in the core dimensions referenced (Papagiannidis et al., 2025).

While adherence to responsible AI principles is increasingly sought and in some places required by law, the operationalisation and realisation of principles in practice may be challenging (Gajjar & Brione, 2024). Transparency and explainability, for example, are widely accepted as principles of responsible AI, with major ethical and social implications. However, judging how much transparency and explainability is

[2] At the time of writing, GPT-5 is the latest model and is being adopted by major applications.

needed, and for whom, is not straightforward. The appropriate, proportionate and necessary level may differ for designers of an AI implementation, decision-makers, regulators, operational staff running and maintaining it, and end users. Achieving the functional and ethical outcomes intended means carefully considering the specific context in which AI is used (Vredenburgh, 2024). This is one of the motivations for the ongoing work described in this report.

Structure of this report

Sections 1 to 3 aim to provide a baseline for the remainder of the report, by summarising important concepts, results, and arguments that are relevant to the question of AI use in high stakes marking. Section 1 aims to highlight relevant characteristics of current AI. Section 2 discusses the important characteristics of the high stakes marking context, and section 3 outlines different methods of using AI in marking processes. In particular, it summarises the major approaches to automated generation of marks, and contextualises use of current AI against other automated marking technologies. Section 4 outlines validity frameworks relevant to the use of AI and automarking systems. These are research-based, principled approaches that offer constructive ways to think through and evaluate the use of AI in high stakes marking processes.

Section 1: Current AI

The section aims to highlight relevant characteristics of deep artificial neural network (DNN) models, and particularly large language models (LLMs), which underpin currently leading AI systems.

Artificial neural network models are a form of machine learning, and the terms ‘deep learning’ and ‘deep neural networks’ refer to models constructed from multiple layers of network nodes. Current leading LLMs are ‘transformer models’, a subset of DNN models with a particular architecture that allows them to learn context from the relationships in sequential data (like words in a sentence). Good summaries of LLMs and foundation models are given by Bommasani et al. (2021, pages 3 to 5) and the Ada Lovelace Institute’s explainer ‘What is a foundation model?’ (Jones, 2023), while a more in-depth technical account of deep learning and neural network models can be found in Wolfram (2024). In theoretical terms, what LLMs (like other language models) do is predict “the conditional probability of a token—which might be a character, word, or other string” given its context (Liao & Wortman Vaughan, 2024, page 5).

As noted in the introduction, deep neural network (DNN) models generally, including LLMs, differ from other technologies – and specifically earlier forms of machine learning technologies – in the extent to which we can explain their working (Zerilli et al., 2019). The structure of a DNN is sufficiently complex that it functions as a ‘black box’ and without further work (for example, using the techniques of explainable AI) is impenetrable not just to ordinary users but to model designers. The structure of a DNN model is made up of layers, each containing parameters (weights and biases which shape how the model processes data to make predictions). These parameters are automatically adjusted by an algorithm as it learns from the training data. For a complex task a DNN may have millions or billions of parameters. The result is that “humans, even those who design them [the models], cannot understand how variables are being combined to make predictions” (Rudin & Radin, 2019, page 3).

Capabilities

A distinguishing feature of LLMs is that they appear to demonstrate capabilities well beyond solving the next-token (for example, next-word) prediction task that is the basis of their training (Lu et al., 2024; Vafa et al., 2024). The extent to which LLMs demonstrate emergent abilities is a matter of ongoing research and often heated debate, not least because it relates to whether or how far LLMs represent (or are interpreted as) a step towards Artificial General Intelligence[3]. There is, further, longstanding philosophical debate about the relationship between being hypothetically able to accurately model semantic understanding, and true semantic understanding (see, for example, Durt, 2022).

LLMs can often complete tasks in ways that appear to demonstrate functional linguistic skills. Research continues to emphasise, however, that "despite their usefulness in various tasks, current AI models fall short of understanding language" (Dentella et al., 2024, page 1; see also Bottazzi Grifoni & Ferrario, 2025; Floridi, 2023; Hicks et al., 2024). Analysing LLMs from the perspective of semiotics (the study of signs, representation and meaning) is a useful way to think about how and why in more detail. Drawing on Charles Peirce's theory of signs, Legg for example notes that LLM training allows the models to learn (important) aspects of language meaning, but not grasp the full meanings of human concepts "because their [LLMs'] learned associations between terms, however rich and fine-grained, are insufficiently scaffolded by robust indices to real-world objects and also insufficiently disciplined by iconic structures – most crucially, logical form" (Legg, forthcoming, page 2).

Current AI models can excel at advanced tasks, but also continue to falter, and this includes generating some factually inaccurate outputs. Many have pointed out that this should be expected as a result of how LLMs are designed and trained, namely, to "predict the likelihood of different next utterances based on prior utterances", which does not reliably or necessarily relate to factuality (Weidinger et al., 2021, page 22, see also Hicks et al., 2024; Webb, 2023; Waldo & Boussard, 2025).

A range of recent studies have highlighted gaps between successful task performance by current AI systems, and what these tasks may have been expected or assumed to demonstrate about AI capabilities. Route-finding problems in New York city, for example, revealed interesting results about the ability of AI to deduce structural representations of a domain or problem (Vafa et al., 2024). The research found that successfully completing route-finding problems did not imply that the AI model had deduced or constructed a valid working model of the city's street plan. Relatedly, recent studies have shown that completing reasoning tasks need not mean that AI systems are in fact using analogic or mathematical reasoning processes (Lewis & Mitchell, 2025; Mahdavi et al., 2025; Mirzadeh et al., 2024; Shojaee et al., 2025; see also Shanahan, 2024). It is important to stress that current AI systems can achieve excellent results on reasoning and mathematics tasks. What increasing numbers of studies show, however, is the importance of not assuming

[3] Engaging in this question is clearly out of scope of this report. Some experts argue that AGI, or indeed substantial improvements to current AI abilities, will not emerge from connectionist systems such as DNNs without incorporating some use of explicit symbols (for example, a hybrid between language-based and symbol-based AI, see Lenat & Marcus (2023); see also the long history of connectionism vs computationalism debates).

(from task performance) that the means used resemble human-like mathematical or algorithmic reasoning (Zhao et al., 2025). The steps reported by AI systems as a ‘chain-of-thought’ or ‘reasoning’ do not necessarily reflect the actual processes used to reach a task solution (Chen et al., 2025; Turpin et al., 2023).

There is not yet consensus on a scientific or principled answer, nor overwhelming empirical evidence, to divide tasks into those that are definitely and reliably solved by current AI, and those that are not. Currently, users and potential users require empirical evidence to find out what AI systems can reliably achieve. This presents challenges for those trying to apply AI, since without understanding how AI succeeds at a task, the empirical evidence has limited generalisability – the key matter of which task features will cause changes in task performance (a major component of AI robustness) remains as a separate empirical question. In the route-finding task for example, Vafa and colleagues found that performance was fragile and dropped substantially when any detours were introduced; in assessment and reasoning tasks, studies have shown sensitivity to surface level features such as replacing a word with a synonym (Aloisi, 2023; Mirzadeh et al., 2024).

There are several factors that cause variable behaviours in current AI systems. Some researchers have referred to the idea of “capability unpredictability”, which describes “the idea that an LLM’s capabilities cannot be fully anticipated, even by the model’s creators, until its behavior on certain input is observed” (Liao & Wortman Vaughan, 2024, page 6). Variations in AI system performance can also occur due to updates to underlying models, which providers may not explain or announce (Liao & Wortman Vaughan, 2024; Tate et al., 2024). Furthermore, non-determinism in LLM models themselves can mean that the same prompt or input results in different outputs. Interestingly, this has been observed even with LLM model settings intended to maximise determinism (that is, with the ‘temperature’ parameter set to zero) (Atil et al., 2025).

Current AI systems can report levels of confidence to users, but these – like their outputs more generally – are not always accurate or reliable (Liao & Wortman Vaughan, 2024; Wang, 2024). For LLMs, the likelihood assigned to a particular ‘next token’ provides one built-in measure of uncertainty. However, this is typically not the measure of uncertainty that a human user needs: the ‘next token’ probability corresponds to something more like ‘lexical confidence’, that is, how likely it is that that token follows, rather than how likely it is that the meaning of the resulting text is factually accurate (Kuhn et al., 2023; Liao & Wortman Vaughan, 2024). These ‘next token’ probabilities are therefore not well calibrated to real-world probabilities. Other techniques for reporting levels of confidence to users include fine-tuning an LLM to describe its own confidence, and asking the LLM to evaluate multiple different answers to the task or query, but again, these are not always accurate or reliable (Liao & Wortman Vaughan, 2024, pages 27 to 28). Achieving calibration – the ability to produce probabilities associated with model outputs that correspond to reality –

would support trust, and enable users to make informed, appropriate and safe use of model outputs. This is clear if we consider the alternatives: an ‘underconfident’ model that produces accurate outputs but with low confidence, or an ‘overconfident’ model that rates the likely accuracy of its own outputs more highly than is warranted. Better calibration of LLMs would improve their real-world usability, since users take into account both the accuracy of predictions and the confidence associated with predictions as part of decision-making (Wang, 2024; Zhu et al., 2023).

AI capabilities and judgment

The apparently intelligent output that is produced by generative AI systems can cause difficulty in grasping the nature of current AI. In particular, it is “increasingly tempting” to anthropomorphize current AI, frame AI capabilities in human terms, and attribute human-like powers to the AI system as an explanation for its output, despite the fact that AI need not use the same means as humans (Shanahan, 2024, page 68; Dentella et al., 2024; Durt, 2022).

The use of anthropomorphizing language may hinder users in accurately understanding AI systems, and contribute to over-trusting their outputs (Liao & Wortman Vaughan, 2024; Shanahan, 2024; Weidinger et al., 2021). Describing non-factual AI outputs as ‘hallucinations’ is an example of well-established but probably unhelpful language that lacks accuracy. Specifically, “Since an LLM has no awareness and no senses or cognition, it cannot be said to perceive and thus its outputs cannot qualify as hallucinations” (Graydon & Lehman, 2025, page 15). Relatedly, Shanahan argues that it does not make sense to talk about an LLM believing or judging:

> Only in the context of a capacity to distinguish truth from falsehood can we legitimately speak of *belief* in its fullest sense. But an LLM—the bare-bones model—is not in the business of making judgements. It just models what words are likely to follow other words. (Shanahan, 2024, page 73).

The distinction between reckoning and judgement can be helpful in discussing current AI: the term ‘reckoning’ describes a calculation-like process, while ‘judgement’ can be reserved for the deliberative process (see, for example, Pamuk, 2023; OED, 2025).

Transparency, explainability and interpretability

The related terms ‘transparency’, ‘explainability’ and ‘interpretability’ do not have fixed agreed meanings (Vredenburgh, 2024). This report uses the terms ‘transparency’ and ‘explainability’ in line with the working definitions provided in the UK Government’s principles-based framework for AI regulation, which in turn reflect the OECD principles (OECD, 2019, Principle 1.3). Transparency refers to “the communication of appropriate information about an AI system to relevant people (for example, information on how, when, and for which purposes an AI system is being used)” (UK Government, 2023, page 28). Transparency is a core part of responsible AI frameworks because it is a necessary foundation for public accountability, effective governance and further scientific progress in AI. Specifically, unless there is transparency, “stakeholders cannot understand foundation models, who they affect, and the impact they have on society” (Bommasani et al., 2025, page 2).

Explainability refers to “the extent to which it is possible for relevant parties to access, interpret and understand the decision-making processes of an AI system” (UK Government, 2023, page 28). The main motivations for trying to explain the logic of an AI model include to justify (especially, to improve the justifiability of AI-generated decisions in sensitive or high stakes contexts); to improve control and allow debugging; to improve the model by refining its accuracy and efficiency; and to discover new domain knowledge (Ali et al., 2023; Marcinkevičs & Vogt, 2023; Vilone & Longo, 2021). The concept of explainability and the field of explainable AI research are necessary because current AI models are generally not explainable in themselves, and their design and training do not themselves guarantee the properties human users are likely to want them to have. Above all, machine learning models including current AI locate and exploit associations in data, which may or may not relate to real-world causal relationships.

The AI literature discusses both ‘explainability’ and ‘interpretability’ at length. While sometimes used interchangeably, ‘interpretability’ is generally used to describe the potential for (some expert human) to understand the decision-making model itself and how AI model outputs are generated (Ali et al., 2023; Bommasani et al., 2021; Rudin, 2019). Where used in this report, this is the sense intended by ‘interpretable’. The broad definition of ‘explainability’ expressed in the UK Government principles makes interpretable models a subset of explainable models. It is worth noting, however, that usage in the research literature can differ: for example, ‘explainable’ is used by many to describe AI models whose working can only be explained to users by a separate post-hoc model, in contrast to ‘interpretable’ models where the model itself must be understandable, namely, “built from a small set of clearly defined and comprehensible variables, where these variables stand in a linear or other easily graspable relationship with the outcome variable” (Vredenburgh, 2024, page 3).

As noted in the Introduction, there is consensus that transparency and explainability are important for responsible use of AI, but translating these principles into practice and achieving them may not be at all straightforward. It may require balancing the needs and competing priorities of different groups, where risks and power are unequally distributed. Decisions may need to be made in a rapidly-changing context with significant unknowns, and where levels of necessary transparency and explainability are agreed it may still be technically challenging to achieve them.

Recent research has observed that the best way to improve transparency in AI remains unclear, despite widely recognised rising concerns about opacity in AI and use of AI (Bommasani et al., 2025, page 3). Relatedly, Laux et al note that measures to increase transparency with the aim of strengthening trust can have “ambiguous effects”, and in some contexts even “backfire” (Laux et al., 2024, pages 9, 27). Considering explainability, a key point is that not all proffered explanations of AI models are equally accurate, appropriate or effective. In particular, there is consensus among regulators and those issuing guidance that automatically generated explanations must be carefully assessed, not just accepted (Ali et al., 2023, page 30; see for example Information Commissioner’s Office, 2023; Leslie, 2020).

Bias

Unfairness caused by bias is a major risk to the responsible and ethical use of AI. Like other forms of machine learning, current AI models can both encode and perpetuate biases of multiple types, from a range of mechanisms (Liao & Wortman Vaughan, 2024; Mehrabi et al., 2021). Foundation models can have intrinsic biases (such as misrepresentation, underrepresentation and overrepresentation) that result in harm when the foundation model is used or adapted. They may also have properties that result in biased performance or other harms when the model is used in certain contexts (for example, performing less well for some subgroup of people compared with others). Bias is especially concerning for foundation models because these models are precisely intended to be used widely and in diverse applications, meaning that both positive and negative properties can be highly leveraged. In so far as problematic intrinsic bias exists in a foundation model, it may be perpetuated very widely through any application built using that model (Bommasani et al., 2021, pages 5, 131 to 132).

As many have pointed out, the risk of bias from training data can be anticipated for foundation models, given that the data sources used to train the models (so far as these sources are known) are not representative of all societies, regions and groups (for example, Bender et al., 2021). However, the precise relationship between training data, data practices and intrinsic biases in the resulting foundation model are not fully understood (Bommasani et al., 2021, page 132). For most current AI models there is low transparency about training data (Bommasani et al., 2025), presenting challenges for both research and suggested mitigation strategies that involve evaluating training data.

AI and decision-making

The way in which AI is incorporated into decision-making processes matters for both risk and accountability. In particular, the risk of an inaccurate or unacceptable output from AI is not the same as the risk that this output will be implemented but depends on the checks in place – including the design of the 'human in the loop'.

Thinking in terms of responsibility, legal researchers have argued that AI takes a decision-making role in a process if and only if the decision generated by AI is automatically implemented. The implementation mechanism may in fact involve a human, but the AI is still the decision-maker if a human agent automatically implements or waves through the decision generated by the AI without challenging it, for example if instructed to do so by their employer (Tjong Tjin Tai, 2022, page 3). One consequence of this point is that phrases such as 'human oversight' and 'human-in-the-loop' need very clear specification if they are to be meaningful. It is useful to note here, too, that the term 'automated decision-making', used in UK government guidance and documentation, includes "both solely automated decisions (no human judgement involved) and automated assisted decision-making (assisting human judgement)" (Cabinet Office, 2023, cited by Bhatnagar & Gajjar, 2024).

Human review is often cited as a method to reduce risk and improve accountability where AI is involved in decision-making. Besides the already-noted need to define what qualifies as a human-in-the-loop or human review in a given process, there is ongoing debate about whether human review should be included by default in all automated decision-making, or only for contested decisions (Bhatnagar & Gajjar, 2024, pages 11 to 12). Either way, Laux et al conclude that, overall, the empirical literature "contains ample support for humans remaining in the AI-supported decision loop" (Laux et al., 2024, page 27). Gajjar and Brione (2024) observe that "It remains unclear if it is technically possible to successfully render such systems safe and responsible without direct human oversight."

An important consideration for the design of processes integrating AI and human work is that human performance at checking may not equal their performance on the original task (Graydon & Lehman, 2025). This is particularly so when the materials to be checked are largely correct, and the process design permits or encourages the human involved to omit the thinking required by the original task. Achieving well-calibrated AI models would help in the design of effective decision-making processes that integrate AI and human review, by enabling effort to be directed to where it is most needed (Shen et al., 2024). For example, process designers could build in additional support and checking of AI outputs on the basis of accurate confidence estimates indicating where AI outputs were more likely to be inaccurate.

Section 2: Characteristics of high stakes marking

This section aims to highlight important characteristics of high stakes marking as an activity and a context. It considers marking tasks, cognitive activity by human markers, the ways in which academic expertise and academic judgment is input into the marking process, and the types of judgments that use of AI might supplement or replace.

Marker judgments

With human markers, the marking process is underpinned by the application of academic judgment[4] by a trained expert. Not all human judgments made during marking are the same, however, and academic judgments are not always involved in the same way or at the same point in the marking process. To think through the implications of using AI in a marking process, it is useful to think precisely about what human judgements we would be replacing or supplementing.

The process of marking has been extensively researched, primarily in the context of research into marker agreement (for example, Baird et al., 2013; Black et al., 2011; Bramley, 2009). The first point to make is that different marking tasks (that is, different items to be marked, which may vary in terms of item type, constructs assessed, and mark scheme) ask markers to make different types of judgement. A top-level distinction can be made between objective, points-based and levels-based mark schemes (Massey & Raikes, 2006). An objective mark scheme for an item is one that specifies a single correct answer (such as a single number, or a multiple-choice answer option); the marker must verify whether the student's response matches. A points-based mark scheme, meanwhile, asks the marker to judge whether specific credit-worthy elements are present in the student's response, and award points for the number that are located. A levels-based mark scheme, finally, asks the marker to judge which level descriptor in the mark scheme best describes the student's response. An analytic levels-based mark scheme separately describes levels of performance for different aspects of the response (for example, different assessment objectives) while a holistic levels-based scheme incorporates all aspects

[4] 'Academic judgment' is a subset of judgment involving specific expertise (in addition to the capabilities supporting judgment in general). In an article unpacking the idea of academic judgment, Hinchcliffe (2020) draws on Geach's philosophy of mind in which judging is construed as "the exercise of mental *concepts*". Academic judgment is then classed as "a cognitive activity that has … as its object the product of the cognitive activity of another (e.g. a student)" and could not be carried out by an entity or system that does not have cognitive activity or concepts.

into one holistic performance description per level (Ahmed & Pollitt, 2011). Having identified the level that best describes the student response, markers using a levels-based mark scheme must then make a finer-grained judgement to determine the exact mark to award within that level's mark range.

The nature of marker judgement in terms of psychological activity is contested (Brooks, 2012). Empirical research with markers of high stakes qualifications has identified matching, scanning, evaluating and scrutinising as cognitive marking strategies markers draw upon to varying extents in the process of reaching a mark (Suto & Greatorex, 2006, 2008). The matching strategy is invited by an objective mark scheme, and by a points-based mark scheme that specifies a visually recognisable pattern (for example, a word or pair of words, a letter, a number) for a specific and constrained response space. In this situation, the "examiner looks at a short answer line or another predetermined spot in the answer space for that question part, and simply compares what the candidate has written there with the correct response, making a judgement about whether or not the two match up" (Suto & Greatorex, 2008, page 220). The scanning strategy is often similar, except that the marker must scan the whole of the response space to determine whether the sought-for element is present or absent. The evaluating strategy is characterised as the marker considering a whole response or response part, and "evaluating its meaning in at least one of a number of ways" (page 223). This involves semantic processing, and may also involve considering the "structure, clarity, factual accuracy and logic" of the response, as well as "whether it meets the criteria, descriptors and targets specified in the mark scheme" (page 223). These strategies differ[5], and in particular, evaluating and scrutinising strategies "rely on subject knowledge, past marking and/or teaching experience, and on advice from the Principal Examiner" (Suto & Greatorex, 2006, page 10). While there are differing views on the psychological explanations, Brooks notes that confidence can be drawn from the fact that different studies have produced similar characterisations of marker judgement processes (Brooks, 2012, page 67).

The judgments and strategies required by a marking task are not directly determined by the item type or sophistication of the construct being assessed. However, there are relationships, not least because certain types of items are more frequently used to assess certain types of knowledge, skill and understanding, and certain mark scheme designs are more frequently used for certain item types (Ahmed & Pollitt,

[5] As Suto and Greatorex note, some psychologists reject the 'dual-processing' theory of judgment, and the idea that types of judgment qualitatively differ at a cognitive level. From this perspective, the slower, more effortful and reflective strategies of evaluating and scrutinising – which Suto and Greatorex explain as underpinned by 'System 2' thinking – are in fact collections of rapid 'System 1' type judgements, of which only the combined results are reported by markers (Suto & Greatorex, 2008, pages 224 to 225).

2011). The strategy employed by a marker may also change depending on whether they are marking a strong or weak student response. For example, where a matching strategy might be sufficient to mark a strong response that is well aligned with the mark scheme, scrutinising or evaluating may be necessary for a response that is less well aligned to the mark scheme, for example one that is “long or unexpected” (Suto & Greatorex, 2006, page 10).

Academic judgment specifically is not implemented in the same way across all marking tasks. At one extreme, for objectively marked items such as multiple-choice questions, academic judgment is necessary to construct a high-quality item and corresponding mark scheme, but not required in the moment of marking itself. The marking can be performed by any person or system that is able to ‘read’ the answer option indicated by the student, whether on paper or on screen, and has been provided with the right answer key. Similarly, for some items where the marking task is essentially a matching task, the major part of the academic expertise and judgment has been input at the earlier stages of designing the item, designing the mark scheme, and in standardisation (which includes refining the mark scheme, and should address upfront questions of judgment – such as which synonyms or alternative terms should be judged as equivalent to the specified correct answer). The point of marking itself requires relatively less academic judgment than these prior stages. By contrast, for other items such as an extended piece of writing marked by an analytic levels-based mark scheme, the marking task itself involves significant evaluation of quality or qualities. Hence, a significant degree of academic judgment is required at that point of marking itself, as well as at earlier stages.

Despite a good deal of research, there are limits to how finely we can describe marker cognition. This is of course not specific to marking; for human decision-making generally, it is widely understood that we cannot give complete and faithful cognitive accounts of the processes underlying decisions (Nisbett & Wilson, 1977; Zerilli et al., 2019). It is clear, however, that not all marking tasks ask markers to make the same type of judgment, and empirically, there are features of tasks and mark schemes (as well as of people and systems) that help predict the easiness or difficulty of a marking task (Black et al., 2011).

Marking as categorisation

The views of marking noted so far begin from the perspective of marking as judgment. There is evidence, however, for markers drawing on processes other than the application of written assessment criteria in the process of reaching a mark. In particular, there is evidence for markers drawing upon internalised understandings of levels of performance as reference points, and using heuristics (Brooks, 2012; Elliott, 2017). This aspect of marking is something that a human category learning (HCL) perspective incorporates more naturally. From an HCL perspective, marking with objective and points-based mark schemes corresponds to rules-based category learning, that is, where rules of belonging (to a category) can be straightforwardly verbalised (Bauer & Zapata-Rivera, 2020, page 17). ‘Information integration’ category learning, on the other hand, more accurately describes the process of marking with a rating scale or holistic levels-based scale, which the marker must learn to apply using a ‘best fit’ approach (2020, pages 17 to 18). There is an interesting question about the extent to which more complex levels-based mark schemes (implicit and explicit analytic schemes) do and do not resemble rating scales. A rating scale describes a single ordered set of levels of performance (typically, around 3 to 7 levels) and requires the marker to categorise the response to one of these levels (integrating complex information to do so). Implicit and explicit analytic mark schemes ask the marker to think about the level of performance demonstrated across multiple dimensions or criteria and to integrate these according to rules to reach the final mark – for an explicit analytic mark scheme, marking involves determining a ‘best fit’ category for multiple criteria then applying combination rules. The overall mark tariff for items marked by complex levels-based schemes may be high (for example, 40 marks), and a particular mark point can represent different profiles of performance across the different criteria, lessening the likelihood that mark points themselves represent coherent categories. Notably, the HCL perspective agrees that marking tasks are not all the same; they make different demands on markers in terms of the cognitive activity required for successful performance. Further, while a HCL perspective emphasises the similarities between human and machine learning, it is still not the case that the psychological activity underpinning human scoring is the same as automated scoring (Bauer & Zapata-Rivera, 2020).

Availability of understanding

For all the marking tasks described above, it is useful to make explicit that there is a fundamental fall-back or underpinning provided by human understanding. Whatever the strategies or heuristics employed by a human marker, and whether or not they are a very good or very weak marker, they have the capacity to understand the content of the sentences they are reading in a response (Williamson, 2013).

Assessment context

This report is focused on marking in the context of high stakes qualifications in England. These are largely curriculum-based qualifications, and though it is not always foregrounded, they are underpinned by a distinct logic. In particular, “Curriculum-based assessment has often been contrasted with psychometrics in the literature and simply termed ‘examinations’ or ‘assessment’” (Baird & Opposs, 2018, page 11). For curriculum-based assessments, the interest is in the knowledge and skills the student has gained through studying the curriculum (2018, page 12).

Baird and Black (2013) stress that for marking in a curriculum-based qualification, “There is no fully blossomed theory of learning and progress constituting a latent trait underlying our examination scores” (page 12). Rather, marking is understood as a values-based activity: “examiners assign scores on the basis of their values, arising from a number of sources, including disciplinary knowledge, understanding (from theory and practice) of learning of that discipline and educational and cultural values” (Baird & Black, 2013, pages 12 to 13). There may be no assumption of unidimensionality, and assessment constructs may be “very weakly described” (page 5). One consequence is that the kinds of validity evidence that are relevant and available, and the process of evidence accumulation to support a validity argument, may be different and more complex than in a psychometrics paradigm.

Another important aspect of the assessment context of focus is that items are almost never pre-tested. Exam papers for high stakes qualifications in England are written specifically for each assessment session, and items are not generally re-used.

Social context and purpose

High stakes qualifications in England can be understood both as examples of and components of implied social contracts. Focusing on the assessment aspect of qualifications, Baird et al describe the implied contract whereby “individuals submit themselves to normative, standardised testing on the promise of their results having currency in the labour and education markets” (Baird et al., 2024, page 6). Newton, meanwhile, emphasises the role of qualifications within the larger social contract of education, describing how “the implicit social contract rules that students will be celebrated and rewarded for the breadth and depth of the competence that they are able to acquire” – specifically, with “more, or better, opportunities” (Newton, 2023, page 231).

Maintaining public confidence matters for social contracts generally, since without it “expectations of fulfilment of the social contract are broken, bringing critique and calls for reparation” (Baird et al., 2024, page 6). Even more directly, however, qualifications (and other forms of currency) depend heavily on public confidence to hold their value. Ofqual and awarding organisations have important roles in ensuring that both public confidence and the value of qualifications are maintained.

Transparency is critical for assessment and qualifications in this context. Baird and colleagues note the importance of transparency around the content of the curriculum, and the tradition of publishing past exam papers in order that students have full visibility of what they will be assessed upon and how (Baird & Black, 2013; Baird & Opposs, 2018). Newton’s account reiterates this point:

> To be fit for a purpose of this sort, qualifications need to be designed so that their assessment procedure and learning outcomes are as transparent as possible, and so that teaching, learning and assessment arrangements follow the established 'rules of the game' without partiality, prejudice or blatant malpractice. (Newton, 2023, page 231).

Baird et al also stress the importance of transparency and shared understanding of qualification standards; a situation where stakeholders have different understandings "poses problems for the fulfilment of the contract and management of the system" (Baird et al., 2024, page 6).

Marking quality and properties sought

Ofqual's General Conditions state that the criteria differentiating candidates taking regulated qualifications should be accurately and consistently applied, in particular, applied consistently "regardless of the identity of the Assessor, Learner or Centre" (Ofqual, 2017, Condition H1.1). This formulation foregrounds the major aspects of validity and fairness in marking, namely, that students taking a qualification should be differentiated by the qualification's assessment criteria and not something else, and that the way in which the criteria are applied should not depend on the identity of the marker, the student or where they are taking the qualification.

A broad term that appears in much of the research on marking is 'marking quality', defined in previous Ofqual research to mean "the accuracy and reliability of marking" (Ofqual, 2014, page 5). The terms 'accuracy' and 'reliability' are especially common, but Bramley (2007, page 22) notes that "reliability, accuracy, agreement, association, consistency, consensus, concordance, correlation" are all used in discussions of marking quality, and not always related with precision to particular definitions or metrics. Core questions include "Is there a known 'correct' mark with which we are comparing a given mark or set of marks?" and "Where does this marking situation fall on the continuum from completely objective (e.g. multiple-choice item) to subjective (e.g. holistic high-tariff essay)?" (page 22). Where there is an objectively correct marking decision, it is straightforward to talk about the 'accuracy' of a marker in achieving this correct marking decision. Where there is subjectivity involved, there may be a range of legitimate marks (Ofqual, 2014). A single 'definitive mark' can be defined by taking the most experienced or senior examiner's mark (for example, the principal examiner's mark), but this has a different status to an objectively correct mark, and in these scenarios it is more appropriate to measure the 'agreement' of the marker (with the principal examiner) than 'accuracy'. Bramley (2007) avoids the term 'reliability' for a single marker, reserving it strictly for describing a set of marks, since the theoretical definition in classical test theory is the proportion of variance in marks explained by systematic variation in the test-takers, or equivalently, the ratio of true-score variance to observed score variance.

As suggested in the previous section, ‘transparency’ and ‘fairness’ – principles central to responsible AI frameworks – are particularly critical for high stakes marking, in the sense that they matter not just ethically, or for social good in a general sense, but functionally[6]. That is to say, transparency and fairness, and the perceptions of transparency and fairness, are important factors in high stakes marking fulfilling its task of upholding the meaning and value of high stakes qualifications as currency.

Newton offers an analysis that links the contractual understanding of high stakes qualifications to important observations about public views on marking quality. Specifically, Newton notes “that parents, students and members of the public more generally tend to be less tolerant of systematic, non-random error (bias) and therefore more tolerant of unsystematic, random error (unreliability)” (Newton, 2023, page 233). Newton’s explanation is that the social contract implicit in high stakes qualifications represents a collective decision to reject other means of distributing opportunities, such as bribery or nepotism. Newton argues that assessment errors which “appear to arise from partiality, or prejudice, or blatant malpractice” are for this reason likely to be viewed extremely negatively, because “Bias of this sort undermines the most fundamental purpose of a qualification system when viewed from the contractual perspective.” (page 234) The same is not true of marking unreliability, because this unreliability does not threaten the core contract’s rationale. In fact, Newton notes, “it is only slightly ironic to say that unreliability (unsystematic, random error) is almost fair by definition in the sense of demonstrating neither prejudice nor partiality” (page 234).

Practical context

High stakes marking takes place in a structured wider decision-making process. Markers take part in training and standardisation to align their judgements with the mark scheme and the assessment’s principal examiner, and during live marking, the marks given by human markers may be monitored statistically. This wider decision-making process can be understood as an illustration of the “social and institutional character” of many practical human decision-making processes, which are intended to mitigate against many of the known weaknesses of individual human cognition (Maclure, 2021, page 430).

The decision-making process has not been designed to achieve perfect marks in abstraction but to optimise marking within constrained timescales and affordability. For high stakes summative assessments, marking must usually take place in a

[6] The justice system, clearly, is another context in which transparency has a critical functional purpose and has been extensively theorised (Garrett & Rudin, 2023).

narrow window, and it is not considered efficient or feasible for all scripts to be double-marked or checked by a principal examiner. Consequently, the mark given by a human marker is not likely to be checked by a second marker unless selected for additional quality assurance checks during live marking, or as part of an investigation after a post-results enquiry.

A final practical aspect to note is that although examiners often mark responses onscreen, most student responses in Ofqual’s context are handwritten and scanned. Few of the high stakes qualifications regulated by Ofqual are currently on-screen assessments, and in paper-based assessments, students would only produce a digital text response as a result of an access arrangement (such as use of a word processor or scribe).

Summary

Section 2 set out to clarify important characteristics of high stakes marking as an activity and a context, a critical step in considering the use of AI. It highlighted differences in marking tasks, outlined what we understand about the cognitive activity of human markers undertaking these tasks, and noted differences in how academic judgement is input into overall marking processes. Perhaps most importantly, Section 2 considered the specific assessment context and social context of high stakes marking for regulated qualifications in England. Understanding high stakes marking in these terms is essential for informed future policy decisions.

Section 3: AI and automated mark generation

This section considers different automated and AI systems that can be used to generate marks. The use of current AI in marking is contextualised by setting it alongside existing methods of automated marking, which are still in use.

Approaches, models, features

The phrases 'AI marking, 'automated scoring', 'AI mark generation' and other similar terms do not have fixed technical meanings. They can refer to quite different methods for automatically generating marks or scores, which share some properties in common, but differ in other important ways. As noted in the introduction, research and development on automated marking technologies was traditionally divided into automated essay scoring (AES) and automated short answer grading (ASAG), according to the targeted item type.

Marking using AI systems, more generally referred to as automated scoring, has already proven successful in some high stakes assessments as well as in formative products (Richardson & Clesham, 2021). AES for example is well established in the Graduate Record Examination (GRE) and in Test of English as a Foreign Language (TOEFL) assessments.

At the highest level, human markers and AI marking systems both take a candidate response as an 'input' and generate a mark or score. The view taken in this working paper is that no current AI system can be described as exercising academic judgement, for the reasons discussed in Section 1 and 2. In brief, this report uses the ordinary meaning of 'judgment' that implies some deliberative aspect, and takes academic judgment to be a cognitive activity about the cognition of others. At the point of writing, current AI systems lack the required cognitive abilities. However, there are various ways in which an automated system can be designed and trained to approximate human judgment (if the judgment has been operationalised into a rules-based system) or generate scores that closely align with the outcomes of human judgments. In the first case, an AI system could be described as implementing the academic judgment of human experts. Where AI scoring systems are built upon DNN and LMM models, they inherit the 'black box' characteristic, making it very difficult to understand exactly how marks have been generated. The evidence we have, however, does not support the idea that current AI systems reach their outputs using the same means as humans, and it is probably helpful to avoid anthropomorphic language.

All automated marking systems base their mark generation on mathematical representations of the input response text. The descriptions of the automarking methods that follow all assume that student responses are available in a digital text format. As noted in Section 2, most high stakes marking in Ofqual's context currently involves handwritten student responses, meaning that a preliminary phase of handwriting recognition (such as using optical character recognition or OCR methods) would be necessary before automarking[7]. In the future, it is likely that more written assessments will be computer-based and more student responses therefore in digital text formats already. The current reality of handwritten exams and the details of handwriting recognition technology are important considerations and should not be overlooked, but they are not the focus of the current report. In what follows, we assume that student responses can be obtained in a digital text format.

In the simplest kind of automated marking (which we might nowadays refer to as a 'scoring engine', rather than AI), the system applies a determinative rule that contains the academic human judgement operationalised as a scoring rule. The major question for these systems is how well the rules capture the assessment construct. For an item requiring a single word response, specified in the mark scheme, this would mean looking at how effectively the scoring engine assigns one to everyone who wrote "oxygen" (for instance) and zero to everyone who did not write "oxygen". If there are no permissible synonyms, the rules will focus on implementing the principal examiner's judgments of acceptable and non-acceptable misspellings, and whether the chemical symbol for oxygen is acceptable.

In the systems we would more typically label AI, marks are generated using a trained machine learning model. The model is trained via supervised machine learning on human-marked responses: during this process, it learns to predict marks that align with the human markers, on the basis of 'features' or variables derived from mathematical representations of the response text. AI marking systems may use lists of manually selected text features (for example, variables developed on the basis of natural language processing (NLP) and linguistics research) or 'decide' on features itself (the case in deep learning approaches). Some features may be interpretable and explainable (for example, a manually selected measure of vocabulary complexity, or the length of the response text), but features can also include abstractions of the response text that are not interpretable by humans and do not

[7] Though not within the scope of the current report, it is worth noting that recent developments in handwriting recognition include use of AI, and LLMs specifically. The probabilistic approach involved in LLM-based methods can cause different errors than in older OCR methods, including preferring common words over exact transcription, "correcting" perceived errors, reordering information based on learned patterns, and producing different outputs for the same input due to sampling. The potential for these error types is especially relevant for assessment contexts, where we want to evaluate what students have actually written in their response – whether or not it appears likely or correct.

have any substantive meaning when viewed through human eyes (Bennett & Zhang, 2015). The way in which features are combined to predict human marks also varies significantly. A simple model may combine features in a simple logistic regression model, which is fully interpretable: a suitably trained person can understand how different features are being combined to produce the predicted mark. At the other extreme, features may be combined within a many-layered DNN, in which (as noted in section 1) it is not possible – even for experts – to grasp how features are combined within the model to produce its outputs. However complex, these AI systems have in common that they have been trained using machine learning on human-marked responses to generate marks that align with human-assigned marks.

Current foundation models can be used to generate marks in several ways. One approach involves taking a pre-trained LLM and fine-tuning it using human marked responses. For example, a BERT model can be fine-tuned on marked responses so that it 'learns' how those responses are classified into different score points assigned by humans, and can then classify new (unmarked) responses to a score point (for example, Ludwig et al., 2021). Alternatively, current AI systems can also be used to generate marks without training on human-marked responses at all. Instead, a generative pre-trained LLM (such as GPT-4) can be given a mark scheme to apply to response texts (for example, Henkel et al., 2024). It may also be given a few example marked responses as prompts, but the approach avoids building, training or fine-tuning a model specific to the assessment task being marked. The flexibility and anticipated cost-saving of such an approach is a very significant attraction (Tate et al., 2024).

Figure 1 shows a high-level view of 5 kinds of automated marking for text responses, focusing on the path from response text to mark generated. The text in red highlights the places where human academic judgment is input into the marking process.

Route A, at the top, represents a simple rule-based scoring engine route. Human academic judgment is encapsulated in the scoring rules, and the system applies these rules in a deterministic way.

Route B represents a feature-based machine learning approach. Supervised machine learning on human-marked responses is used to develop an algorithm that generates marks based on manually selected response features (for example, 1,500 NLP variables); the trained model assigns marks to input text in a deterministic way. Human academic judgement enters route B indirectly via the human-marked responses in the training set (which in turn should represent the judgement in the mark scheme). Human judgment is also input via the decision about which features to include and exclude from the model. In particular, one important decision is whether to include non-meaningful and construct-irrelevant feature variables (such as length of response) that may improve the accuracy of the model, in terms of generating marks that more closely align with human marking.

Route C represents a deep learning approach in which a DNN is trained (from scratch) using marked responses. Human judgment enters this route only indirectly, via the human-marked responses in the training set.

Route D represents a deep learning approach in which a pre-trained LLM is fine-tuned using marked responses. Human academic judgment enters this route indirectly via the human-marked responses in the training set used for fine-tuning.

Route E represents a deep learning approach with no training on human-marked responses. A generative AI system based on a pre-trained general purpose LLM (such as GPT-4) is given a mark scheme and/or a specific prompt requesting the generation of marks for response texts. Human academic judgement is included in the mark scheme and/or prompt. Directing the AI model to apply this does not mean that the mark scheme becomes incorporated into AI model itself.

There are also hybrid approaches that combine aspects of feature-engineering and deep learning. One approach is to use feature engineering to extract (possibly thousands of) known features and then input these into a deep learning neural network (for example, Kumar & Boulanger, 2020a). This offers more certainty about what is fed into the ‘black box’, but further work (using xAI) is needed to understand which of the features influence the marks generated, and to what extent, and how the ‘black box’ links those variables to the output marks. Another approach is to use deep learning to assist with the feature extraction and feature identification stage. The features identified this way can then become the inputs into a more traditional machine learning approach; for example, marks could be generated using a logistic model (fully interpretable), where the inputs are features developed through deep learning such as LLM embeddings.

Other hybrid approaches combine AI or automated systems with human marking by using the AI or automated system to identify where new responses match an existing response that has already been marked by a human. Known as similarity-based approaches, these have proven very successful for marking tasks where marks depend on the presence of specific concepts, and they have been extensively researched in ASAG. The methods used to determine similarity between student responses range from the simple matching of responses to a key or dictionary, to calculating similarity using a fine-tuned BERT model (Clauser et al., 2024, pages 211 to 215).

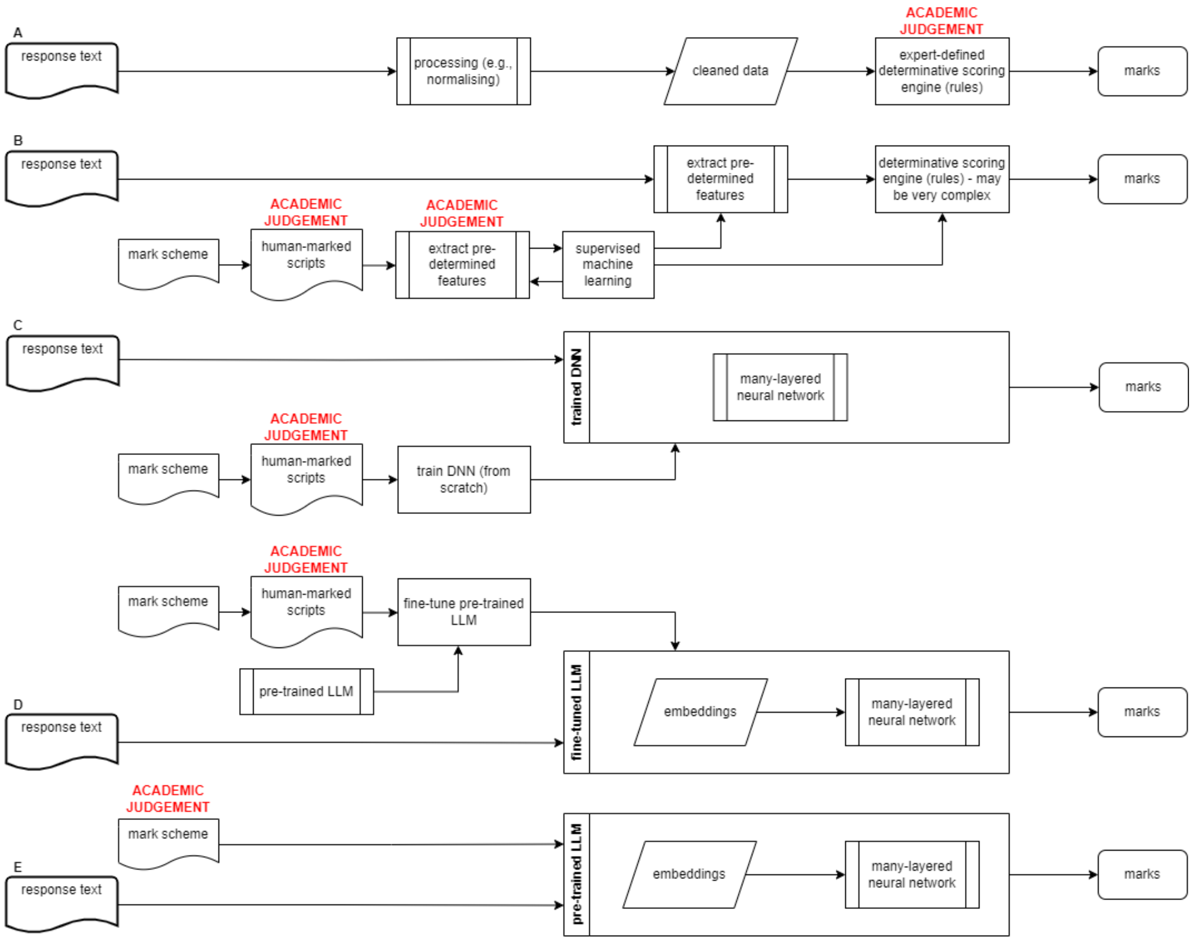

Figure 1: Types of automated and AI mark generation, and input from academic judgement.

Properties of different types of automated and AI marking

Figure 1 showed which of the different automated or AI marking systems incorporate DNNs or LLMs (routes C, D, E and hybrid approaches). This can be a useful dividing line due to the significant impact it has on the prospect of explainability for a marking system.

Another useful classification is whether the automated or AI marking system is general-purpose (also referred to as a ‘cross-prompt’ system), or specific to one assessment item (‘prompt-specific’ or ‘task-specific’). As already noted, the potential to use generative AI systems as general-purpose marking systems is a very significant attraction. For high stakes curriculum-based assessments, the relevance of general-purpose solutions is even greater, since the tasks set in exam papers are rarely re-used or pre-tested (and hence, task-specific responses are rarely available to use in training). However, general-purpose automated marking systems are not yet the norm. For short-answer questions, in particular, “most models used for ASAG still have in common that they are explicitly trained or fine-tuned for particular grading tasks” (Kortemeyer, 2024). In AES, recent research has found that prompt-specific models (trained to mark a particular assessment item) still tend to outperform general-purpose or cross-prompt solutions (Nam, 2025; Yang et al., 2024). In between general-purpose and task-specific automarking models are ‘domain adaptive’ models, which are trained on responses to multiple tasks in a related area (Nam, 2025). These models are more flexible and cost-effective than task-specific models, but also tend to achieve higher accuracy than general-purpose models (Nam, 2025).

Fairly clearly, the characteristics of high stakes marking mean that the reliability and robustness of any AI system used in marking will be important, as they can directly impact the reliability and validity of marks produced. The concept of reliability in marking is not the same as the concept of reliability in AI. To achieve reliability of marking, it is necessary to have both AI system reliability (that is, the system will continue to generate marks in the expected way, under the expected conditions of use, without changing or breaking) and AI system robustness (that is, the system can successfully handle inputs not present in the training data; in particular, can handle unexpected student responses, and not generate different marks for equally valid responses that differ in surface-level features).

It can also be helpful to classify automated and AI marking systems according to whether or not they are deterministic. For deterministic systems, the same input (that is, the same response text) always leads to the same output. This is true for rule-based systems (routes A and B), but this is not true for all AI marking systems, particularly those using generative AI to output marks (such as route E). This point may be very obvious to AI experts but less so to non-specialists, who may expect (based on experiences with earlier technologies) that automated systems will provide consistent, even if flawed, output. Earlier generations of AES systems, in particular, could be compared to human markers in these terms: “the most notable distinction between human and AES scoring is that with human scoring we have greater confidence in the *potential* for the construct to be well-represented but less confidence in the conscientiousness and consistency of scoring, while AES scoring may entail some inadequacies in construct representation but provide highly conscientious measures and complete consistency” (Williamson, 2013, page 174).

Table 2: Some properties that vary by automarking route.

	Route B	**Route C**	**Route D**	**Route E**
Type of AI/automarking system	‘Classical' feature-based	DNN trained on marked responses	Fine-tuned LLM	Pre-trained LLM genAI system
Mark generation fully interpretable (for experts)	Yes	No	No	No

Decisions based on construct-relevant features	Under the control of system designers	Unknown. Investigation required (for example, with xAI methods)	Unknown. Investigation required (for example, with xAI methods)	Unknown. Investigation required (for example, with xAI methods)
Consistency: same candidate response leads to same output every time	Yes	Yes	Not necessarily (but more deterministic than Route E)	Not necessarily
Reliability of marks generated	Empirical question	Empirical question	Empirical question	Empirical question
Training data fully specifiable	Yes	Yes	No	No
Training data under control of those building the scoring system	Yes - fully	Yes - fully	Data for fine-tuning - yes; data to train LLM - no.	No
Potential bias learned from human marking	Yes	Yes	Yes	No
Potential bias from other training data	No	No	Yes	Yes

In their work on the validity of AI use in marking, Casabianca et al (2024) group automarking approaches according to their assessment of their transparency and explainability, and training. Their definitions and rationales for grouping together are summarised in Table 3:

Table 3: Definitions and groupings of automarking approaches from Casabianca et al (2024).

Marking approach	**Transparency**	**Rationale for grouping**
Human marking	Most transparent	Baseline for comparison
Feature-based marking model using hand-crafted construct-related features	More transparent	Trained on human-marked responses, and features in model are interpretable.
1. Feature-based marking model using general linguistic features (for example, n-grams) 2. Marking model trained to predict human marks using LLM embeddings	Less transparent	Trained on human-marked responses, but features used in model are not readily interpretable for humans.
LLM-based generative AI, with prompting	Least transparent	No training on human-marked responses at all, no principles-based mechanism for generating marks

Integrating AI

Automated and AI marking systems can be implemented into an overall marking process in different ways. The ITC and ATP guidelines for technology-based assessment list 5 different approaches for combining automated and human marking (ITC & ATP, 2025, page 64):

1. Fully automated marking with no human review
2. Fully automated marking with some human review
3. Partially automated marking with some responses routed for human marking
4. Combined automated and human marking whereby each response receives a mark from both sources
5. Fully human marking

In their framework for the evaluation of automated marking, Williamson et al. (2012, page 5) also describe 5 routes, ranging from fully human marking to use of an automated marking system as the sole marker (the most liberal use of automated marking technology). In between, Williamson et al. list 3 approaches in which responses receive both a human and automated mark. All 3 can be read as more finely-specified subcategories of the fourth ITC and ATP approach:

1. **Automated quality control of human marking**
   The mark from an automated marking system is compared against the mark from a single human marker. A discrepancy of a certain magnitude means that the response is then sent to a second human marker.
2. **Equally weighted human and automated marks**
   The marks from a single human marker and the automated marking system are averaged or summed. Again, discrepancies over a certain magnitude could cause the response to be routed to a second human marker.
3. **Composite scores**
   The mark from a human marker is treated as an additional 'feature', then weighted and combined with the automated mark to produce a composite score.

To add to this list, it would of course also be possible to implement fully automated marking with full human review, an admittedly expensive approach. It would also be possible to implement an automated or AI marking system as part of quality control for human marking, but in slightly different ways than as described by Williamson et al here. For instance, a discrepancy of a certain magnitude could flag a marker for further standardisation training.

In all these cases, it is essential to make clear who holds accountability for the final mark awarded, the relative status of human and AI marking systems as decision-makers and checking tools, and to fully specify the role of any human intended to be a 'human in the loop' alongside AI mark generation. The extent to which different monitoring arrangements would be classified as human review is one part of this; for instance, to what extent would the monitoring of item parameters and statistical properties by expert psychometricians (for example, Cardwell et al., 2024) be considered human review of the automated marking of these items?

Table 4 lists some initial observations for the different approaches.

Table 4: Observations on different models of AI use in marking.

AI input	Observations
AI generates marks, human checks sample	Large burden on awarding organisation to prove validity, reliability, and lack of bias; lack of consensus so far on what would constitute sufficient evidence. Effectiveness of human checking needs ongoing monitoring If inappropriate mark identified (without a satisfactory explanation as to why), would a human re-mark of all responses be required? If inappropriate mark identified, and if a satisfactory explanation is achieved for why inappropriate mark was assigned, possible for human re-mark of the affected responses only– or re-mark by alternative AI? Affected responses would then be marked differently to other responses; unless an objective item, they would in practice be assessed with slightly different criteria.
AI generates marks, all responses reviewed by human	No efficiency gain Anchoring effect and/or risk of uncritical human review Disagreement should occasion review by second marker Potentially very helpful for research and development of automarking in our context; great deal of opportunity to learn from the resulting data.
AI generates marks in parallel to human marker, discrepancies flag a response for review	Avoids anchoring Useful additional QA / marker monitoring tool Reviewer(s) must be confident weighing merits and likely risks from human and AI-generated marks Ongoing monitoring of discrepancies required to establish whether certain subgroups systematically more affected (for example, due to different prevalence of certain error types, or construct-irrelevant language patterns)
AI marks most responses, a human marks certain flagged responses (for example, with features associated with lower AI marking accuracy)	Risk to fairness as responses would be marked differently, and unless an objective item, they would in practice be assessed with slightly different criteria; careful evaluation required to establish fairness not reduced by this approach; careful communication also required. Ongoing monitoring of flagging/routing mechanism required, both in terms of effectiveness (correctly identifying responses not suited to AI marking), and fairness across subgroups.
AI marks some assessment objectives and human marks others	Combination of AI and human marks needs careful design and evaluation to avoid undesirable artefacts (for example, double-crediting).

Section 4: Evaluating AI use in high stakes marking

As in other contexts, evaluation of AI use in high stakes marking has to weigh both potential benefits and risks (Weidinger et al., 2021, page 7). The principles of responsible AI frameworks are somewhat helpful but perhaps more so are the principled approaches already in existence for evaluating the use of automated marking systems.

This section outlines these existing and well-established approaches to evaluation, which address many of the major questions raised by the prospect of AI use in high stakes marking. Current AI systems do introduce new considerations, and alter some of the emphases in existing frameworks, but experts in the field have stated clearly that earlier work on validity arguments for automated marking systems remains relevant to today's current AI (Johnson & Zhang, 2024, page 7). Not least, this is because the earlier work was principles-based, and readily adaptable to changes in technology. The latest research in autoscoring validity demonstrates how existing frameworks can be built upon to evaluate use of current AI in marking on well-grounded principles (Casabianca et al., 2024; Lottridge et al., 2024).

This section begins by explaining the most relevant existing frameworks. It then goes on to address some particular considerations that arise or take on new dimensions when using current AI: these include both core measurement topics, as well as considerations relating to consequences and the wider social context.

Evaluating automated marking system use

Evaluating the use of automated marking systems in assessment requires an assessment validity perspective, for the straightforward reason that "scoring and validity cannot be meaningfully separated" (Bennett & Zhang, 2015, page 142).

Modern discussions of assessment validity broadly support the definition of validity given by the *Standards for Educational and Psychological Testing*[8], and argument-based approaches to validation (Kane, 2006, 2013). There is, however, variation in how different accounts propose collecting, organising and evaluating the relevant

[8] "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA & NCME, 2014, page 11). Ofqual's formulation is more explicit: "The validity of a particular qualification is the degree to which it is possible to measure whatever that qualification needs to measure by implementing its assessment procedure" (Newton, 2017, page 16).

evidence. As a reference point, Table 5 shows the framework designed by Shaw and colleagues for traditional curriculum-based examinations (Shaw et al., 2012; Shaw & Crisp, 2015). Evidence supporting validity and evidence identifying threats to validity should be assembled against each of the validation questions; weighing up the strength of evidence at each of these logical stages is then a structured way to approach the overall validity question.

Table 5: Revised framework for the argument of assessment validation (Shaw & Crisp, 2015, page 36).

Interpretive argument – Inference	**Interpretive argument – Warrant justifying inference**	**Validity argument – Validation questions**
Construct representation	Tasks elicit performances that represent the intended constructs	1. Do the tasks elicit performances that reflect the intended constructs?
Scoring	Scores/grades reflect the quality of performances on the assessment tasks	2. Are the scores/grades dependable measures of the intended constructs?
Generalisation	Scores/grades reflect likely performance on all possible relevant tasks	3. Do the tasks adequately sample the constructs that are set out as important within the syllabus?
Extrapolation	Scores/grades give an indication of likely wider performance	4. Do the constructs sampled give an indication of broader competence within and beyond the subject?
Decision-making	Scores/grades give an indication of likely success in further study or employment	5. Do scores/grades give an indication of success in further study or employment such that they can be used to make appropriate decisions?

Frameworks for automated scoring

The proposed use of an automated marking system is obviously relevant to the ‘scoring’ step of the interpretive argument in Table 5 and the corresponding validation question of whether scores are dependable measures of the intended constructs. However, it is necessary to go beyond this, because the use of automated scoring introduces validity challenges that are absent or manifest very differently in the case of human marking (Williamson et al., 2012). Besides construct under-representation and construct misrepresentation, Williamson et al list vulnerability to cheating, impact on test-taker behaviour and “score users’ interpretation and use of scores” as challenges in this category (page 4).

Williamson et al present a framework for evaluating the use of an automated marking system (Table 6), targeted particularly at use of an automated system as a second marker, where the marks will be combined with those from a human marker. Their framework is organised around the inference stages in an argument-based validity framework, though with terminology that differs from Table 5. Fairness is operationalised as “comparable validity for *identifiable* and *relevant* groups” (2012, page 4; following Xi, 2010), and is therefore a facet of the areas already listed in Table 6. It is incorporated into the framework by specifying that analyses for certain criteria (those asterisked) should be repeated at subgroup level to check for subgroup differences.

Table 6: Summary of the evaluation framework (Williamson et al., 2012, page 5, Table 1).

Emphasis area	**Corresponding inference in an argument-based validity framework**	**Guidelines and criteria**
Construct relevance and representation: evaluating the fit between the capability and the assessment	Explanation	Construct evaluation Task design Scoring rubric
Empirical performance – association with human scores	Evaluation	Human scoring process and score quality Agreement with human scores* Degradation from human-human agreement* Standardized mean score difference* Threshold for human adjudication Human intervention of automated scoring Evaluation at the task type and reported score level*
Empirical performance – association with independent measures	Extrapolation	Within-test relationships* External relationships* Relationship at the task type and reported score level*
Empirical performance – generalizability of scores	Generalization	Generalizability of scores across tasks and test forms* Prediction of human scores on an alternate form*
Score use and consequences – impact on decisions and consequences	Utilization	Impact on the accuracy of decisions* Claims and disclosures Consequences of using automated scoring

The next existing framework to note is the ETS best practice guide for scoring constructed-response items (McCaffrey et al., 2022). This framework is extremely comprehensive, and addresses the principles, standards and collection of evidence recommended for validating both human marking and automated marking of constructed responses. An outline of the framework, listing the questions to address and relevant forms of evidence, is provided in the Appendix of this report. The ETS framework is informed by the *Standards*, and work on validity and autoscoring by Bennett and Zhang (2015).

The final framework to specifically note here is the most recent, again developed by ETS researchers to address the validity of scoring constructed responses (Casabianca et al., 2024). This framework is presented as a table for the structured organisation of validity evidence, and it compares the potentially available and appropriate evidence for four different marking approaches:

1. A human marker or markers
2. Automated: construct feature-based
3. Automated: based on general linguistic features and/or embeddings
4. Automated: marks produced by generative AI

The framework uses the high level guidelines of the *Standards* (AERA, APA & NCME, 2014) but the detail has been developed by building upon the more concrete recommendations laid out by Bennet & Zhang (2015), McCaffrey et al (2022) and Williamson et al (2012). It is organised by evidence type: the *Standards*' 5 sources of validity evidence, plus 'Fairness'.

In addition to these frameworks, other highly relevant guidance can be found in the 'roadmap' for the validity of autoscoring detailed by Dorsey, Michaels and Ferrara (2024), and the guidelines on autoscoring of constructed response items within the ITC and ATP guidelines for technology-based assessment (ITC & ATP, 2025). Both can be found in the appendix to this report. The roadmap by Dorsey and colleagues is a practical guide designed to help ensure the necessary thinking and collection of evidence for validity arguments for the automated marking of essay questions and short-answer questions. It includes considerations that are new or altered by the characteristics of current AI. The ITC and ATP guidelines again have a practical focus, with clear recommendations.

The appendix contains an example by Williamson (2013) of the kind of reasoning involved in using these frameworks; the example lays out in detail the backing of the scoring warrant for use of a feature-based AES system. Another highly instructive example is the example given as a demonstration by Casabianca et al (2024), considering the use of using generative AI to mark a high stakes writing assessment. The example is lengthy and not repeated in this report, but highly recommended as a working example that illustrates the principled thinking applied to current AI.

The following sections summarise some foundational themes from these frameworks for evaluating autoscoring:

- the relationship to human marking
- human marking validity
- the significance of the assessment stakes and implementation model
- metrics and thresholds

- consequences

Relationship to human marks

At the outset, it is helpful to be precise about what relationship is claimed between the assessment construct, human marks, and the marks produced using an automated system. Different relationships can imply a different purpose for the automated marking system, and consequently “the nature and degree of evidence” needed in validation (Williamson, 2013, page 177). The potential roles of AES scores that Williamson lists are the following (page 177):

1. “Prediction. The automated score predicts the human score that would have been provided.
2. Complement. Automated scores represent some aspect of the construct of writing better than human raters, and other aspects less well than human raters.
3. Replication. Automated scores do exactly what human raters do, replicating both the positive aspects and the negative elements of human scoring.
4. Improvement. Automated scores represent the construct of writing better than human raters, having incorporated all of the desirable characteristics and eliminated the undesirable characteristics of human raters.”

Roles (1) and (3) differ, because role (1), prediction, does not specify how the automated scores or marks are produced. Any method that achieves good agreement rates with human markers meets the description, although this method used by the automated system may be entirely different to the means used by human markers, and may not be construct-relevant or even meaningful in human terms.

Where agreement between automated scores and human scores is the only validity evidence presented, the validity argument must by default be for a prediction model, since there is no basis for making any of the other 3 arguments (Williamson, 2013, page 178). The best practice guidance is clear that agreement with scores from human markers should not in fact be the only evidence provided; McCaffrey et al, for example, state unequivocally that validity evidence for automated scores “should include more” than the level of agreement with marks from human markers (McCaffrey et al., 2022, page 8).

The use of AES scores for high stakes assessment in combination with scores from human markers is commonly supported by an appeal to the complementary model. Typically, this “references the contributions of human raters to ensure full construct representation and the contributions of AES for specificity and consistency of scoring” (Williamson, 2013, page 178). Since the replication model means the continuation of both good and bad aspects of human marking, including the inconsistencies that AES could improve upon, the replication model is not usually cited as a goal (page 178). The improvement model, by contrast, is considered the ultimate AES goal. The ITC guidelines for technology-based assessment state that automated marking systems are typically designed to "emulate" human marking, which seems closest to the replication model (ITC & ATP, 2025, page 56).

Human marking validity

The fact that human marking is not completely understood is fully recognised by the autoscoring validity literature (for example, Bennett & Zhang, 2015, pages 151 to 154). Validity for automarking of constructed responses begins with the validity of human marks, since they are either the target for training the automated system, or the criterion measure for evaluating it, or both. In particular, the correspondence of automated scores to those from human markers is "by far" the most important source of evidence for the construct relevance of automated scores, and often the only source provided (McCaffrey et al., 2022, page 7) – meaning that the strength of validity evidence for the human marking often represents a ceiling for the possible strength of the argument for the automated marking system.

McCaffrey et al stress that strong evidence for the validity of human marking and strong correspondence between human and automated marks is still "not sufficient for concluding that automated scores support the desired inferences" (McCaffrey et al., 2022, page 7). Without further evidence, there is no reason for confidence that the 2 sets of marks are measuring a common construct.

Bennett and Zhang suggest that with a perfect validity argument for human marking (that is, all relevant validation questions evidenced fully, clearly and positively) and automated scores that agreed "close to perfectly" with the human scores, then the automated scores could be considered "exchangeable" with the human scores for that qualification. However, they go on to stress the unlikeliness of these conditions: "In reality, however, the answers to those questions are not very likely to be unequivocally 'Yes' …if the answers are known at all, precluding the use of human rating as the sole (or, arguably, even the primary) validation criterion" (Bennett & Zhang, 2015, page 153). Furthermore, even if the criteria were fully met at a given moment in time, there remains the need to monitor and update validity evidence for automarking over time. The evidence for validity at one moment in time could become less relevant due to changes that potentially occur to a qualification over its lifespan. These include updates to the curriculum, changes in the assessment design, changes in teaching and learning, and changes to the student cohort taking the qualification.

Stakes and implementation

Two major factors affecting the need for validity evidence are how the automated scores will be used, and their weight in consequential decisions. There is consensus that more robust evidence is needed when the stakes are higher – which is a product of both the stakes of the assessment, and the way automated marking is implemented. In particular, “more support is needed when automated scoring is the sole score than when it is employed as a check score” (Bennett & Zhang, 2015, page 162). Williamson et al spell out that this means “a higher burden of both the amount and quality of evidence” (2012, page 5). In particular, when an automated marking system is proposed as a sole marker, this “demands a much greater degree of congruence between the automated scoring features and the construct of interest” compared with its use in combination with human markers (Williamson et al., 2012, pages 6 to 7).

Bennett and Zhang (2015) also note that the level of supporting evidence required depends on the impact the automated score will have on a candidate’s final score, for instance, whether the automated score will be combined with a heavier-weighted assessment component that is objectively marked, or whether the automated score largely determines the final score and is separately reported. They describe it as “obvious” that there is greater need for “broad and deep evidentiary support” in the latter case compared to the former (Bennett & Zhang, 2015, page 162).

Metrics and thresholds

The agreement rates between human markers and automated marking systems and the accuracy of automated marking (for items with objectively correct marks) are commonly reported using simple statistical measures including:

- percentage agreement between marks from human and automated markers
- kappa
- quadratic weighted kappa (QWK)
- mean squared error (MSE)
- percent reduction in mean squared error (PRMSE)
- correlation coefficients

Some frameworks stress the importance of looking at score distributions, and the size and nature of marking discrepancies (including outliers), not just correlations or mean differences (for example, ITC & ATP, 2025, page 57). For evaluating fairness and bias in automated scoring, Johnson & Zhang (2024) demonstrated a useful and testable operational definition of fairness. Their straightforward approach compared the raw mean difference between automated and human marks, across different demographic groups. They also compared an adjusted mean score difference conditioned on (human) true score (following Johnson et al., 2022).

The ETS best practice guidelines state that the evaluation metrics listed (including measures of agreement with human markers) do not generally have a single threshold for acceptability. Instead, an appropriate threshold needs to be defined and justified depending on the assessment design, assessment construct, intended use of the constructed response scores and overall assessment scores, and potential unintended and negative consequences (McCaffrey et al., 2022, page 4).

Thresholds may be defined in relative terms (compared with human markers) or absolute terms. Schneider and colleagues, for example, state that in terms of accurate marking decisions for short-answer questions, “at least human-level performance should be aimed for in general” (Schneider et al., 2023, page 113). Williamson et al. (2012), meanwhile, suggest appropriate thresholds for use of an automated marking system to quality control human marking or to contribute scores alongside human marks in a high stakes assessment. They also note that “more stringent criteria may be advisable” if the automated scoring system is implemented as a sole marker (Williamson et al., 2012, page 9).

In terms of interpreting disagreements between marks, the ETS guidelines stress that discrepancies between automated marks and human marks are “not parallel” to the discrepancies between marks from 2 human markers. In the case of human marks, discrepancies indicate the “idiosyncratic judgments” of individual human markers. In the case of an automated system compared to human markers, however, the discrepancies include these idiosyncratic aspects in combination with, potentially, “systematic differences between human ratings and automated scores” (McCaffrey et al., 2022, pages 7 to 8). Systematic differences between human and automated marks are not just a technical issue − they can have implications for construct validity (discussed later). There can also be implications for fairness, if the systematic difference means that responses from certain subgroups of students are more likely to be relatively over- or under-rewarded in comparison with the marks they would receive from human markers (for example, due to greater sensitivity to spelling and grammar).

While evaluating autoscoring systems, it is important that test-taking conditions are consistent with the intended implementation model. If they are mismatched, evaluation metrics “may misrepresent the quality of the automated scores” (Williamson et al., 2012, page 9). Williamson et al state that this issue has “largely been ignored” in the literature, but judge this to be a mistake, since students are “very likely to adapt their test-taking behavior” if aware that their responses will be automatically marked. In particular, they may divert their energy to those aspects of the construct that the automated marking system is able to analyse (that is, ‘game’ the assessment) (Williamson et al., 2012, page 9).

Consequences

The evaluation frameworks are clear that the consequences of using automarking systems in a qualification should be systematically considered. For the cases of replacing a human marker with an automated marking system, or using an automated marking system as quality control for human marking, the following areas of consequences are specifically mentioned:

- possible changes to the nature of the construct assessed (due to use of the automarking system)
- possible changes to the reliability of assessment results
- perceptions of the assessment by users and stakeholders
- possible changes to the meaningfulness and usefulness of the reported results
- how the assessment results are interpreted and used for decision making
- how students prepare for the assessment
- teaching and learning of the knowledge, skills and understanding targeted by the assessment
- the feasibility of implementing the automarking system, especially noting the requirements of ongoing maintenance

(McCaffrey et al., 2022, page 4; Williamson et al., 2012, page 10)

Impact of (non) understanding

In the literature on validity of automated marking, there are relatively few direct discussions of understanding and its impact on validity. Dorsey et al argue that where there is an absence of human-level understanding of natural language in current AI systems, “we should ask ourselves, “how does this fundamentally limit our validity claims involving such models?” (Dorsey et al., 2024, page 481). This is an interesting point to raise, given that earlier automarking systems (for example, feature-based scoring engines) also lacked human-level understanding of natural language. Pack and colleagues, for instance, note that the fact that automated scoring engines do not (and cannot) “read essays, nor analyze them, as a human would” has invited doubt over their validity for many years (Pack et al., 2024, page 2).

Bennett and Zhang connect understanding to the extent of validity evidence required, and seem to argue that there is a limit on the extent of authority that can defensibly be given to AI and auto-scoring systems until understanding is present. They recommend:

> unless a broad base of validity evidence justifies the sole use of automated scoring, [awarding organisations should] incorporate well-trained and carefully monitored human raters into the process. The main reason for incorporating human raters is that human raters can understand what they read. Computers, at least as yet, cannot. (Bennett & Zhang, 2015, page 170)

Where the marking task requires evaluative judgements from human markers, the prospect of using AI and automated marking can be met with particular doubts from stakeholders (Wood, 2020). One reason may be because they qualitatively differ from other marking judgements: Bearman and colleagues, for example, argue that “Evaluative judgement takes on particular significance with respect to AI, because of the role that humans play as the arbiters of quality” (Bearman et al., 2024, page 894).

Specific validity considerations for use of current AI systems

This section aims to highlight specific considerations that arise or alter significantly when automated marking is carried out using DNN, LLM-based and generative AI approaches.

Transparency, explainability and interpretability

There is strong agreement from assessment experts that students should be informed when AI scoring systems are used, and that AI scoring systems ought to be explainable (Dorsey et al., 2024; ITC & ATP, 2025; Lottridge et al., 2023, page 21). For classical feature-based automated scoring systems, explanations of how scores are generated are possible, though they may be complex and challenging to communicate, and need to deal with the unwillingness of commercial stakeholders to disclose proprietary details of their models. Achieving explainability and interpretability for AI systems based on DNNs and LLMs is much more difficult, with practical consequences for users. Lottridge and colleagues give a good summary of the specific implications for assessment:

> transformers [i.e., DNN models with a transformer architecture, including LLMs] have “implicit features” learned during engine training …These features are not readily interpretable because they are not manually crafted; rather, they are

> model values generated after the engine training process and are low-dimensional representations of the language used by examinees. This makes describing how the engine produces scores difficult to explain and vet, except at a theoretical level. This is a validity concern and a logistical concern; if there are difficulties in matching human scoring, there is very little recourse in addressing this issue other than by hyperparameter tuning [i.e., altering the parameters that control the training process, such as the rate at which model weights are updated] or adding data. (Lottridge et al., 2024, page 499)

Casabianca and colleagues note specific consequences of opacity for the task of validation. The first is that, both for generative AI and automarking engines using LLM embeddings, there is greater reliance on the validity of human marking. For an automarking engine using LLM embeddings, the lack of interpretability of the model weights means validity of the marks produced "relies more heavily on the quality of the human ratings used to train the models" (Casabianca et al., 2024, page 15). For automarking using generative AI, there is again no way to decompose the model itself, and as a result "we must rely heavily on the link between human ratings and the LLM scores" (page 19). The second consequence of the opacity is a general increase in the need for validity evidence. The lack of explainability means "we must be even more diligent with fairness checks when using an LLM rather than other AI scoring models" (page 19). They also note that automarking using generative AI specifically is "distinctly different" in the sense that "there is no principled or explicit system of deriving a model or selecting features and the score is not based on a prediction of a human rating". As such, they conclude that "the types of validity evidence we might collect for feature-based AI scoring should be expanded for generative AI applications" (Casabianca et al., 2024, page 15).

Table 7 lists the additional pieces of validity evidence that Casabianca et al. suggest should be collected for automarking systems involving LLMs: either where marks are produced by a trained machine learning model using general linguistic features and LLM embeddings (column 2), or where marks are produced by prompting an LLM-based generative AI system (column 3). Table 7 excludes the validity evidence that is also recommended for classical feature-based scoring or human marking or has a direct parallel for these methods (for example, "expert review of scores and responses at all score levels").

A validity challenge specifically relevant to Ofqual's context is that for any AI marking system with a black box component, it would be difficult to show that mark generation depended on or took account of all admissible evidence (Condition H5.2). The assessment process could guarantee to input all allowed response evidence into the AI system but further work (such as investigations using post-hoc methods) would be needed to build an evidence base showing that all admissible evidence was being considered. For example, without additional evidence it would be impossible to know if the AI system was basing its score only on the first half of a particular student's response, disregarding the rest.

As explained in Section 2, transparency is critical for the successful functioning of high stakes curriculum-based qualifications. It is therefore especially important that transparency and explainability are considered from the perspectives of students and members of the public as well as those designing, implementing, and regulating AI systems. As in other contexts, it is possible that AI could achieve accurate and appropriate outcomes in high stakes marking with evidence that satisfies some groups (for example, assessment designers, psychometricians, regulators) while being fundamentally opaque to students and the users of qualifications. This point and the implications for trust are discussed further under 'Consequences'.

Before considering some possible solutions to the transparency and explainability challenges of AI, it is worth briefly noting here that some researchers consider the relationship between AI capabilities and non-explainability to be either misunderstood or frequently mis-represented. Specifically, some describe it as a "widely-held myth" that black box systems are always or usually more accurate, and that there is necessarily a trade-off between interpretability and accuracy (Garrett & Rudin, 2023; Rudin & Radin, 2019). In assessment specifically, it is interesting to note that 'black box' scoring systems do not always perform better than other automated marking systems (see Johnson & Zhang, 2024, pages 2 to 3).

Table 7: Additional validity evidence for scoring methods incorporating LLMs and generative AI, adapted from Casabianca et al. (2024, Table 1).

Type of validity evidence	General linguistic features-based and embeddings-based AI scores	Generative AI scores
Internal structure	Document the support for the selection of the scoring engine/feature set and/or LLM embeddings	Documentation on prompting strategy, in-context learning, and finetuning decisions. Document the support for the selection of the LLM Studies showing reproducibility and consistency of LLM scores over time. Documentation of the variability of LLM scores and how that affects reported score reliability. Analysis of chain-of-thought output from LLM to consistency with construct definition.
Relations to external evidence	(no additional evidence listed)	(none – same analyses as for other methods)
Response processes	(no additional evidence listed)	Expert evaluation of chain of thought feedback and comparison to expert annotation, if available.

Test content	(no additional evidence listed)	(no additional evidence listed)
Consequence of use	(no additional evidence listed)	(no additional evidence listed)
Fairness	Consult with subject matter expert on saliency results to understand if the scores are biased. Use saliency methods to understand differences in responses and assigned scores for different subgroups. Report on the data used to pre-train the LLM.	Consult with subject matter expert on saliency results to understand if the scores are biased. Use saliency methods to understand differences in responses and assigned scores for different subgroups. Report on established fairness metrics for the LLM. Report on the data used to pre-train the LLM.

Possible solution 1: avoid or reduce use of non-explainable systems

One possible solution is to limit the role of deep learning in automarking systems, either by insisting on the use of interpretable models only, or taking a hybrid approach that uses deep learning to construct features, but then incorporates those into a ‘classical’ scoring system (that is, a feature-based scoring engine) (Lottridge et al., 2023, page 22). This hybrid approach could still leave considerable non-interpretability (since the features constructed using deep learning would not necessarily be interpretable, even if their subsequent combination within a scoring engine was). There are, however, options that can improve the interpretability of features created this way. One is to train the deep learning model to predict human-understandable concepts first, and use these as the features for the subsequent feature-based model; models created this way are known as ‘Concept Bottleneck Models’ (see Koh et al., 2020). Another option is to investigate the features created through deep learning using explainable AI methods.

Possible solution 2: re-think understanding of validity

Another approach is to seek an achievable level of explainability that is less than full interpretability, but appropriate to the context of high stakes marking. Dorsey and Michaels, looking ahead in 2022, wrote that:

> If the field moves from automated scoring models that involve relatively straightforward regression models to those involving deep learning neural networks that calculate functions of functions across millions and billions of parameters, we must be prepared to identify the levels of transparency and

> explainability that will be sufficient to build robust validity arguments. (Dorsey & Michaels, 2022, pages 389 to 390).

More directly, Ferrara and Qunbar argued that AI scoring will not be possible if we maintain the current insistence that scoring be transparent and reflect only construct-relevant features: “we must accept validity arguments that leave room for degrees of opaqueness, limited explainability, and what we call for now indirect construct relevance” (2022, page 289). The amended view proposed is that scoring should be considered valid if (1) “significant portions of the engine’s features are explainable and construct relevant” and (2) “the automated scores produced by the engine are at least as accurate as human scoring” since the conjunction of these facts together mean that “evidence exists to support the intended interpretations of the automated scores in relation to the targeted construct” (Ferrara & Qunbar, 2022, page 292).

The main attraction of this approach is that it is achievable. Its challenges, meanwhile, include pinning down exactly what counts as “significant portions of the engine’s features”, and what counts as “explainable”. If consensus could be reached on a working definition of “explainable” that is suitable for the context, then “significant portion” could be defined in terms of the proportion of variance explained by explainable features[9]. Researchers in the field of explainable AI are working on methods that quantify the extent to which AI model outputs are explained, and if successful, these could form the basis of turning this approach into a working standard. This is not yet resolved. Dorsey et al. (2024, page 476) note that the field has indeed moved towards use of DNNs and DNN-based technology, but that agreeing the validity requirements remains work for future research.

Possible solution 3: develop use of explainable AI methods

Another approach to improving explainability is to empirically investigate the validity of marks after they have been generated, making use of techniques from explainable AI to try to demonstrate what aspects of student responses the AI system either has rewarded or tends to reward. This approach is the only one of the 3 outlined that is suitable for automated marking via pre-trained AI models, with no element of training on human-marked responses or feature-engineering.

As an example of validity reasoning of this type, Bulut and colleagues (2024, page 14) suggest a combination of saliency methods and qualitative analyses by subject

[9] At face value, Ferrara and Qunbar’s wording suggests it is about the number of explainable (versus non-explainable) features but this makes little sense, since the manner in which different features are combined into a scoring algorithm means that features can have widely varying levels of influence on the score assigned (Williamson, 2013, page 168).

matter experts to provide evidence about what AI-generated scores have rewarded. Saliency methods fall within the broader area of explainable AI, and include SHAP and LIME - a further demonstration in the context of AI marking can be found in Kumar and Boulanger (2020b), and good explanations of explainable AI methods are given by Molnar (2025). There are also other methods to be found in explainable AI. Even where explanations fall short of what users consider a full explanation (of how the AI system generated scores), explainable AI methods could still be used to produce helpful quantitative information about which features (which may or may not be interpretable) are more and less closely aligned with the AI-generated scores. These analyses can only be carried out once some student responses have been collected and marked (by the automarking system). If the automarking system was task-specific, this would mean a need for pre-testing the task.

It is important to note there is a possible disjunct between what assessment experts want or expect from 'explainable AI' and what it is able to deliver. Drawing on their research specifically addressing the psychometric considerations of AI use in marking, Lottridge et al. (2024) judged that the "field of explainability for deep neural networks is nascent and there are no solutions available yet that provide excellent results" (page 500). In particular, they noted that saliency values (which summarise the importance of individual features, such as particular words, in determining an AI system's output) were not yet capable of satisfactorily explaining AI-generated marks. Some explainable AI methods can be actively unhelpful (Bommasani et al., 2021; Watson, 2022). In particular, many approaches to AI explanation consist of separate models that are designed to explain how outputs were generated in the original AI model (for example, the black box model generating assessment scores). Unfortunately, due to being produced by separate models, the explanations generated "can lack faithfulness, by being unreliable and misleading about the causes of a behavior" – posing an obvious problem for those trying to establish how assessment marks have been generated (Bommasani et al., 2021, pages 125 to 126).

The reasoning reported by generative AI and 'chains of thought' produced by AI models, similarly, do not necessarily correspond to the process used to produce the model output (Turpin et al., 2023). In marking specifically, Naismith and colleagues note that "GPT-4's rationales may not accurately reflect the process used to assign the ratings", and that "rationales may present rationalizations for decisions actually grounded in biasing features" (Naismith et al, 2023, page 399). The authors remind readers that "Rationales should therefore not be treated as offering insight into the *process of generating* ratings" (page 399), even though their presentation as a method for demystifying the 'black box' of an AI scoring system seems to strongly invite this interpretation.

Possible solution 4: re-design assessments and automarking together

A final idea to note is that of re-designing assessments and automarking systems together, using the principles of evidence-centred design (ECD, Mislevy et al., 2003; Bennett, 2004). This approach ensures – by design – that the evidence for assuring the validity of an automarking approach is present and collected.

In essence, Bennett proposed re-designing assessments simultaneously with an associated expert-informed feature-based scoring engine. Human expert judgment remains within the overall marking process but is simply used differently. In particular, human judgment is used in defining the characteristics of good performance, the behaviours that evidence it, and the tasks that should elicit those behaviours, as well as the response features to include in marking and the direction of their relationship with the assessment construct (Bennett, 2004, page 3). As written, the ECD approach seems to assume an assessment process similar to our current high stakes assessment context. It also seems to implicitly exclude features that are not interpretable or do not have substantive meaning as viewed by humans.

Some commentators, however, expect that AI will reach its full assessment potential in more radically re-imagined assessments (Swiecki et al., 2022). Principled approaches to this kind of assessment already exist, not least in the form of computational psychometrics. Computational psychometrics is entirely aligned with ECD, and represents a natural extension of Bennett’s proposal for the validity of AES, since it incorporates machine learning and the benefits of data-driven methods, but within a strongly expert-driven, psychometric framework (von Davier et al., 2021). Further research exploring constructs (for example, within such a computational psychometrics framework) might identify and characterise new features, and thereby expand the range of features that experts understand to have substantive meaning.

The explainability ‘double standard’

Within the literature on AI use, some have argued that current AI systems are being held to an unfairly higher standard than humans when it comes to transparency and explainability (for example, Zerilli et al., 2019). There is agreement that transparency and explainability of automated decision making is socially important, but strong disagreement about what this should mean for regulation, what levels of explainability should be mandated, and how this relates to existing standards of transparency and explainability in decision making carried out by human actors. For instance, Zerilli and colleagues are clear that DNN-based AI systems are opaque, but argue that they are not substantially more opaque than humans. They state that “If human decision-making represents the gold standard for transparency, we think

AI can in some respects already be said to meet it" (2019, page 664). From this perspective, those voicing concerns about the lack of transparency in machine learning and AI models appear to be holding machine learning and AI models to an "unrealistically high standard" (2019, page 661).

This position has been named the "argument from the limitations of human reasoning" (Maclure, 2021, page 428). It is worth noting here because both this argument and critiques of this argument are relevant to transparency and explainability in marking. Maclure offers one such critique. Maclure's case is that the "argument from the limitations of human reasoning" is atomistic, with "crude" and unwarranted individualism and internalism. That is, the argument implicitly assumes that reasoning is best understood by focusing on the properties of individuals, and assumes that the mind can be understood by focusing only on facts internal to the brain or mind. Maclure argues that this approach is inadequate for comparing AI and human decision-making, because "the significance of the social and institutional dimensions of human reasoning and cooperation is lost from sight" (2021, page 430). Maclure's critique stresses that there are strong social and intersubjective aspects to reasoning, and we build institutional processes precisely to assist with consequential decision-making. Using the case of judicial decision-making as an illustration, Maclure explains:

> A particular judge, as I will try to show below, may be influenced by noise or biases, but judicial reasoning, as an institutionalized form of social reasoning, is designed with the purpose of neutralizing to the greatest extent possible the cognitive limitations of individual judges. (2021, page 430)

There are, clearly, parallels to the context of high stakes marking (see section 2). The cognitive processes of any individual marker are not fully transparent, and the human potential for fallibility is known. The details of high stakes marking processes (including the design of materials, training, standardisation, professional development and marker monitoring processes) are designed with these limitations in mind. As far as the available resources allow, the intention is to prevent, detect and resolve likely human errors and biases. The logical conclusion from Maclure's argument is that acknowledging the limitations of reasoning in individual humans "does not vindicate leniency with regard to the regulation of AI" (2021, page 430).

A related critique of the "argument from the limitations of human reasoning" is given by Adams (2023). Adams, like Maclure, observes that the "argument from the limitations of human reasoning" fails to account for social factors, and draws attention to the role of internalised social norms. Adams notes that those who assert equivalence between the opacity of AI and human decision-making "tend to focus on their shared inability to present reasons for their decisions *themselves*" (2023, page 621). However, there is no obvious reason why we should prefer or demand a self-explaining system: "in terms of intelligibility and accountability, it is difficult to tell why an explanation that we are given would be preferable to one that we find" (2023, page 621). Furthermore, Adams argues that "Epistemically, our having access to knowledge about social norms that explain human decision-making processes seems to make them straightforwardly more intelligible than their AI counterparts" (2023, page 621).

Ability to represent the construct

Dorsey and colleagues state outright that "In our view, it is generally accepted at this point that automated scoring engines typically evaluate only a small fraction of elements within construct definitions (e.g., writing)" (Dorsey et al., 2024, page 471). They specifically emphasise that "high correlations can be achieved even when humans and machines focus on different features" (2024, page 471). This directness seems to contrast with earlier statements in the literature, which often appeared more optimistic about the ability of automated scoring engines to achieve full or near-full construct representation. The validity approach described by Williamson (2013), for example, explained that the warrant "AES covers the construct appropriately" required collecting evidence to show that "AES measures the aspects of writing that, in the aggregate, embody the meaning of writing for the assessment" (2013, page 166). Earlier accounts seemed to imply that construct representation would be solved by further research and technological developments, as in this comment from Wood (2020): "the ability of an engine to adequately represent the important constructs of interest are also unique issues for AES algorithm developers and data scientists to address through advances in technology."

One of the most important points for the validity of automated scoring using DNN and LLM approaches is that construct representation, along with marking consistency and marking fairness, remains an empirical question. Current AI systems using DNN and LLM methods cannot 'ensure' that the assessment construct is represented by the scoring, or that non-construct-relevant features are not represented. Empirical investigation is required to find out what response characteristics have impacted the score generated, in the specific assessment context.

Reliability and robustness

As noted in section 1, LLMs exhibit what human users interpret as inconsistency and unreliability, including changes to outputs following trivial (to humans) changes in input. The 'black box' aspect of LLMs makes it more difficult to anticipate what changes will occur, to understand and explain the changes that do occur, and to take steps to avoid or mitigate them.

There are multiple causes of variability when using LLMs, and this poses a challenge for marking reliability (Casabianca et al., 2024; Henkel et al., 2024). A primary cause is the in-built element of randomness or non-determinism in an LLM's internal algorithms, and further variability for users arises when proprietary LLMs are updated by their owners; Tate et al. observe that currently, models may be "changed without warning or notice" (2024, page 3). Systematic marking investigations have confirmed that "all models are subject to fluctuations in their performance" (Pack et al., 2024, page 1).

Some qualities of the variability in current AI that are important for marking include:

1. That the 'errors' made by LLMs are not the same as the errors made by humans (Dentella et al., 2024). In marking, for example, an automated marking system based on generative AI may produce a non-numerical score such as a punctuation character (Johnson & Zhang, 2024, page 7).
2. That changes to AI system parameters and prompts may result in different changes than expected. Tate and colleagues, for instance, found that reducing the 'temperature' of the LLM-based scoring engine (which reduces the inherent randomness, and was expected to improve the consistency of marks generated) in fact led to more variable marking for some AI models and some populations, for reasons that are not clear (Tate et al., 2024, pages 3 to 4).

Another issue is that AI output may appear (more) inconsistent and unreliable due to expectations that implicitly assume too much similarity between AI and human understanding. In marking, this means that assessment users can be surprised by the effect of differences in responses that are trivial to human readers. This is particularly evident from research on adversarial examples, that is, examples specifically designed or modified to exploit (and expose) the weaknesses of an AI system, such as an essay that contains a paragraph repeated multiple times, or a short answer response where the word order has been shuffled (Aloisi, 2023, page. 102). Results may also be surprising in comparison with earlier feature-based automated scoring. Lottridge and colleagues have found in comparisons of LLM and 'classical' automated scoring that the "BERT engines are more susceptible than a classical engine to word shuffling, on-topic, non-sense essays using complex language, and off-topic essays, giving higher scores than warranted" (Lottridge et al., 2023, page 26). Overall, Lottridge et al. judge that the robustness of deep learning AI scoring systems in live scoring situations remains "unclear", with particular doubt over how they perform on "adversarial writing or on unusual writing", and research evidence suggesting that they are vulnerable to gaming behaviours (page 26).

As noted in section 1, achieving better-calibrated AI models (where the probabilities associated with model outputs accurately correspond to reality, rather than reflecting overconfidence or underconfidence) would significantly benefit users. For assessment designers and regulators, calibrated AI models would be very useful for accurately understanding the reliability of AI output. Calibration for current AI models is more challenging than for earlier machine learning methods, however, and is still a developing field.

Context specificity

Bennet and Zhang make a strong case that the validity of automated scoring is context specific. They write:

> for any given application, the validity argument for automated scoring should be particularized to the testing program in question (i.e., purpose and claims, task types, populations). In other words, that argument cannot be made *in general* for the scoring engine. That engine may not behave invariantly across task types or populations …; its performance may be simultaneously adequate for one purpose but not for others; and there will always be purposes, populations, and task types that differ significantly from past experience and for which sufficient evidence has not yet been gathered. (Bennett & Zhang, 2015, page 162)

Although Bennet and Zhang's account was written before the advent of current AI systems, the characteristics of current AI suggest that much of the same argument would apply. While the great attraction of applying the latest AI models to marking is precisely that they are general-purpose and do not have to be constructed anew for each question, the current extent of 'capability uncertainty' and potential for variability suggest that validity evidence for an AI scoring system in one assessment context should not be **assumed** to give backing to use in another assessment context.

More recent guidance is still clear that validity evidence should relate to the relevant student population (for example, McCaffrey et al., 2022). The question of when validity evidence must be task-specific, however, is less clear-cut. Like the question of whether the automated marking system itself is task-specific, this is an important practical question for Ofqual's context, because so few of the assessment items in this context are re-used or pre-tested.

The aim of general-purpose and domain-adaptive automarking systems is that they can mark a range of items, but their performance (in terms of any of the previously-mentioned metrics) may vary across items due to item-specific features (Schneider et al., 2023). If evidence from trialling, operational use, or dedicated generalisability studies shows that automarking performance is very stable across variations of a particular item type in a particular domain, this evidence can contribute to the case for using that automarking system to mark similar (untested) items of this type. The empirical evidence of generalisable performance would still need support from other forms of evidence (for example, expert analysis explaining why the automarking system should be expected to correctly assess the construct in untested items).

For automated short-answer grading, Schneider et al. argue that validation should in fact include collecting performance evidence on the specific items to be marked:

> Even if the autograder's performance has been assessed properly during development, i.e., using an (independent) large test set, additional validation for each usage, i.e., checking the autograder's performance for each exam, is recommended (Schneider et al., 2023, page 104)

The crux of their argument is that even where LLMs are fine-tuned for marking on a very large set of human-marked question-answer pairs, covering a wide range of questions, the performance of the LLM-based automarking system on new questions cannot be assumed similar to performance on other questions: "classifier accuracy should always be assessed explicitly for a set of novel question-answer pairs, since due to a large diversity of questions, there might be considerable differences in performance on training, validation, and test data" (Schneider et al., 2023, page 115). The authors are clear that their recommended approach to validation creates the need for human marking on the specific new items intended to be automarked. They acknowledge that this is a "major hurdle and significantly reduces the benefits of autograding" (pages 113 to 114).

Bias and fairness

As noted in section 1, machine learning models, including current AI, can encode and perpetuate forms of bias. The non-transparency of current AI, including particularly the data sources it is trained upon, make avoiding and mitigating potential bias in the models particularly difficult.

A typical way to investigate potential bias in pre-trained models is to examine "whether words are more closely associated with gender, ethnicity, or sexual orientation in ways that may impact downstream interpretation" and such biases in LLMs have been demonstrated (Lottridge et al., 2023, page 24). Lottridge and colleagues observe that more generally, the 'downstream' effects arising from LLM training data are not yet well understood; for example, it is unclear the extent to which biases would persist in a pre-trained LLM that is subsequently fine-tuned.

As a practical example in AI scoring, Johnson and Zhang (2024) investigated potential bias by comparing the ability of GPT-4o to score essays and predict candidates' race or ethnicity, considering the marks awarded and human-AI mark difference across different predicted race or ethnicity groups, and the accuracy of the race or ethnicity group prediction. Johnson and Zhang found differences according to predicted race or ethnicity group, concluding that "it appears that whatever aspects of the essay that are used to predict race/ethnicity are also associated with the aspects that cause fairness issues. Unfortunately, it is difficult to determine exactly what those aspects are" (Johnson & Zhang, 2024, page 7).

Human accountability

The principle of human accountability is a common thread in responsible AI guidelines, but some have pointed out that most studies have "paid little or no attention to meeting such demands when using and evaluating AI-based autograders" (Schneider et al., 2023, page 113). Since the chief benefit of AI use in marking is to efficiently generate a large volume of decisions (in contrast to idea generation, or other business applications of AI), there is an obvious challenge in how to include a human-in-the-loop that achieves meaningful oversight while retaining the intended efficiency of AI use and intended quality of decision-making.

Schneider et al (2023, page 114) state that a human should verify the automated scores, despite the fact this will "significantly" reduce the efficiency of the autoscoring system. A possible alternative model is to involve human experts only at a higher level of oversight, in monitoring item parameters and the statistical properties of autoscored items (for example, Cardwell et al., 2024, pages 33 to 35). In regulating AI use in high stakes marking, a key question is the extent to which different monitoring arrangements are considered as meaningfully including a human-in-the-loop, or human accountability, for consequential decisions. Casabianca et al speculate about how future assessments and automarking systems will evolve as AI technologies and AI use both continue to develop. Determining "the minimum amount of human involvement needed for validity evidence, especially in the high-stakes context" is noted as a specific question to address (2024, page 29).

Like transparency and explainability, this is an area where the public perception of accountability must also be taken into account. Public perception may not necessarily be in step with the legal and operational understandings of accountability. Furthermore, the implementation of AI or automated marking with human accountability may be perceived as complex in comparison with human marking, where the examiner is perceived to hold authority and joint accountability with the awarding body.

Verdicts on current AI use in marking

Recent studies exploring AI use in marking typically conclude that AI systems based on LLMs are impressive, but not yet suitable for use in high stakes marking without careful and detailed evaluation in the specific context:

- Tate et al concluded that "we would not recommend that AI be used for high-stakes summative purposes nor as a replacement for teachers' evaluation" (Tate et al., 2024, page 7).

- Johnson and Zhang concluded that “the black-box model in GPT-4o can better predict the ethnicity of a writer than to predict a score on an essay.” Their study specifically investigated using GPT-4o with ‘zero-shot’ prompting, that is, asking the AI model to score an essay using a prompt that contains no examples. The authors “found a number of issues that should make one think twice about zero-shot approaches for scoring essays without careful evaluation” (Johnson & Zhang, 2024, page 7).
- Kortemeyer concluded that “artificial intelligence is capable of providing meaningful formative feedback to completely open-ended assessment responses, but might not be ready yet for high-stakes summative assessment” (2024, page 2).
- Henkel et al concluded that “generative LLMs could be useful for low-stakes assessment tasks in real-world educational settings” based on finding that “GPT-4 can reach equivalent performance levels to expert human raters on ASAG tasks”, most notably, without fine-tuning the model and with little prompt engineering (Henkel et al., 2024, page 20).
- Casabianca and colleagues concluded from their validity study that “the evidence is too weak to support the use of these scores in operational score reporting unless they are used in combination with e-rater scores and/or human ratings” (Casabianca et al., 2024, page 25). In particular, Casabianca et al noted the need for better evidence on:
    - The reproducibility/reliability of GPT4, to allow informed decision making
    - Concordance with human marks, from a larger sample

Consequences

The frameworks for evaluation of AI and automarking systems note the importance of thinking through possible consequences of implementing an AI or automarking system in a particular qualification. This section highlights particular areas of consequence to consider, for the marking context set out in section 2.

Positive impacts on assessment

The use of AI in marking could lead to significant efficiency gains, in some implementations, potentially both lowering costs and reducing the time required to mark exams. These efficiency gains could unlock further benefits such as the possibility of faster release of results, and an increase in the accessibility of qualifications to schools and pupils.

The use of AI in marking also has the potential to improve the accuracy of marks awarded to students in some cases, ensuring that student performance is evaluated more precisely by removing the element of random error associated with human markers. Some forms of AI marking systems have the potential to reduce bias and construct-irrelevant variance, supporting fairer and more objective assessment.

A critical point for all of these potential advantages is that they require empirical investigation in the particular qualification context, and should not be assumed. Furthermore, research literature on AI use and automarking validity emphasises the need to evaluate whole processes rather than only isolated stages (Graydon & Lehman, 2025; McCaffrey et al., 2022, page 4) – the use of AI at one stage can have costs and effects elsewhere in an overall process (for example, on the effectiveness of human checking; on the cost of investigating marking errors). It is therefore critical that evaluations of AI use in marking consider the whole process. As an example, Casabianca and colleagues note that while using off-the-shelf generative AI to generate marks is itself low-effort, performing the validation work necessary to support the use of those marks for their intended purpose "is time consuming and costly" (2024, page 29). As a result, "using an off-the-shelf LLM for CR [constructed response] scoring may be less cost-effective than expected or hoped" (2024, page 29).

A final point is that the possible positive effects on assessment may need to be weighed against threats to validity, including a potential narrowing of construct representation. The properties of LLMs include characteristics that could also introduce inaccuracies and biases in marking: for example, the possible unknown biases rooted in the training data and pre-training process, and inherent randomness in output. These risks are not present in the same way for a (determinative) feature-based scoring engine.

Impact on assessment expertise within the system

Widespread implementation of AI in marking as a replacement for human marking could mean a large reduction in the employment of teachers and assessment specialists as examiners. There is a risk that this could decrease the total assessment expertise within the education system, and in particular, reduce the extent to which assessment expertise is integrated back into teaching and learning by practising teachers who are employed seasonally as examiners. From an operational perspective, a reduction in assessment expertise in the system would also create a vulnerability for the live assessment process. Maintaining and indeed increasing the ability of humans to critically assess the output of automated decision-making systems is noted in the literature as a safeguard "in case of errors or malfunctioning", particularly relevant for an overall process that would remain time pressured (Maclure, 2021, page 433).

To mitigate this risk, some other means would be required to ensure that large numbers of teachers gained regular assessment expertise. Without the necessity of paying teachers for their work as examiners, there would be a danger of this not happening or not being maintained (for example, if set up as professional development, it might be regarded as lower priority than today’s high stakes marking, and/or suffer loss of funding after initial enthusiasm).

Washback on education

Washback effects on teaching and learning could be positive, especially if AI generated scoring was implemented in a way that encouraged deeper reflection about understanding of assessment constructs, and in such a way that teachers were systematically included in this work. An indirect positive impact on education could also occur if teacher time was freed up by the use of AI to assess components (such as performances and coursework) that are currently marked by teachers.

Washback effects could also include some negative impacts. One area for concern is the motivation to ‘game’ the assessment process: students may develop inauthentic responses designed to capitalise on the behaviour of the AI or automated marking system. More subtly, this could include students learning to write responses in a particular style that is rewarded by the marking system but not part of the assessment construct (such as a particularly bland or flowery style). This concern is particularly relevant to high stakes assessment, which increases the need for ongoing monitoring since “even if automated scoring algorithms or automatic item generation are initially unbiased, increasing examinee awareness of scoring procedures or subdomain weights can lead to gaming” over the longer term (Ho, 2024).

Constructs and their social context

Evaluation frameworks for automarking note that use of an AI or automarking system could result in changes to the nature of the construct assessed (for example, McCaffrey et al., 2022). This is a particularly interesting area to consider for curriculum-based assessments, for the reasons set out in section 2. A curriculum-based assessment is not generally designed to measure distinct and fully-theorised latent abilities. The performance of students on the curriculum-based assessment is itself what is of interest, and marking for a curriculum-based assessment is a values-based activity that draws on disciplinary knowledge, markers’ understanding of learning in the relevant discipline or domain, educational values, and cultural values (Baird & Black, 2013, pages 12 to 13). If the humans currently involved in this marking process were replaced in large numbers by automated systems, it is worth asking how the constructs assessed by curriculum-based assessments would be understood – what they would be understood to mean, how that understanding would be shared and maintained, and how it might change. While the cognition of individual markers is not perfectly understood, markers for high stakes qualifications make up communities of practice embedded in the education system. The overall marking process is a social and institutional decision-making process operationalised by awarding bodies; the marks given currently have their meaning in relation to this overall social process.

While the validity literature specifically notes the risk of AI or automated marking changing the nature of the construct assessed (when change is not intended), a parallel consideration for curriculum-based assessments is that use of AI or automated marking systems could prevent the evolution of the construct assessed (where change is expected or intended). For example, an automarking system trained on marked responses to replicate human markers would capture a particular moment in time. Using this system for high stakes marking in later years would mean applying a system designed to replicate expert judgement from one moment in time to student responses produced in a different, later time. This consideration can be understood in terms of the more general responsible AI concerns of deployment bias and historical bias.

A further question about the impact on constructs is related to possible washback. What would be the effect on students of losing both the human audience for their written responses, and the knowledge that their mark is overseen by human authority? Thinking about how communication could be altered by the change in audience, Bennett and Zhang ask “What is the effect on the composer’s writing of knowing that the audience cannot, in any real sense, understand and react to it?” (Bennett & Zhang, 2015, page 161). Understanding the possible impact would require monitoring over time.

Perceptions and trust

The high stakes summative assessment context in England positions external assessments as an important part of the overall social contract of schooling, as described in Section 2. There are, currently, strong expectations that at ages 16 and 18, students from different walks of life all have their best performances judged by external experts. Thinking about high stakes qualifications in terms of social contracts strongly underscores the role of public confidence (Baird et al., 2024, page 16). This would not necessarily be damaged by using AI in high stakes marking. However, as demonstrated when assessment arrangements in England changed during the COVID-19 pandemic, changes to high stakes assessment that alter the assumed terms of the social contract (for example, through the use of algorithms, or teacher assessment) can cause strong reactions, and public confidence in how standards are set and applied must be kept in view. More generally, researchers in

AI governance have highlighted an imbalance in attention when looking at the relationship between AI use, trust and public life:

> While making AI trustworthy has garnered substantial political momentum, equal attention needs to be paid to AI's potential to erode the trustworthiness of public institutions and, with it, their own ability to produce trust in the population (Bodó, 2021). Without trust, the public sector risks losing support and compliance by citizens. (Laux et al., 2024, page 4)

In the context of high stakes curriculum-based qualifications, section 2 and earlier parts of this section stressed the critical importance of transparency. An implication of this is that the validity considerations of transparency, explainability and accountability must be considered from the perspective of students and the public and the possible implications for trust. An implementation of AI in high stakes marking could achieve high accuracy in marking, supported by evidence judged to be high quality and convincing to experts and decision makers, but fundamentally not be understandable to students and the public. It is also possible that the views of experts on the relative opacity of human and AI-supported marking would not be shared by non-experts. It would be important to monitor whether AI use in high stakes marking led to perceived changes in accountability for high stakes marks, compared with the current arrangement of examiner and awarding body responsibility.

Some commentary implies that the increasing use of AI in work and everyday life will necessarily increase its acceptability. Research into perceptions of AI has found a nuanced picture, in which attitudes are sensitive to precise context, perceived trustworthiness and usefulness, and the nature and proximity of different risks (Dreksler et al., 2025; Modhvadia et al., 2025). As Laux and colleagues stress, attention must be given to the risk of eroding trustworthiness. For high stakes curriculum-based qualifications, a marking process that is fundamentally not understandable to the public, or where the accountability for high stakes decisions is not very clear (with recourse to appeal), introduces risk to trust. It is important that trust in marking and qualifications remains robust, both when the system is under challenge and when normal operations are 'running smoothly'.

From a slightly different angle, the context of 2020 and 2021 grading has also been investigated in terms of the role of the implied social contract as a facet of assessment fairness (Crisp et al., 2024, pages 8 to 9). This work highlights the importance of students' "legitimate expectations", which may derive from either explicit promises or established practices (Nisbet & Shaw, 2019). Students' legitimate expectations currently include that the work they produce for high stakes summative assessment is marked by experts applying their academic judgement, who understand their work and the criteria to apply. An implementation of AI use in high stakes marking needs to take these expectations into account.

A related point to keep in mind is that expectations and understandings of AI developed through experiences with formative assessment tools could affect understanding about AI use in high stakes summative marking. For example, Khosravi and colleagues (2022, pages 14 to 15) cite a statement that accompanies feedback given to students by a classroom writing tool: “Remember, AcaWriter does not really understand your writing the way people do [...] If you think it got it wrong, that’s fine – now you’re thinking about more than spelling, grammar, and plagiarism.” This kind of messaging is more accurate than that implying human-like understanding in AI systems but should probably be kept in view when planning how to communicate and justify the use of AI systems in high stakes marking.

A final area of perceptions to note are those to do with possible washback effects on constructs (discussed in the previous subsection). A perception among students, teachers or the public that AI use in high stakes marking has altered assessment constructs could have impact independently of whether expert analysis concludes that washback has occurred.

Considerations not specific to marking

Data and information security

How would candidate work be protected and secured? Would it be used for training awarding organisation scoring systems, or general-purpose AI models? Would candidates have sufficient transparency, and the option to exercise choice over this? These concerns have been raised explicitly by some considering the use of LLMs in marking (Kortemeyer, 2023, 2024).

Environmental concerns

The training and use of current AI models is resource intensive. Energy consumption is probably the most high-profile aspect, and has motivated efforts to improve AI’s energy efficiency, measure the carbon footprint of AI models, and compare the energy required to fully complete tasks with and without AI use - to properly contextualise the energy use (for example, de Vries, 2023; Jiang et al., 2024). Other researchers have drawn attention to the water consumption of current AI technology, and argued for much greater transparency over AI water use and much greater consideration of this impact (Li et al., 2025). Li et al also stress the need to consider energy and water use holistically, since considered in isolation they may lead to opposite conclusions. For example, focusing solely on the carbon footprint of AI may suggest “follow[ing] the sun” (that is, carrying out intensive computation to maximise solar energy use), while focusing solely on water consumption would suggest "unfollow[ing] the sun" to avoid high temperature conditions and hence reduce the water needed for cooling.

Because the physical infrastructure supporting AI is located out of sight, and the benefits of AI use accrue most to those who are least affected by negative environmental impacts, there is a risk that the environmental costs of AI technology are insufficiently weighed in considering AI use (Bender et al., 2021). While it may not be required for regulation of qualifications, there should be awareness that environmental concerns are a factor to be weighed in decisions about whether AI use is appropriate, proportionate and justified. Previous research on the environmental impact of assessment for high stakes qualifications noted that exam boards in England have made commitments to reducing the carbon footprint of their operations, and have demonstrated numerous examples of good practice (PA Consulting et al., 2023).

Discussion

This working paper explored the principles of using AI in high stakes marking and the implications for our context, through considering the nature and capabilities of current AI, important characteristics of the marking context, different models for AI integration, and approaches to evaluating the acceptability of AI use in marking.

Section 1 outlined some of the important characteristics of current AI technologies with particular relevance to marking. While LLMs and generative AI systems based on LLM models can produce human-like outputs and perform complex task, they do not possess understanding in the human sense, and they do not replicate the cognitive processes of human thought. The distinction between reckoning and judgement can be particularly helpful here; it is important to recognise the capabilities of current AI while avoiding accidental or implicit assumptions about their functioning that are linked to anthropomorphism. The variability of AI outputs, and crucially their lack of transparency, complicate efforts to understand and apply current AI in high stakes contexts. Potential users currently require extensive empirical evidence to have confidence that AI is able to function in a given role, and to have an informed grasp of its reliability and accuracy in that context.

Section 2 explored high stakes marking in the context of England's curriculum-based qualifications. It emphasised that marking is not a uniform activity, but that different marking tasks ask markers to make different kinds of judgments, and that the academic judgment underpinning marking may be involved at different stages for different item types. This section outlined some of the research on marker cognition, and acknowledged that marker cognition is not fully understood. However, there is empirical evidence for the variability of marking tasks and their difficulty in different contexts, and this section also emphasised that high stakes marking takes place in a highly structured social apparatus – the overall marking process is much more than the decision-making of individual human experts. This section also noted that high stakes qualifications can usefully be understood in terms of social contracts, and that their value depends on public confidence in the fairness, transparency and legitimacy of assessment. Changes to the 'terms' or the experience of the contract – especially those involving automation – must be carefully considered in terms of their impact on trust, legitimacy and the meaning of marks awarded.

Phrases such as "AI marking system", "autoscoring system", and "automated AI marking" do not refer to a fixed technical model or approach. They can be used to describe different approaches and refer to different combinations or integrations of AI scoring with human marking. Section 3 outlined different approaches, and some important differences in their properties including their explainability, consistency, and possible causes of bias. Given recent developments in AI, there is particular interest in the use of AI marking systems based on LLMs, either as part of an automarking system fine-tuned on human-marked responses, or within a general-purpose generative AI system (applied using prompts). A significant attraction is the possibility of using a single model to generate marks for different tasks and populations, rather than building or fine-tuning a dedicated automated marking model for each marking task.

Grouping together AI and automarking systems according to their properties is helpful in identifying the trade-offs implied by different choices. For example, feature-based systems may offer greater explainability and control, but less flexibility, while generative AI systems may offer scalability and adaptability, but at the cost of explainability and potentially consistency of marking. For all methods based on DNNs (including LLMS), it is more difficult to determine how and why the model has generated particular marks, which presents a problem for marking validity for all the reasons laid out in section 4. It also presents challenges for operational use, since without explainability, it is difficult to control, anticipate, diagnose, or mitigate any problems with the marking system. A further challenge that arises with the use of pre-trained LLMs is that the training data is typically not fully specified, but remains opaque to users including any assessment experts designing scoring systems. It is well documented that LLMs (like machine learning models more generally) can encode and perpetuate biases found in their training data, and in the case of LLMs, it is not yet known what impacts the training data might cause ‘downstream’ (for example, in terms of how these models carry out marking tasks).

An important question for AI use in marking is how to evaluate the acceptability of AI as a supplement or replacement for marking by human judgement. This paper argues for building upon existing approaches to the validity of automated marking systems, as explained and outlined in section 4. Existing validity frameworks and the associated research remain highly relevant and offer structured approaches for evaluating the validity of AI use in high stakes marking. These approaches do not expect or require that AI systems have human-like understanding or that they generate marks in the same way as human markers. Rather, they involve systematically gathering evidence to support (or refute) the argument that the marks generated by the automated system allow the intended uses and interpretations of the qualification results. Both the research and industry guidance in this area stress that agreement with human marks is not sufficient on its own, and that construct representation, fairness, and consequences must also be considered.

Both responsible AI principles and assessment validity frameworks point towards a need to understand how and why AI systems generate the marks that they do in a particular assessment scenario. At present, this requires a substantial amount of empirical investigation and compilation of evidence for AI scoring based on DNNs or LLMs. Current AI capabilities, robustness and reliability are not sufficiently well understood to (in themselves) give confidence that the performance seen in one assessment context will transfer to another. In assessments where an AI marking system can be re-used extensively to test the same performance skill in the same population (for example, in English language testing), there is a far better return on investment in producing and compiling the evidence base to demonstrate the validity of AI-generated marks – as there was for developing earlier generations of automated marking systems. For this reason, AI use in marking may continue, for now, to be more viable in some qualifications than others. It is also important to acknowledge that even with a suitable evidence base, technological developments may enable more accurate AI and automarking systems at the cost of lower transparency and explainability, posing a risk to trust in marking.

The principle that humans should retain oversight of and responsibility for consequential decisions is a point of consensus found in both responsible AI guidance and law. This principle is not always foregrounded in the research into AI marking. There is a clear challenge in determining how to include meaningful human accountability and oversight while retaining the efficiency gains of AI use. Possible models include parallel or second marking by a human marker (which would significantly reduce efficiency) or involving human experts only at a higher level of oversight, for example in monitoring the statistical properties of autoscored items. A key question in regulating the acceptability of AI use in high stakes marking is the extent to which different monitoring arrangements should be considered as meaningfully achieving human oversight and accountability for consequential decisions.

Summary of working position

There are positive and welcome potential uses of AI to improve high stakes marking. These include use of AI as a quality assurance tool for marking (as part of a combination of measures) and to support the training of markers.

Ofqual’s General Conditions specify that students taking a regulated qualification in England should be differentiated by criteria that are understood by assessors, accurately applied, and consistently applied. The use of AI as the sole mechanism for determining a student’s mark would not comply with Ofqual’s current regulations, as previously explained (Ofqual, 2024).

The use of AI-generated marks is likely to continue to develop and become more acceptable in contexts such as formative and low-stakes assessments. There are also principled arguments that can be made for accepting AI-generated marks in high stakes assessments, in the right conditions and with the right safeguards in place. Assessment validity frameworks exist and are well-placed to support the design, documentation and evaluation of valid implementations of AI and automated marking in high stakes assessment. Our understanding is that current evidence does not support an overall or general case for the validity of AI marking in high stakes qualifications, and at present points to the need for context-specific evidence (that is, specific to the particular qualification, in terms of addressing the relevant candidate population, constructs and item types that the AI marking system will encounter). The necessary empirical evidence is not yet widely available for high stakes qualifications. However, this does not mean such evidence could not be produced and compiled in future, especially as technologies continue to develop.

The core principle running through this working report is the need for clarity if valid use of AI in high stakes marking is to be achieved. On the qualification side this includes clear understanding of the qualification's purpose; the intended uses and interpretations of results; and the constructs being assessed. On the AI side this includes the properties (strengths, limitations and risks) of the proposed AI or automated system; how this system is proposed to be implemented in the overall marking process, including how its role relates to the tasks and responsibilities of humans in the marking process; and a clear rationale explaining how this system and its implementation will support a valid marking process for the qualification.

The automarking validity frameworks explored in section 4 and the appendix offer constructive examples of the types of evidence and organisation of evidence that could back the valid use of AI or automarking systems in high stakes marking. As a high level summary of the important areas to think through, Table 8 lists the key recommendations from the ITP and ATP's guidance for automated marking of constructed responses (whether as a sole marker, joint marker, or as part of quality assurance), and the logical steps that could parallel these for other proposed uses of AI in a high stakes marking process (where the AI is not used to generate a mark).

Table 8: High level guidance for valid implementation of AI or automated marking systems, adapted from the ITC and ATP guidelines (ITC & ATP, 2025, pages 62 to 64).

	AI or automated system generating marks	*AI or automated system with other proposed role in marking process*
4.9	The rationale for using automated marking should be clearly articulated and appropriate for the qualification.	The rationale for use of the AI or automated system should be clearly articulated and appropriate for the qualification.
4.10	The automated marking system design should appropriately reflect the constructs assessed.	The capabilities and proposed implementation of the AI or automated system should appropriately reflect the constructs assessed.
4.11	If an AI marking system requires training, it should be trained on a representative sample of responses that have been human marked with the highest level of quality.	If the AI or automated system requires training, it should be trained on a high-quality representative sample of responses.
4.12	The validity, reliability, and fairness of automated marking should be evaluated using sound methodological and statistical approaches and clear evaluation criteria.	The validity, reliability, and fairness of marks resulting from the overall marking process (with the AI or automated system implemented) should be evaluated using

		sound methodological and statistical approaches and clear evaluation criteria.
4.13	The approach for using automated marking or human marking should be based upon the marking quality of each method and consistent with the goals and stakes of the qualification.	The decision to use the AI or automated system in the overall marking process should be based upon the marking quality with and without its implementation and consistent with the goals and stakes of the qualification.
4.14	A well-defined process for reviewing automated marking performance during and after assessment series should be developed, documented, and implemented. There should also be a process in place for handling errors or disruptions.	A well-defined process for reviewing the impact of the AI or automated system during and after assessment series should be developed, documented, and implemented. There should also be a process in place for handling errors or disruptions.
4.15	For AI-based marking systems, algorithmic predictions, recommendations, outcomes, and prescriptive actions should be derived via transparent, ethical, and bias-free methods that can be explained and evaluated by internal and external experts or expert systems.	For AI-based systems, algorithmic predictions, recommendations, outcomes, and prescriptive actions should be derived via transparent, ethical, and bias-free methods that can be explained and evaluated by internal and external experts or expert systems.

References

Adams, J. (2023). Defending explicability as a principle for the ethics of artificial intelligence in medicine. *Medicine, Health Care and Philosophy*, *26*(4), 615–623. https://doi.org/10.1007/s11019-023-10175-7

AERA, APA & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice*, *18*(3), Article 3. https://doi.org/10.1080/0969594X.2010.546775

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, *99*, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Aloisi, C. (2023). The future of standardised assessment: Validity and trust in algorithms for assessment and scoring. *European Journal of Education*, *58*(1), 98–110. https://doi.org/10.1111/ejed.12542

Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., Rajagopal, G. R., Sloan, A., Tudrej, T., Ture, F., Wu, Z., Xu, L., & Baldwin, B. (2025). *Non-Determinism of 'Deterministic' LLM Settings* (No. arXiv:2408.04667). arXiv. https://doi.org/10.48550/arXiv.2408.04667

Baird, J.-A., & Black, P. (2013). Test theories, educational priorities and reliability of public examinations in England. *Research Papers in Education*, *28*(1), 5–21. https://doi.org/10.1080/02671522.2012.754224

Baird, J.-A., Godfrey-Faussett, T., Allan, S., MacIntosh, E., Hutchinson, C., & Wiseman-Orr, L. (2024). Standards as a social contract in curriculum-based qualifications: Stakeholder views in Scotland. *Cambridge Journal of Education*, 1–20. https://doi.org/10.1080/0305764X.2024.2377965

Baird, J.-A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability* (No. Ofqual/13/5261). https://assets.publishing.service.gov.uk/media/5a7e4fa5e5274a2e8ab4732d/2013-01-21-marker-effects-and-examination-reliability.pdf

Baird, J.-A., & Opposs, D. (2018). The Standard Setting Project: Assessment paradigms. In J. Baird, T. Isaacs, D. Opposs, & L. Gray (Eds.), *Examination Standards: How measures and meanings differ around the world* (pp. 2–25). UCL Institution of Education Press.

Bauer, M., & Zapata-Rivera, D. (2020). Cognitive Foundations of Automated Scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of Automated Scoring: Theory into Practice* (pp. 13–28). Chapman & Hall / CRC.

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, *49*(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots*. 610–623. https://doi.org/10.1145/3442188.3445922

Bennett, R. E. (2004). *Moving the Field Forward: Some Thoughts on Validity and Automated Scoring* (ETS Research Memorandum No. RM-04-01; pp. 414–423).

Bennett, R. E., & Zhang, M. (2015). Validity and Automated Scoring. In F. Drasgow (Ed.), *Technology and Testing* (pp. 142–173). Routledge.

Bhatnagar, A., & Gajjar, D. (2024, January 9). *Policy implications of artificial intelligence (AI) [POST-PN-0708]*.

Black, B., Suto, W. M. I., & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy & Practice*, *18*(3), Article 3. https://doi.org/10.1080/0969594x.2011.555328

Bodó, B. (2021). Mediated trust: A theoretical framework to address the trustworthiness of technological trust mediators. *New Media & Society*, *23*(9), 2668–2690.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). *On the Opportunities and Risks of Foundation Models*. https://crfm.stanford.edu/assets/report.pdf

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., & Liang, P. (2025). The 2023 Foundation Model Transparency Index. *Transactions on Machine Learning Research*, *February 2025*. https://openreview.net/forum?id=x6fXnsM9Ez

Bottazzi Grifoni, E., & Ferrario, R. (2025). The bewitching AI: The Illusion of Communication with Large Language Models. *Philosophy & Technology*, *38*(2). https://doi.org/10.1007/s13347-025-00893-6

Bramley, T. (2007). Quantifying marker agreement: Terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, *4*, 22–28.

Bramley, T. (2009). Mark scheme features associated with different levels of marker agreement. *Research Matters: A Cambridge Assessment Publication*, *8*, 16–23. https://doi.org/10.17863/CAM.100494

Brooks, V. (2012). Marking as judgment. *Research Papers in Education*, *27*(1), Article 1. https://doi.org/10.1080/02671520903331008

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C. M., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). *The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges* (No. arXiv:2406.18900). arXiv. http://arxiv.org/abs/2406.18900

Cabinet Office. (2023, November). *Ethics, Transparency and Accountability Framework for Automated Decision-Making [Guidance]*. GOV.UK. https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making

Cardwell, R. L., Naismith, B., LaFlair, G. T., & Nydick, S. W. (2024). *Duolingo English Test: Technical Manual*. Duolingo. https://duolingo-papers.s3.amazonaws.com/other/technical_manual.pdf

Casabianca, J. M., McCaffrey, D. F., Johnson, M. S., Alper, N., & Zubenko, V. (2024). Validity Arguments for Constructed Response Scoring Using Generative Artificial Intelligence Applications. *Eprint arXiv*. https://doi.org/10.48550/arXiv.2501.02334

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S., Leike, J., Kaplan, J., & Perez, E. (2025). *Reasoning Models Don't Always Say What They Think*. Anthropic. https://doi.org/10.48550/arXiv.2505.05410

Clauser, B. E., Yaneva, V., Baldwin, P., An Ha, L., & Mee, J. (2024). Automated Scoring of Short-Answer Questions: A Progress Report. *Applied Measurement in Education*, *37*(3), 209–224. https://doi.org/10.1080/08957347.2024.2386945

Crisp, V., Elliott, G., Walland, E., & Chambers, L. (2024). A structured discussion of the fairness of GCSE and A level grades in England in summer 2020 and 2021. *Research Papers in Education*, 1–28. https://doi.org/10.1080/02671522.2024.2318046

de Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, *7*(10), 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004

Dentella, V., Günther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, *14*(1), 28083. https://doi.org/10.1038/s41598-024-79531-8

Dorsey, D. W., & Michaels, H. R. (2022). Validity Arguments Meet Artificial Intelligence in Innovative Educational Assessment: A Discussion and Look

Forward. *Journal of Educational Measurement*, *59*(3), 389–394. https://doi.org/10.1111/jedm.12330

Dorsey, D. W., Michaels, H. R., & Ferrara, S. (2024). Validity argument roadmap for automated scoring. In M. D. Shermis & J. Wilson (Eds.), *The Routledge International Handbook of Automated Essay Evaluation* (1st ed., pp. 469–483). Routledge. https://doi.org/10.4324/9781003397618

Dreksler, N., Law, H., Ahn, C., Schiff, D. S., Schiff, K. J., & Peskowitz, Z. (2025). *What Does the Public Think About AI?* Centre for the Governance of AI. https://dx.doi.org/10.2139/ssrn.5108572

Durt, C. (2022). Artificial Intelligence and Its Integration into the Human Lifeworld. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Eds.), *The Cambridge Handbook of Responsible Artificial Intelligence* (1st ed., pp. 67–82). Cambridge University Press. https://doi.org/10.1017/9781009207898.007

Elliott, V. (2017). What does a good one look like? Marking A-level English scripts in relation to others. *English in Education*, *51*(1), Article 1. https://doi.org/10.1111/eie.12133

Ferrara, S., & Qunbar, S. (2022). Validity Arguments for AI-Based Automated Scores: Essay Scoring as an Illustration. *Journal of Educational Measurement*, *59*(3), 288–313.

Figueras, C., Farazouli, A., Cerratto Pargman, T., McGrath, C., & Rossitto, C. (2025). Promises and breakages of automated grading systems: A qualitative study in computer science education. *Education Inquiry*, 1–22. https://doi.org/10.1080/20004508.2025.2464996

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*, 201–224. https://doi.org/10.1002/berj.4069

Floridi, L. (2023). AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology*, *36*(1), Article 1. https://doi.org/10.1007/s13347-023-00621-y

Foltz, P. W., Yan, D., & Rupp, A. A. (2020). The past, present, and future of automated scoring. In *Handbook of automated scoring* (pp. 1–10). Chapman and Hall/CRC.

Gajjar, D., & Brione, P. (2024). *Artificial intelligence: Ethics, governance and regulation [POST briefing]*. Parliamentary Office of Science and Technology. https://post.parliament.uk/artificial-intelligence-ethics-governance-and-regulation/

Garrett, B. L., & Rudin, C. (2023). The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice. *Cornell Law Review*, *202*(03). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4275661

Graydon, M. S., & Lehman, S. M. (2025). *Examining Proposed Uses of LLMs to Produce or Assess Assurance Arguments* (No. NASA/TM–20250001849; NASA STI Program Report Series). National Aeronautics and Space Administration.

Haller, S., Aldea, A., Seifert, C., & Strisciuglio, N. (2022). *Survey on Automated Short Answer Grading with Deep Learning: From Word Embeddings to Transformers* (No. arXiv:2204.03503). arXiv. http://arxiv.org/abs/2204.03503

Henkel, O., Hills, L., Roberts, B., & McGrane, J. (2024). Can LLMs Grade Open Response Reading Comprehension Questions? An Empirical Study Using the ROARs Dataset. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00431-z

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, *26*(2), 38. https://doi.org/10.1007/s10676-024-09775-5

Hinchliffe, G. (2020). What is an Academic Judgement? *Journal of Philosophy of Education*, *54*(5), 1206–1219. https://doi.org/10.1111/1467-9752.12431

Ho, A. D. (2024). Artificial Intelligence and Educational Measurement: Opportunities and Threats. *Journal of Educational and Behavioral Statistics*, *49*(5), Article 5. https://doi.org/10.3102/10769986241248771

IBM. (2024, March 15). *What is Fine-Tuning?* https://www.ibm.com/think/topics/fine-tuning

Ifenthaler, D. (2022). Automated Essay Scoring Systems. In *Handbook of Open, Distance and Digital Education* (pp. 1–15). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-0351-9_59-1

Information Commissioner’s Office. (2023). *Guidance on AI and data protection*. UK Government. https://ico.org.uk/media2/ga4lfb5d/guidance-on-ai-and-data-protection-all-2-0-38.pdf

ITC & ATP. (2025). *Guidelines for Technology-Based Assessment* (1.1). International Test Commission & Association of Test Publishers. https://www.intestcom.org/upload/media-library/tba-guidelines-ver-11-july-2025-1754044782Z3vV9.pdf

Jiang, P., Sonne, C., Li, W., You, F., & You, S. (2024). Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots. *Engineering*, *40*, 202–210. https://doi.org/10.1016/j.eng.2024.04.002

Johnson, M., Liu, X., & McCaffrey, D. F. (2022). Psychometric Methods to Evaluate Measurement and Algorithmic Bias in Automated Scoring. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12335

Johnson, M., & Zhang, M. (2024). Examining the responsible use of zero-shot AI approaches to scoring essays. *Scientific Reports*, *14*(1), 30064. https://doi.org/10.1038/s41598-024-79208-2

Jones, E. (2023). *What is a foundation model?* Ada Lovelace Institute. https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/

Kane, M. (2006). Validity. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Praeger.

Kane, M. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), Article 1. https://doi.org/10.1111/jedm.12000

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept Bottleneck Models. *Proceedings of the 37th International*

Conference on Machine Learning, 5338–5348. https://proceedings.mlr.press/v119/koh20a.html

Kortemeyer, G. (2023). *Performance of the Pre-Trained Large Language Model GPT-4 on Automated Short Answer Grading* (No. arXiv:2309.09338). arXiv. http://arxiv.org/abs/2309.09338

Kortemeyer, G. (2024). Performance of the pre-trained large language model GPT-4 on automated short answer grading. *Discover Artificial Intelligence*, *4*(1), Article 1. https://doi.org/10.1007/s44163-024-00147-y

Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/pdf?id=VD-AYtP0dve

Kumar, V., & Boulanger, D. (2020a). Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value. *Frontiers in Education*, *5*, 572367. https://doi.org/10.3389/feduc.2020.572367

Kumar, V., & Boulanger, D. (2020b). Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value. *Frontiers in Education*, *5*, 572367. https://doi.org/10.3389/feduc.2020.572367

Laux, J., Wachter, S., & Mittelstadt, B. (2024). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, *18*(1), 3–32. https://doi.org/10.1111/rego.12512

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Legg, C. (forthcoming). Peirce and Generative AI. In R. Lane (Ed.), *Pragmatism Revisited*. Cambridge University Press.

Lenat, D., & Marcus, G. (2023). *Getting from Generative AI to Trustworthy AI: What LLMs might learn from Cyc* (No. arXiv:2308.04445). arXiv. http://arxiv.org/abs/2308.04445

Leslie, D. (2020). Explaining Decisions Made with AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4033308

Lewis, M., & Mitchell, M. (2025). Evaluating the Robustness of Analogical Reasoning in GPT Models. *Transactions on Machine Learning Research*. https://openreview.net/pdf?id=t5cy5v9wph

Li, P., Yang, J., Islam, M. A., & Ren, S. (2025). Making AI Less 'Thirsty'. *Communications of the ACM*, *68*(7), 54–61. https://doi.org/10.1145/3724499

Liao, Q. V., & Wortman Vaughan, J. (2024). AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, *Special Issue 5*. https://doi.org/10.1162/99608f92.8036d03b

Lottridge, S., Ormerod, C., & Jafari, A. (2023). Psychometric Considerations When Using Deep Learning for Automated Scoring. In V. Yaneva & M. Von Davier, *Advancing Natural Language Processing in Educational Assessment* (1st ed., pp. 15–30). Routledge. https://doi.org/10.4324/9781003278658-3

Lottridge, S., Ormerod, C., & Patel, M. (2024). Redesigning Automated Scoring Engines to Include Deep Learning Models. In M. D. Shermis & J. Wilson

(Eds.), *The Routledge International Handbook of Automated Essay Evaluation* (1st ed., pp. 487–503). Routledge.

Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., & Gurevych, I. (2024). Are Emergent Abilities in Large Language Models just In-Context Learning? *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5098–5139. https://doi.org/10.18653/v1/2024.acl-long.279

Ludwig, S., Mayer, C., Hansen, C., Eilers, K., & Brandt, S. (2021). Automated Essay Scoring Using Transformer Models. *Psych*, *3*(4), 897–915. https://doi.org/10.3390/psych3040056

Maclure, J. (2021). AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind. *Minds and Machines*, *31*(3), 421–438. https://doi.org/10.1007/s11023-021-09570-x

Mahdavi, H., Hashemi, A., Daliri, M., Mohammadipour, P., Farhadi, A., Malek, S., Yazdanifard, Y., Khasahmadi, A., & Honavar, V. (2025). *Brains vs. Bytes: Evaluating LLM Proficiency in Olympiad Mathematics* (No. arXiv:2504.01995). arXiv. https://doi.org/10.48550/arXiv.2504.01995

Marcinkevičs, R., & Vogt, J. E. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *WIREs Data Mining and Knowledge Discovery*, *13*(3). https://doi.org/10.1002/widm.1493

Massey, A. J., & Raikes, N. (2006). *Item level examiner agreement*. Paper presented at the 2006 Annual Conference of the British Educational Research Association (BERA), University of Warwick.

http://www.cambridgeassessment.org.uk/Images/111065-item-level-examiner-agreement.pdf

McCaffrey, D. F., Casabianca, J. M., Ricker-Pedley, K. L., Lawless, R. R., & Wendler, C. (2022). *Best Practices for Constructed‐Response Scoring* [ETS Research Report Series]. ETS. https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12358

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), Article 6. https://doi.org/10.1145/3457607

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models* (No. arXiv:2410.05229; Issue arXiv:2410.05229). arXiv. https://doi.org/10.48550/arXiv.2410.05229

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A Brief Introduction to Evidence-Centred Design*. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), Article 2. https://doi.org/10.1016/j.rmal.2023.100050

Modhvadia, R., Sippy, T., Field Reed, O., & Margetts, H. (2025). *How Do People Feel About AI?* Ada Lovelace Institute and The Alan Turing Institute. https://attitudestoai.uk/

Molnar, C. (2025). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (3rd ed.). https://christophm.github.io/interpretable-ml-book/

Nam, S. (2025, July). *Domain-Adaptive Automated Essay Scoring with Topic Relevance Learning*. EvalLAC'25: 2nd Workshop on Automatic Evaluation of Learning and Assessment Content, Palermo, Italy. https://ceur-ws.org/Vol-4006/paper2.pdf

Newton, P. (2017). *An approach to understanding validation arguments* (Ofqual/17/6293). Ofqual.

Newton, P. (2023). Certificating Learning on the Basis of Teacher Assessment. In C. Harrison, C. Leung, & D. Pepper (Eds.), *Educational Assessment: The Influence of Paul Black on Research, Pedagogy and Practice* (1st ed., pp. 221–240). Bloomsbury Publishing Plc. https://doi.org/10.5040/9781350288522.ch-14

Nisbet, I., & Shaw, S. D. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, *26*(5), 612–629. https://doi.org/10.1080/0969594X.2019.1586643

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

OECD. (2019). *Recommendation of the Council on Artificial Intelligence* (No. OECD/LEGAL/0449). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OED. (2025). Judgement | judgment. In *Oxford English Dictionary*. Oxford University Press.

Ofqual. (2014). *Review of Quality of Marking in Exams in A Levels, GCSEs and Other Academic Qualifications* (Ofqual/14/5379). Ofqual.

Ofqual. (2017). *Ofqual Handbook: General Conditions of Recognition* (2025 ed.). Ofqual. https://www.gov.uk/guidance/ofqual-handbook

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

PA Consulting, Moores, L., & Handford, J. (2023). *The carbon footprint of a GCSE*. Ofqual. https://www.gov.uk/government/publications/the-carbon-footprint-of-a-gcse/the-carbon-footprint-of-a-gcse

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, *6*, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Pamuk, Z. (2023). Artificial Intelligence and the Problem of Judgment. *Ethics & International Affairs*, *37*(2), 232–243. https://doi.org/10.1017/S089267942300014X

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, *34*(2), 101885. https://doi.org/10.1016/j.jsis.2024.101885

Richardson, M., & Clesham, R. (2021). Rise of the machines? The evolving role of AI technologies in high-stakes assessment. *London Review of Education*, *19*(1). https://doi.org/10.14324/LRE.19.1.09

Rotou, O., & Rupp, A. A. (2020). Evaluations of Automated Scoring Systems in Practice. *ETS Research Report Series*, *2020*(1), 1–18.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. *Harvard Data Science Review*, *1*(2).

Schneider, J., Richner, R., & Riser, M. (2023). Towards Trustworthy AutoGrading of Short, Multi-lingual, Multi-type Answers. *International Journal of Artificial Intelligence in Education*, *33*(1), 88–118. https://doi.org/10.1007/s40593-022-00289-z

Shanahan, M. (2024). Talking about Large Language Models. *Communications of the ACM*, *67*(2), 68–79. https://doi.org/10.1145/3624724

Shaw, S., & Crisp, V. (2015). Reflections on a framework for validation – Five years on. *Research Matters: A Cambridge Assessment Publication*, Issue 19.

Shaw, S., Crisp, V., & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice*, *19*(2), https://doi.org/10.1080/0969594x.2011.563356

Shen, M., Das, S., Greenewald, K., Sattigeri, P., Wornell, G., & Ghosh, S. (2024). Thermometer: Towards Universal Calibration for Large Language Models. *Proceedings of Machine Learning Research*, *235*, 44687–44711. https://proceedings.mlr.press/v235/

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*. Apple. https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf

Suto, W. M. I., & Greatorex, J. (2006). A cognitive psychological exploration of the GCSE marking process. *Research Matters: A Cambridge Assessment Publication*, Issue 2.

Suto, W. M. I., & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, *34*(2), https://doi.org/10.1080/01411920701492050

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, *7*, 100255. https://doi.org/10.1016/j.caeai.2024.100255

Tjong Tjin Tai, T. F. E. (2022). Liability for AI Decision-Making. In L. A. DiMatteo, C. Poncibò, & M. Cannarsa (Eds.), *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics* (pp. 116–131). Cambridge University Press.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. https://arxiv.org/abs/2305.04388

UK Government, D. for S., Innovation &. Technology. (2023). *A pro-innovation approach to AI regulation [White paper]*. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

Vafa, K., Chen, J. Y., Kleinberg, J., Mullainathan, S., & Rambachan, A. (2024). *Evaluating the World Model Implicit in a Generative Model* (No. arXiv:2406.03689). arXiv. http://arxiv.org/abs/2406.03689

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

von Davier, A. A., DiCerbo, K., & Verhagen, J. (2021). Computational Psychometrics: A Framework for Estimating Learners' Knowledge, Skills and Abilities from Learning and Assessments Systems. In A. A. von Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment: With Examples in R and Python* (pp. 25–43). Springer International Publishing. https://doi.org/10.1007/978-3-030-74394-9_3

Vredenburgh, K. (2024). Transparency and Explainability for Public Policy. *LSE Public Policy Review*, *3*(3), 1–11. https://doi.org/10.31389/lseppr.111

Waldo, J., & Boussard, S. (2025). GPTs and Hallucination. *Communications of the ACM*, *68*(1), 40–45. https://doi.org/10.1145/3703757

Wang, C. (2024). *Calibration in Deep Learning: A Survey of the State-of-the-Art* (No. arXiv:2308.01222). arXiv. http://arxiv.org/abs/2308.01222

Watson, D. S. (2022). Conceptual challenges for interpretable machine learning. *Synthese*, *200*(2), 65. https://doi.org/10.1007/s11229-022-03485-5

Webb, M. (2023). *A Generative AI Primer*. National Centre for AI, JISC. https://nationalcentreforai.jiscinvolve.org/wp/2023/10/16/generative-ai-primer/#3

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from Language Models* (No. arXiv:2112.04359). arXiv. http://arxiv.org/abs/2112.04359

Williamson, D. M. (2013). Probable cause: Developing warrants for automated scoring of essays. In *Handbook of automated essay evaluation* (pp. 153–180). Routledge.

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

Wolfram, S. (2024, August 22). What’s Really Going On in Machine Learning? Some Minimal Models. *Stephen Wolfram Writings*. https://writings.stephenwolfram.com/2024/08/whats-really-going-on-in-machine-learning-some-minimal-models/

Wood, S. W. (2020). Public Perception and Communication around Automated Essay Scoring. In *Handbook of Automated Scoring: Theory into Practice* (1st ed., pp. 133–150). Chapman & Hall.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170. https://doi.org/10.1177/0265532209349465

Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., & Chen, G. (2024). Unveiling the Tapestry of Automated Essay Scoring: A Comprehensive Investigation of Accuracy, Fairness, and Generalizability. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(20), 22466–22474. https://doi.org/10.1609/aaai.v38i20.30254

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, *32*, 661–683.

Zesch, T., Horbach, A., & Zehner, F. (2023). To Score or Not to Score: Factors Influencing Performance and Feasibility of Automatic Content Scoring of Text Responses. *Educational Measurement: Issues and Practice*, *42*(1), 44–58. https://doi.org/10.1111/emip.12544

Zhao, C., Tan, Z., Ma, P., Li, D., Jiang, B., Wang, Y., Yang, Y., & Liu, H. (2025). *Is Chain-of-Thought Reasoning of LLMs a Mirage? A Data Distribution Lens* (No. arXiv:2508.01191). arXiv. https://doi.org/10.48550/arXiv.2508.01191

Zhu, C., Xu, B., Wang, Q., Zhang, Y., & Mao, Z. (2023). *On the Calibration of Large Language Models and Alignment* (No. arXiv:2311.13240). arXiv. http://arxiv.org/abs/2311.13240

Appendix

Validity Argument Roadmap for Automated Scoring

The "Validity Argument Roadmap for Automated Scoring" (Dorsey et al., 2024, page 470) specifies 8 stages, depicted in an iterative cycle. They are:

1. Describe assessment constructs, including rationale for using automated scoring
2. Document intended score interpretation and uses (including scoring rubrics, processes, and features)
3. Describe test blueprints and specifications with an eye toward clarifying item types that use automated scoring
4. Describe item/test development procedures, including evidence supporting validity arguments
5. Document approach for developing AI scoring model(s)
6. Describe test administration procedures, including evidence supporting validity arguments
7. Document psychometric analyses supporting validity, reliability, and fairness claims
8. Document reporting procedures, ensuring alignment with score interpretation and uses

ITC and ATP guidelines for automated scoring of constructed-response items

Table 9: Guidelines for automated scoring of constructed-response Items (ITC & ATP, 2025, pages 62 to 64)

4.9	The rationale for using automated scoring should be clearly articulated and appropriate for the program in which it is used. Documentation of the design and use of the automated scoring system should be developed so stakeholders can make reasonable decisions about the scope of its possible application and use.

4.10	The automated scoring system design should appropriately reflect the constructs assessed via the items, rubrics, and other scoring materials. The relationship between the scoring system and the constructs assessed should be documented, including the flow of responses through the 3-stage process (normalization, feature extraction, statistical modeling), detection of unusual responses, and design limitations.
4.11	If an AI scoring system requires training, it should be trained on a representative sample of responses that have been human scored with the highest level of quality the program supports. The rating process should be monitored and aligned with the program's operational practices, with clear measurement and construct validity criteria used to evaluate the quality of the scoring.
4.12	The validity, reliability, and fairness of automated scoring should be evaluated using sound methodological and statistical approaches and clear evaluation criteria. The methods and procedures should be documented and provide recommendations about appropriate use of the automated scoring system.
4.13	The approach for using automated scoring or human scoring methods during test administrations should be based upon the scoring performance of each method and consistent with the goals and stakes of the program. Describe the rationale for the approach and the methods for combining automated scoring or human scoring.
4.14	A well-defined process for reviewing automated scoring performance during and after test administrations should be developed, documented, and implemented. There should be a process in place for handling errors or disruptions.
4.15	For AI-based scoring systems, algorithmic predictions, recommendations, outcomes, and prescriptive actions should be derived via transparent, ethical, and bias-free methods that can be explained and evaluated by internal and external experts or expert systems.

ETS best practice for scoring constructed responses

The ETS guide on best practice for scoring constructed responses offers a framework for establishing validity evidence (McCaffrey et al., 2022, pages 8 to 50). Sections 1, 2 and 3 describe best practices for "planning, designing, executing, and documenting" scoring of constructed responses, with a view to creating a validity argument based on a human marker (section 2) or an automated marking system (section 3). Section 4 then addresses the "principles and decisions" to address when creating reported scores from the marks (in ETS terminology, the ratings), and section 5 addresses evaluation metrics and thresholds for decision-making.

1. Principles/standards and evidence for constructed-response task and test development
    a. Construct relevance of tasks and the constructed-response section of the test

        i. When constructed-response tasks/tests are created, task relevance as well as content coverage should be considered.
    b. Evidence based on relations to external variables
        i. Evaluate the concordance between the task/test scores and related statistics.
2. Principles/standards and evidence for human rating
    a. Construct relevance of human scoring process and materials
        i. Design and implement a human scoring system that is based on the explicit link between the construct to be measured, scoring rubric, and support materials.
    b. Exemplar selection and scoring process
        i. Principled selection of exemplars
            1. Support rater training and scoring by selecting exemplars that represent the variety of possible response types, scores levels, and test-taker characteristics.
        ii. Incorporating Exemplars Representing Atypical Responses
            1. Provide exemplars for raters that demonstrate how to appropriately score responses deemed difficult to score.
        iii. Exemplar consensus scores
            1. Ensure that the consensus scores assigned to exemplars are based on a true consensus of content experts and reflect the gold standard application of the scoring rubric and guidelines.
    c. Development of training and scoring materials for raters
        i. Rater training and operational scoring materials should be clear, precise, and aligned with the tasks to be scored.
    d. Rater selection
        i. Recruit raters who meet the minimum hiring standards for the testing program and who represent a wide range of demographic groups.
    e. Rater training
        i. Raters should be adequately trained before scoring and provided with appropriate support materials to guide their judgments.
    f. Raters’ demonstration of scoring competence
        i. Rater qualifying tests
            1. Prior to operational scoring, use a method to qualify raters that requires them to demonstrate their ability to score reliably at a predetermined level of accuracy.
        ii. Development of Qualifying Tests
            1. Use a method to qualify raters that requires them to demonstrate their ability to score accurately prior to operational scoring.
    g. Rater monitoring
        i. Real-time rater monitoring
            1. Monitor raters in real or near-real time during operational scoring sessions and provide feedback to ensure continued scoring accuracy.
        ii. Ongoing Regular Monitoring of Individual Raters

1. Perform ongoing statistical monitoring of individual raters and use compiled performance data to identify raters for remediation, target real-time monitoring efforts, and provide feedback to raters.
2. Rater Remediation and Rating Invalidation
   a. Identify raters exhibiting less than ideal scoring accuracy or behavior and offer opportunities for additional training that targets their deficiencies.

iii. Ongoing Rater Pool Monitoring and Management
   1. Conduct ongoing monitoring of the rater pool using metrics to provide feedback to raters and determine whether corrective action is needed.
   2. Estimating Inter-Rater Reliability
      a. Implement a system to track inter-rater reliability on a regular basis.
   3. Detecting Rater Drift
      a. Determine if raters change their scoring behaviors over time.

h. Assigning responses to raters
   i. The scoring system should include specifications for the assignment of responses to raters for scoring.

i. Maintenance and Incremental Improvement of Human Scoring System and Materials
   i. Ongoing Maintenance of Test and Scoring Materials
      1. On an ongoing basis, evaluate the test and scoring materials currently in use to ensure that they are still aligned with the construct definition and that the consensus scores are still correct.
   ii. Maintenance of Computer Systems
      1. To ensure accurate data collection and storage of ratings, any computer systems used to implement human scoring should be quality tested on a regular basis and when any changes are made to it.

3. Collecting Validity Evidence for the Use of Automated Scores
   a. Construct relevance of an automated scoring system
      i. Design and implement an automated scoring system based on the link between the construct to be measured and the system components.
   b. Feature and Automated Scoring Engine Development and Maintenance
      i. Feature Development
         1. Develop computational routines to generate features as intended.
         2. Replicability of feature values
            a. Computational routines should consistently yield the same values for the same response.
         3. Human annotations for training components
            a. Human annotations that are used to create and test components should be of high quality.
   c. Automated Scoring Engine Development, Testing, and Maintenance

    i. Base the set of components of an automated scoring engine on empirical research from the appropriate discipline and have as clear a link as possible to key aspects of the scoring rubric.
    ii. Atypical Input Detection
        1. Evaluate the robustness of the automated scoring system to atypical inputs at various design stages that are relevant to the particular use context.
  d. Automated Scoring Model Development (Building and Evaluation)
    i. Characteristics of Samples
        1. Use independent samples that are of sufficient size to yield accurate predictions of human scores and are representative of the testing population and intended score-use contexts to build and train automated scoring models.
    ii. Criterion Variable Quality
        1. Evaluate the statistical and substantive quality of the variables used as prediction criteria prior to using them in model building or evaluation.
    iii. Appropriateness of Modeling Techniques/Algorithms and Selection Criteria
        1. Statistical methods used to develop and select the statistical prediction models should be appropriate for the properties of the data.
    iv. Model Evaluation Analyses
        1. Model evaluation analyses should yield sufficient evidence that predicted scores provide an interchangeable proxy for human scores/judgments that lead to fair conclusions about test takers.
    v. Comparison of Prediction Model Performance by Subgroup
        1. To ensure fairness, evaluate the model performance for all relevant subgroups.
  e. Automated Scoring Model Monitoring and Maintenance
    i. Determine operational quality control procedures to monitor the performance of the automated scoring model both prior to and after the system is implemented.
  f. Implementation and Ongoing Maintenance of the Automated Scoring Software and Architecture
    i. The implementation and maintenance of the automated scoring software should follow up-to-date software industry standards and best practices.
  g. Utilizing Human and Automated Scores in Ongoing Quality Control Measures
    i. Use human and automated scores to perform quality-control checks and detect scoring trends.
  h. Generalizability of Automated Scoring Models
    i. Account for potential generalizations of automated scoring models beyond the evaluation conditions and sample.

4. Evidence for reported scores
  a. Scoring mode and design
    i. Determining the Mode for Rating Constructed Responses

1. The choice of mode for rating constructed responses—human, automated, or both—should be informed by several considerations.

ii. Scoring With Human Ratings Only
   1. Carefully consider how many ratings are needed when scoring with humans only.

iii. Scoring With Both Human Rating and Automated Scores
   1. Determine how human and automated scores will be used in combination.

iv. Scoring With Automated Scores Only
   1. Ensure that the use of scores based solely on automated scoring can be supported by a validity argument.

b. Ancillary Ratings
   - i. Determine the role of ancillary ratings in score calculation.

c. Ratings Resolution and Scoring Rules
   - i. Determine how to use ratings and calculate scores.

d. Score Transparency
   - i. Ensure transparency in the procedures used for producing reported scores.

5. Evaluation metrics and thresholds for decision-making
   - a. Metrics for Evaluation [of constructed response scores]
     - i. Selection of Metrics
       1. Select appropriate metrics for the purpose of the evaluation.
     - ii. Use of Metrics in Human-Rater Monitoring
       1. Use metrics that are appropriate for the type of data as well as the quantity of available cases.
     - iii. Selection and Use of Metrics in Automated Scoring Model Evaluation
       1. Metrics should be chosen in relationship to the validity argument to be developed for automated scores and the assessment of the statistical properties of scores.
   - b. Thresholds for Evaluation Metrics
     - i. Setting Thresholds for Human Scoring
       1. Thresholds for human rater evaluations should be determined under consideration of the metric, historical performance of the threshold for the metric, and empirical evidence.
     - ii. Setting Thresholds for Automated Scoring
       1. Human rater reliability
       2. Concordance With the Human Rater True Score
       3. Subgroup differences
     - iii. Guiding Questions for Setting Thresholds for Concordance Statistics
       1. How important is the concordance with human ratings, as a source of content evidence, such that the automated scores measure the content or construct the item is claimed to measure, according to its specification?

    2. How are automated scores used in the creation of the final reported score and how close of an approximation to human ratings does this require?
    3. How much damage can be caused by construct-irrelevant variance or construct underrepresentation introduced by automated scores?

iv. PRMSE Is Very High (PRMSE > 0.95) [Guidance]

v. PRMSE Is Below 0.70 [Guidance]

vi. Setting Threshold Between the High and Low Extremes [Guidance]

vii. Using Thresholds for Decision-Making

    1. The way in which thresholds for metrics are applied might vary and, thus, the decision-making process should be clear and documented.

Example: validity for feature-based AES

Williamson (2013) presents a detailed account of a validity argument for AES scoring, pre-supposing a feature-based AES scoring engine. This account states that backing for the AES scoring warrant has three parts (2013, page 166):

1. "AES covers the construct appropriately"
2. "Automated scoring models are appropriately derived", and
3. "Scoring is conducted so as to ensure appropriate scores"

To show that "AES covers the construct appropriately" is to show that "AES measures the aspects of writing that, in the aggregate, embody the meaning of writing for the assessment" (2013, page 166). There are potentially a very large number of features in a scoring engine, each of which may be relevant to only part of the overall construct of writing. The features can vary in 2 important ways:

1. They can vary in how accurately they measure the part of the construct they are supposed to measure. Williamson specifically notes that "some features may be very accurate while other features may be relatively inaccurate in distinguishing higher quality aspects of the feature from lesser quality aspects of the feature."
2. They can vary in the extent to which they are actually construct-relevant: in particular, "some being direct measures of an aspect of the construct of interest and others being proxies that may be marginally (or not at all) related to valued aspects of writing". The fact that features may be statistically very useful predictors of human-marker scores, but not construct-relevant, is precisely why the backing needs to be clear about the extent to which features are construct relevant (2013, page 167).

At a secondary level, the backing needs to show that the features in the AES system can represent "the full range of performance" in whatever aspect or dimension of the overall construct the feature is targeting (page 167). Finally, the backing needs to establish that the features used by the AES system represent the full construct, when aggregated, since "Omission, or incomplete coverage, of aspects of the writing construct threatens the construct representation of AES scores." (page 168) A key point of evaluation is whether the features are "designed and programmed in such a way that they represent a truly valued part of the construct" or whether they are instead "statistical proxies of what is valued" (Williamson, 2013, page 170).

Calibration of the scoring model is typically specific to the population of interest. The backing also requires:

- an "appropriate" method of aggregating features

- establishing that “the statistical influence of features on the resultant score is consistent with the definition and quality of the feature” (Williamson, 2013, page 168)
- calibration methods that are “appropriate”
- a sample large enough to produce reliable results

Williamson particularly highlights how empirical calibration of an AES scoring model can raise tensions between trying to meaningfully represent the writing construct at the same time as achieving accurate prediction of the scores from human markers:

> A feature of AES might be conceptually very important to the construct definition, but be poorly measured by AES. Similarly, another feature may be less important to the construct but measured with great accuracy and have notable variance in the population. (Williamson, 2013, pages 168 to 169)

Because of this discrepancy, an empirical machine learning calibration process may well assign higher weighting to the less-important but accurately measured features, and lower weighting to those features that are important to the assessment construct but poorly measured. As a consequence of these weightings, the “resultant scoring model could have conceptually important features with little influence on the essay score and conceptually minor features with great influence on the score” (Williamson, 2013, page 169). Williamson suggests that such a model may present an “interesting dilemma” to those evaluating its merits (2013, page 169).

The account presented by Williamson argues that choosing between human marking and AES marking means weighing up the random error of human markers against the risk of consistent marking by AES that systematically under-represents (perhaps only slightly) the assessment construct[10] (Williamson, 2013, page 174). This ‘trade-off’ is interesting to reflect upon in today’s context, since it is no longer true of some AI scoring approaches. In particular, some AI systems for generating scores do not provide complete consistency (they provide variable outputs to the same inputs) (Pack et al., 2024).

[10] “In summary, the most notable distinction between human and AES scoring is that with human scoring we have greater confidence in the *potential* for the construct to be well-represented but less confidence in the conscientiousness and consistency of scoring, while AES scoring may entail some inadequacies in construct representation but provide highly conscientious measures and complete consistency in the use of features to determine summary scores. In this way, a transition from human scoring to AES is an exchange of the random error of human scoring for systematic error of AES. Such an exchange may be advantageous if it can become a mechanism for better identification of sources of error in scoring and successful efforts to control those sources of error.” (Williamson, 2013, page 174)

Published by:

ofqual

2nd floor, 1 Friargate
Station Square
Coventry
CV1 2GN

0300 303 3344

www.gov.uk/ofqual

January 2026

Ofqual/26/7292