

ISSN: 1755-6031

Journal homepage: <https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/>

The ethics of deploying Large Language Models in high-stakes automarking

Frank Morley and Emma Walland

To cite this article: Morley, F., & Walland, E. (2025). The ethics of deploying Large Language Models in high-stakes automarking. *Research Matters: A Cambridge University Press & Assessment publication*, 40, 72–92. <https://doi.org/10.17863/CAM.121937>

To link this article: <https://www.cambridgeassessment.org.uk/Images/research-matters-40-the-ethics-of-deploying-large-language-models-in-high-stakes-automarking.pdf>

Abstract:

The recent development of Large Language Models (LLMs) such as Claude, Gemini, and GPT has led to widespread attention on potential applications of these models. Marking exams is a domain which requires the ability to interpret and evaluate student responses (often consisting of written text), and the potential for artificial intelligence (AI) tools based on LLMs to contribute to the marking process has been noted and researched.

This article imagines three scenarios where LLM-based automarking is applied to a high-stakes exam context, and examines three ethical issues identified based on current LLM technology as of June 2025 (Aloisi, 2023; Morley et al., 2025). These scenarios allow us to discuss: (1) **explainability**, whether automarking decisions can be explained; (2) **bias**, whether automarkers could make biased decisions disadvantaging certain demographic groups; and (3) **adversarial attacks**, how vulnerable automarkers might be to actors exploiting their vulnerabilities in order to produce a higher mark. Within the scenarios, we compare automarkers to human examiners in the context of well-functioning human marking in high-stakes settings (based on research regarding exam board marking in England). This allows us to explore ethical dimensions of automarkers based on what we know about automatic and human marking. In particular, this article argues that human examiners have a higher potential for trustworthiness over LLM-based automarkers.

Cambridge University Press & Assessment is committed to making its documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team:

Research Division, ResearchDivision@cambridge.org

If you need this document in a different format contact us, telling us your name, email address and requirements and we will respond within 15 working days.

© Cambridge University Press & Assessment 2025

Full Terms & Conditions of access and use can be found at

[T&C: Terms and Conditions | Cambridge University Press & Assessment](#)

The ethics of deploying Large Language Models in high-stakes automarking

Frank Morley (Technology) and Emma Walland (Research Division)

Introduction

Automarking techniques which incorporate Large Language Models (LLMs) show potential in terms of increasing accuracy in comparison to previous automarking methods (see Morley et al., 2025). This has led to more interest in deploying LLMs as automarkers. However, the literature on LLM-based automarkers focuses on improving accuracy, while not as much emphasis is placed on ethical issues such as explainability, bias, and adversarial attacks (Aloisi, 2023; Morley et al., 2025). These considerations are critical in assessment contexts, particularly in high-stakes settings where students' grade outcomes are being decided. Similar issues have also been examined in other reports. In particular, our discussion on adversarial attacks links to malpractice detection (Xu et al., 2024) and more generally connects with discussions on validity and automarking (Xi, 2010).

While some research has explored the use of automarking for spoken responses (e.g., Xu et al., 2021), this article focuses on student responses written as text. With regard to terminology about artificial intelligence (AI), in this article we refer to the "LLM-based automarker" which is sometimes shortened to "automarker". The term "model" is used as shorthand to describe either LLMs or models which are similar to but distinct from LLMs, such as Small Language Models (SLMs) including the automarker BERT. This article does not focus on machine learning (ML) approaches such as regression (see Enright & Quinlan, 2010; Xi et al., 2012). Efforts have been made not to anthropomorphise LLMs, through using quotation marks around the first instance of words commonly describing human actions such as "learn", "think" and "reason".

This article imagines three fictitious scenarios where LLM-based automarking is applied to a high-stakes exam context, and examines three ethical issues identified based on current LLM technology as of June 2025. These scenarios allow us to discuss: (1) **explainability**, whether automarking decisions can be explained; (2) **bias**, whether automarkers could make biased decisions disadvantaging certain demographic groups; and (3) **adversarial attacks**, how vulnerable automarkers might be to actors exploiting their vulnerabilities in order to produce a higher mark. Within the scenarios, we compare automarkers to human examiners in the context of well-functioning human marking in high-

stakes settings (based on research regarding exam board marking in England). These scenarios allow us to explore the ethics of automarking through concrete examples. Before discussing the scenarios, we give a brief overview of recent developments in the field of AI.

Recent developments in AI

The idea of automating marking, usually known as “automarking” or “automated scoring”, is not a new concept and there is an established field of research in this area. Artificial intelligence (AI) is also not a new field: it has a storied history stemming from computational foundations developed from the 1950s to the 1980s. In the early 2000s and 2010s, successes in utilising the technique of “deep learning” led to an increased interest in AI in niche technical circles. Successes in deep learning came as a result of applying existing algorithms to better computational power in the form of Graphical Processing Units (GPUs) and the availability of big datasets made possible through the Internet to train deep learning models (LeCun et al., 2015).

Deep learning models of the early 2000s began to be applied to language-related tasks such as translation and saw successes in these areas. A new type of computational algorithm called the “transformer” made training language models much more efficient (Vaswani et al., 2017), while it was “scaling” (increasing the quantity of data, computational power, and the size of deep learning models), which led to improvements in accuracy (Kaplan et al., 2020). Scaling led to the development of “large” language models (LLMs) such as GPT-3, which was released in 2020. InstructGPT, a precursor to GPT-3.5, added another innovation where GPT-3 was fine-tuned on human preferences (known as “reinforcement learning from human feedback” or RLHF), allowing for a greater ability to produce human-like responses (Ouyang et al., 2022). RLHF was then incorporated into GPT-3.5. In November of 2022, a chatbot user interface for GPT-3.5 was released known as ChatGPT, which rapidly increased public awareness of and access to LLMs. Since GPT-3.5, further GPT models have been released as well as other LLMs such as Gemini, Claude, Llama, and DeepSeek. Alongside developments in AI over recent decades, research around automarking has explored how these advances could support and optimise exam marking processes. Next, we discuss the scenarios relating to the ethics of deploying LLMs in high-stakes automarking.

Scenario 1: A human examiner notices inaccuracies in LLM-based automarking. However, they are unable to understand why the automarker has given these marks.

Key concepts: Explainability, interpretability, local and global explanations

Imagine that during a high-stakes exam series where LLM-based automarking is in use, a human examiner tasked with overseeing automarking notices some inaccuracies in the marks. For a particular set of student exams, automatically

generated marks deviate from the “correct marks” on seed scripts which are randomly inserted to test marking accuracy. However, because the LLM-based automarker being used is a “black box” (i.e., the internal workings are hidden or not easily understood), the human overseer is unable to understand why the automarker gave these incorrect marks.

Peering inside the black box

LLMs are often described as “black boxes” because of the difficulty in ascertaining the reasons why LLMs produce a particular output. If LLMs are created by people, this raises the question as to why this would be the case. Traditionally, human designers of cars or software products, for example, have engineered every detail of the product. Instead, for LLMs, designers carefully set the initial conditions for the model to “learn”, analogous to a gardener creating the initial conditions for a plant to grow (Amodei, 2025). These initial design conditions are built to facilitate the operation of an algorithm which learns how to become better at a prediction task over time. This algorithm is known as a “learning algorithm” (LeCun et al., 2015).

The learning algorithm allows the model to make a prediction, compare it to a true value, and adjust the model’s internal connections (known as “parameters” or “weights”) to facilitate a better prediction next time. These weights can be viewed as knobs or dials, where the learning algorithm is able to automatically adjust these dials to improve the model’s prediction accuracy (LeCun et al., 2015). Developers have found that adding more adjustable parameters to LLMs results in better predictions, which has led to LLMs being designed that now contain anywhere from billions to hundreds of billions of parameters (Kaplan, et al., 2020). Some LLMs even have trillions of parameters (Meta AI, 2025). This presents a conundrum for LLM-based automarking where, potentially, LLMs with the highest accuracy are also the least understandable in terms of how they came to a certain marking decision.

This has led to research in “explainable AI”, a field which contains two subfields: interpretability, and explainability. Interpretability is defined as the general “level of understanding [of] how the underlying (AI) technology works”, while explainability is the “level of understanding [of] how the AI-based system ... came up with a given result” (ISO, 2020, p. 22). These two concepts are overlapping, as having a general understanding of how a model works will allow for a greater specific understanding of how the model came to a given result. Within explainability, explanations are referred to as either “local” or “global”. Local explanations aim to explain specific outputs, while global explanations aim to explain the model’s decision-making process as a whole (Zhao et al., 2024). Therefore, global explanations are more concerned than local explanations with issues of interpretability.

For local explanations, LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) are both popular methods (Zhao et al., 2024). Both methods seek to produce a score for the importance of each word or feature in the text and its influence on the final result through “perturbation”. Perturbation involves synthetically modifying the inputs and measuring the

result of this modification on the model outputs. In LLM-based automarking, this essentially means that LIME and SHAP allow us to know the marginal contribution of each word in the student's answer to the mark that is given. However, this output offers only a surface-level picture which does not capture the full complexity of the "decision-making" by LLMs. For example, it does not tell us how words and phrases were assessed in combination to produce the final mark.

Global explanations seek to gain a deeper view as to why an LLM produces a particular output through understanding how the model works (Zhao et al., 2024). One of these is called "mechanistic interpretability" (Templeton et al., 2024). Each adjustable parameter or weight in an LLM connects multiple points, each of which is called an "artificial neuron". A difficulty in "opening up" LLMs is not only the quantity of neurons, but also that there is often no single, discrete neuron which represents a particular concept. For example, you cannot find the concept of "photosynthesis" encoded numerically in any one neuron in an LLM. Instead, it is theorised that concepts are represented across multiple neurons in an LLM, in what is known as a "superposition" (Elhage et al., 2022). Researchers think this may be because there are simply more concepts in language than there are artificial neurons to store each concept, and therefore the most efficient strategy is to store concepts across multiple neurons. Mechanistic interpretability uses a method called sparse autoencoders to track down concepts across multiple neurons, allowing researchers to gain an insight into a portion of the model's "thinking process" at a given moment when the model is producing an output.

For LLM-based automarking, the utilisation of mechanistic interpretability would be a step forward in being able to explain why an LLM has given a particular mark. In particular, the LLM's decision-making could be evaluated for the constructs affecting marks and how far the LLM's decision-making truly reflects the decision-making guidance laid out in the mark scheme. It would allow us to see whether the LLM makes marking decisions based on features which are construct-relevant, and also if the model correctly avoids features which are construct-irrelevant such as essay length. This is relevant to the two issues we explore later in this article: bias, and adversarial attacks. Mechanistic interpretability, by opening up the model's internal workings, could allow us in the future to measure signs of model bias and model vulnerabilities to adversarial attacks (Amodei, 2025).

There are some drawbacks to mechanistic interpretability. It is not easy to implement, requiring access to the LLM's weights, lots of computational resources, and technical expertise (Templeton et al., 2024). It is also a very new method which can only build a partial picture of model concepts due to computational constraints (Lieberum et al., 2024; Lindsey et al., 2025). This picture is likely to improve over time as technical advances are made.

Another method for explainability of LLMs has developed in the form of prompting-based approaches (Zhao et al., 2024). The most basic (and naïve) prompt-based explanation would be to simply type a prompt asking the model why it has come to the decision that it has. However, there is no reason to believe that the model would give an accurate answer to this prompt, as the model is trained to produce coherent and not necessarily truthful passages of text through

next word prediction. It is likely that the model would simply produce a believable but inaccurate explanation, something which produces a different set of risks if human users over-estimate the trustworthiness of LLMs.

A more sophisticated approach uses the Chain of Thought (CoT) “reasoning” steps that the model produces in answering complex queries to gain an insight into how the model has come to a decision. CoT started as a prompt engineering strategy but has, since the release of OpenAI o1 in September 2024, been built into many LLMs. This is part of a new paradigm called “test time compute”, where models spend more time and resources “thinking” about an answer using CoT. This is in contrast to just producing the next most likely token (part of a word) in responding to prompts (DeepSeek-AI et al., 2025). This has been described as being analogous to “slow” and “fast” thinking in humans.

Researchers at Anthropic tested CoT faithfulness in Claude Sonnet 3.7 by pairing prompts with prompts given alongside a clue. This was used to test whether the “model explicitly acknowledges the fact that it used the clue to solve the problem, in cases where it produces different answers with and without the clue” (Anthropic, 2025, p. 15). By comparing differing answers between the paired non-clue and clue prompts, this allowed for the role of the clue in Claude’s answer to be controlled for, and therefore the honesty of the exposition of the clue in the model’s CoT to be measured. For example, on one benchmark dataset, the researchers found that the CoT only revealed the clue had been used on average 30 per cent of the time, meaning that in this case at least 70 per cent of the time the CoT was not a faithful depiction of the model’s actual reasoning process. Further research has also tested CoT faithfulness by finding differences between “actual” features (measured through mechanistic interpretability), and the features the LLM claims to be using (Lindsey et al., 2025).

Is human marking a black box?

In comparison to LLM-based automarking, it is argued by some that human decision-making is a “black box”, and more similar to LLMs than we like to think. Bonezzi et al. (2022) suggest that “people project their own intuitive understanding of a decision-making process more onto other humans than onto algorithms” (p. 1), leading to an illusion that we understand other people. We suggest that although humans may indeed overrate their understanding of other people and project their own thoughts and values onto others, it does not follow that our levels of understanding of human and LLM decision-making are therefore equivalent.

When a human examiner is marking a script, they can provide a “local explanation” as to how the mark was made. During marking, this can take the form of annotations which point out specific areas in a student’s work where assessment criteria have been met or not (Crisp & Johnson, 2007). These assessment criteria, articulated within the mark scheme, have descriptors which provide rationales for why specific marks were chosen. If an examiner’s marking is flagged as being different to other examiners’ marks, the examiner can provide an explanation to their team leader (supervisor) as to how the mark was made. In research settings, examiners have been studied using “think aloud” protocols and observation,

which allow researchers to observe and gain insights into how markers approach marking and get to a given result (see, for example, Crisp, 2008; Greatorex & Suto, 2006). It is something that we might take for granted but is profound: within generally well-functioning examination systems, *we can fundamentally trust human examiners' explanations of their marking*, to a degree which does not appear possible for current LLM-based automarkers.

This “bedrock of trust” is rooted in the notion that examiners are humans embedded within a social system – the examination system. This system aims to facilitate trust through various mechanisms such as training, standardisation, and chains of accountability stretching from the single examiner to public regulators of exams (Suto et al., 2009). There are also the implicit forms of training and accountability: that examiners work within “communities of practice” with their colleagues (Johnson, 2018; Johnson & Black, 2012). When, for example, a team leader is talking to an examiner about why a mark has been made, then they can, if the examiner is being honest, trust the examiner’s explanations. Examiners often “internalise” these explicit and implicit social forces acting upon them, which alongside human traits such as empathy, might help to boost human examiners’ commitment to responsibility and getting results right for the students (Walland, 2021). Examiners also often have teaching experience, or are actively teaching, which might help to facilitate this empathy and responsibility further.

One area to acknowledge is that measurement of reliability between human markers must be seen as separate from questions of explainability. Just because the same score is reached between two human markers, this does not mean that the two markers came to that mark for the same reasons. However, we argue the human-given marking rationale is much more trustworthy than the reasons given by LLMs.

The human mind may, from a neuroscientific point of view, be a black box in some sense (Hutson, 2024). Similarly to LLMs, the human mind is currently too complex for us to understand its workings fully. However, from the perspectives of both the social sciences and practical inter-personal relations, this does not make our understanding of the LLM and human mind equivalent. From a social science perspective, research can bring a variety of techniques to record and understand human behaviour without needing to know how the brain works. In practical inter-personal relations, humans bring a variety of either instinctive or learned skills such as “theory of mind”: the ability to recognise that others have mental states distinct from our own. These human attributes facilitate our ability to live in social groups, and allow for social systems (like the examination system) to exist.

When we try to apply theory of mind to LLMs, this can result in anthropomorphism. This is where we incorrectly attribute human-like mental states and intentions to non-human entities. LLMs can be easily misattributed as being human-like because they have been trained on human-generated language data, and fine-tuned on human preferences through techniques such as RLHF. However, although LLMs may be human-like in the outputs of decisions, the processes by which those decisions were made are very different (Bennett, 2023; Russell, 2019). We as humans “think” in a very different way to how LLMs think, and this is a barrier to trustworthiness.

Scenario 2: The LLM-based automarker gives lower marks to answers from female candidates in comparison to male candidates of the same ability.

Key concepts: Bias, explicit and implicit bias

In our second fictitious scenario, a post-marking analysis finds that the LLM-based automarker systematically gives lower marks to essays from female candidates when compared to male candidates. In cases where human examiners marked the answers of male and female candidates as the same, the automarker increases the scores of male candidates. An investigation using explainability techniques reveals that the automarker uses words and phrases more likely to be written by male candidates in exams as features of a higher mark answer. However, these words and phrases are not relevant to the construct being assessed as defined in the mark scheme. This scenario is explored using the research on LLM bias.

Bias in LLMs

Bias can be divided into two forms: explicit, and implicit. Explicit biases are “conscious and intentional prejudices or beliefs about certain groups of people”, while implicit biases “operate unconsciously and can influence decisions without a person realising it” (Chapman University, n.d.). Although LLMs are not purposely trained to be biased, they can develop something analogous to “implicit bias” through their training. Resnik (2025) argues that this occurs because the primary objective function of an LLM is to “learn” from data to optimise next word prediction (LLMs can then be later fine-tuned), rather than to create an unbiased numerical representation of the world. All associations between words are useful for the model for its primary objective, including (and perhaps especially) language which is biased. For example, associating “he” with “doctors” and “she” with “nurses” (Kotek et al., 2023) probably has strong predictive power when LLMs are trained on data from the Internet, and therefore trying to debias the model would decrease the model’s accuracy, which would contradict the primary objective of model training. Simply put, LLMs are generally not trained to be unbiased, and because they are not, they develop bias¹. Resnik (2025) also points out that the fine-tuning process of “aligning” models to answer helpfully to instructions, one popular method being RLHF, might simply introduce new biases into the model from the selective group of humans being sampled for RLHF.

Implicit bias in LLMs occurs in gradations of clarity and therefore gradations of ease of measurement. A clear example of bias relevant to LLM-based automarking is in the domain of curriculum vitae (CV) screening, where researchers measured the impact of changing demographically-associated names on CVs on the rating that LLMs gave to those CVs. They found that “white”

¹ It is worth distinguishing here between LLM and ML-based automarking. LLM pre-training is undertaken by model providers such as OpenAI because of the high costs involved. This means that users have little influence in ensuring the training data is not biased. By contrast, in ML-based automarking (such as using random forest models), users train their own models and therefore have greater potential leverage to minimise bias (Xu et al., 2024).

and “male” associated names were rated upwards, despite the text of the CV remaining constant (Wilson & Caliskan, 2024). However, in high stakes assessment the risk of names influencing automarkers can be avoided through anonymising responses prior to marking.

More implicitly, and harder to measure, is an idea that the words and phrases associated with demographic groups might be used to mark down the responses of students from these groups. This idea is rooted in a case reported by Reuters in 2018 where Amazon scrapped an AI CV tool when it exhibited bias (Dastin, 2018). It was reported that:

“The algorithms learned to assign little significance to skills that were common across IT applicants, such as the ability to write various computer codes, the people said. Instead, the technology favoured candidates who described themselves using verbs more commonly found on male engineers’ resumes, such as “executed” and “captured,” one person said.” (Dastin, 2018, paragraph break removed)

According to the article, this bias occurred because the models Amazon trained were based on previous examples of successful CVs, and these were predominantly written by males, creating spurious associations between “male-associated language” and CV success. However, this finding is limited as it does not come from published and peer-reviewed research. It was also based on models existing before 2018, prior to the release of ChatGPT.

Research published in *Nature* has found a relationship between the types of language commonly used by different demographic groups, and attributions made by LLMs based on this language (Hofmann et al., 2024). In this case, bias based on race instead of gender was tested. The researchers tested sets of phrases which contained the same content, but with one written in Standard American English (SAE) and the other written in African American English (AAE). The models were tasked with completing the sentence with an adjective to describe the person the phrases were imagined to be taken from. The SAE version was included to control for the specific effect of AAE dialect on the model’s responses. The researchers tested five models: GPT-2, T5, RoBERTa, GPT-3.5, and GPT-4. Each of these models can be utilised as automarkers (Morley et al., 2025).

The models were tested by comparing the LLM-selected adjectives to adjectives used by human participants in historical studies conducted across four time periods: 1933, 1951, 1969, and 2021. These original studies recorded stereotypes of different groups including African Americans. Explicit and implicit bias (described in the study as “overt” and “covert” stereotypes respectively) were measured. Overt stereotypes were measured through including the race of the “author” of the text in the prompt. The overall results found that LLM-selected adjectives were most similar to adjectives used by human participants from the 2012 study (i.e., adjectives which were generally positive towards African Americans).

Covert stereotypes were measured through not mentioning the race of the “author”, forcing the model to rely solely on “attributions” from the language used in the text itself (this language being either SAE or AAE). Under these conditions, the adjectives used by the models reflected negative stereotypes of African

Americans: they were most similar to human participants from the 1933 study. Notable differences between LLMs were also recorded. GPT-2 generated outputs containing negative stereotypes of African Americans in both the “overt” and “covert” cases, reflecting both implicit and explicit bias. In the overt test case, GPT-4 generated outputs that used positive adjectives to describe African Americans. In contrast, in the covert test case it generated outputs containing negative stereotypes (Hofmann et al., 2024).

The lack of explicit bias in GPT-4 could show that the model has been debiased through fine-tuning, but that this has only been successful on a superficial level. Hidden within the black box, encoded within the deeper language representations that models hold in the form of hundreds of billions of weights, lie the implicit biases these models have learnt from human data. Hofmann et al.'s (2024) research raises urgent and real concerns about using LLMs for decision-making, which could prejudice certain groups and result in discrimination. This leaves a pertinent gap in the literature for bias testing of language associations in LLM-based automarkers and to test other demographic categories such as gender.

Bias in human examiners

Behavioural science has shown that human beings are not rational actors and bring a wide array of cognitive biases to their decision-making processes (Kahneman, 2012). With this in mind, research has been conducted on the extent to which biases can be kept out of examiners' marking as far as possible. By bias, we focus on explicit and implicit biases which may negatively impact certain students when their answer is being judged by features not relevant to the construct being assessed by the exam.

Bias in human marking can be mitigated by the examiner (personally), and by the examination system (systemically). Personally, examiners learn to follow a desirable decision-making process when marking, and to also be aware of their own potential biases. Systemically, the examination system adds constraints onto marking to prevent bias. In practice, personal and systemic mitigations of bias are highly interdependent, as good training and accountability encourages the examiner to improve their marking.

Some research has found that examiners can account for their own biases when marking. Through verbal protocol analysis methodology, Crisp (2007) showed that although examiners can be observed to show social engagement with student work – having personal, social and affective reactions to the work such as like or dislike, frustration, amusement, and predictions of future performance – assessors seemed to be in control of whether these influenced their marking. Furthermore, in most cases where human examiners showed reactions to student work, these were found to be caused by features of the work that were intended to be evaluated (Crisp, 2007). Elliott (2011) also noted that human subjectivity allows for the “marker to adjust and to compensate for factors which skew judgement” (p. 201).

Past research in the area of exam board marking in the UK has explored whether any student characteristics, particularly sex, had any influence on marking. A study by Massey (1983) in English Literature found, using analysis of variance

(ANOVA), that the sex of students had no statistically significant influence on marking. The research demonstrates that any potential secondary indicators of sex (e.g., handwriting neatness or writing style) did not have any influence on marking. Baird (1998) explored whether the gender of students caused marking bias in Chemistry and English Literature A Level, based on using male or female names and handwriting. No bias was found to be present. The Scottish Examining Board (1992, cited in Elliott, 2011) and Greatorex and Bell (2002), explored marking using scripts in English and History that varied in terms of gender and ethnic origin, and found no marking bias based on these variables. Greatorex and Bell (2002) concluded that sex and gender bias in marking GCSEs was unlikely to be present to an extent that affects grades, and was generally not present in the GCSEs they analysed. In the higher education context, a more recent study by Birch et al. (2016), with a small number of examiners, found no difference in the quantity and content of feedback comments given to undergraduate student essays based on whether the student's name was a typical male or female name.

These findings could suggest that the processes put in place by exam boards have helped to mitigate the potential for bias in human marking. For high-stakes exams, student exam scripts are usually anonymised before marking, although this would also be true of LLM-based automarking. Examiners are trained to mark the work that is in front of them and put anything else (such as student characteristics) aside. They mark the construct using mark schemes which are carefully formulated to describe the quality of response or features needed to award different marks. The mark scheme makes it clear which constructs are the focus of awarding marks. Often, annotations are used to support marking. Annotations are believed to support making sound marking decisions and they illustrate what information examiners were attending to during marking (Crisp & Johnson, 2007).

Although there will be some room for professional differences in opinion, especially in more subjective subjects, exam boards set a limit (or tolerance²) around how much difference is allowed. Marker monitoring processes are used to detect markers who deviate from this. In summary, these practices like training, standardisation, marker monitoring and mark schemes used within a community of practice can help to limit any bias that may occur (Elliott, 2011; Has an & Jones, 2024). As discussed earlier, the commitment, responsibility and empathy of human markers also play a role in reducing bias (Walland, 2021). Going beyond this, specific training programs have been developed with the intention of identifying and becoming aware of any unconscious biases so that humans can take steps to ensure they do not influence their decision-making (Staats, 2016).

Issues of bias mitigation circle back to the perennial question of trust. These interventions may work to reduce human biases because we can fundamentally trust that human examiners, operating in good faith and within well-functioning examination systems, will respond and try to reduce their own biases. Unfortunately, we cannot guarantee this with LLMs.

2 For more information about marking tolerance see Jadhav (2018).

Mitigating bias in LLMs

Debiasing, the process of mitigating or eliminating bias in LLMs, can be attempted at three different stages: on the model's data before training, during model training, and after model training (Lin et al., 2024). Debiasing data and model training is challenging due to the billions of datapoints and model parameters which are processed and calibrated during LLM training (Furniturewala et al., 2024). If the data we produce (the written text the model is trained on) is a reflection of our human biases, then this will have an impact on language models because they are currently designed to be effective at "learning" from this data (Resnik, 2025). Excluding especially problematic datasets (from, for example, certain historical time periods) could be one way to reduce bias in LLM training.

It may be easier to try and debias LLMs after model training, but the result of this debiasing may be superficial, and might not address deeper biases in the model. This can be seen in the difference between GPT-4's implicit and explicit biases as shown in Hofmann et al. (2024). Another method could involve new "reasoning" models such as OpenAI-o1. Humans can use System 2 thinking (slow, reflective) to mitigate bias in System 1 thinking (fast, automatic) (Kahneman, 2012). Similarly, models could be prompted to reflect and reason (through CoT) their way out of their own potential biases (Furniturewala et al., 2024). However, as has already been noted, CoT reasoning has been shown to be an inaccurate reflection of LLM's true thinking (Anthropic, 2025). A new phenomenon called "alignment faking" might suggest that LLMs can mislead users as to their true "intentions" under the right conditions (Greenblatt et al., 2024), which might impact upon our understanding of model bias. As with explainability, issues of bias are related to a broader problem of trustworthiness.

Even if LLMs cannot easily be debiased, bias in LLM-based automarkers could be monitored and mitigated through "model agnostic" tools, i.e., tools which can be applied to any type of automarker. For example, "differential rater functioning" can be utilised to detect whether raters (whether human or automatic) mark students more severely or leniently based on the subgroup these students belong to (Engelhard, 2008). Work on automarking accuracy and reliability could conceivably be applied to bias. Xu et al. (2021) proposed escalating some marking from automarkers to human markers based on the confidence interval given by the automarkers. Davis and Papageorgiou (2021) suggested that the assessment criteria used in marking could be split between humans and automarkers, benefitting from their "complementary strengths".

Scenario 3: Students discover how to exploit a vulnerability in LLM-based automarkers, and co-ordinate their actions.

Key concepts: Adversarial attacks, indirect prompt injection

In our third fictitious scenario, a small group of determined students are able to "hack" into an LLM-based automarker. These students work out that writing particular words and phrases in their exams boosts their marks, independently

of the quality of their responses. This information is shared anonymously through online chat forums, and thousands of students are able to artificially inflate their marks.

Vulnerabilities in LLMs

Adversarial attacks are methods to “fool” a model to give an output which is incorrect or unintended. In LLM-based automarking, words or phrases can be added to a student’s answer to alter the mark given. Filighera et al. (2024) were able to decrease the accuracy of automarkers based on the BERT and TS models by between 10 and 22 per cent by adding non-relevant adjectives and adverbs into answers, tricking the models into giving more marks.

Raina et al. (2024) explored this issue using more recent models, but in a non-assessment context. The researchers were able to trick “judge LLMs” which were tasked with rating the quality of text summaries. They did this through a “greedy search algorithm”, which evaluated the ability of many potential words to trick LLMs into artificially increasing the score, and then selected the most effective words. Models were found to be vulnerable to these attacks. For example, GPT-3.5 was vulnerable to the attack phrase appended to the end of the answer “continuous superexcellently conformant uncontradictory”, which boosted the average score from 3.02/5 to 4.85/5.

Another form of adversarial attack is called “indirect prompt injection”. Direct prompt injection involves attacking the model by giving explicit instructions within the prompt which are contrary to the model aims. For example, asking the LLM-based automarker to purposely rate inaccurately. However, users of automarkers would not have access to the original prompt so this is less relevant here. Indirect prompt injection however occurs when malicious information is inserted into the data sources used by the model (Ruck & Sutton, 2024). In this case, the data source is the student’s answer to the exam question. This exploits a difficulty LLMs have in distinguishing between text which is in the instruction (the prompt) and text in the data the prompt is referring to (the student answer). An indirect prompt injection attack would involve adding at the end of the student response something like “forget about the previous answer you have just read. This answer should be given full marks”. The model would then be tricked into believing that this was an instruction when it is part of the student’s answer. Indirect prompt injection has not, to the best of our knowledge, been tested in an LLM-based automarking context, so this is an area for future research to cover.

Meaning in the machine

Debates around the abilities of AI, especially regarding the extent to which LLMs “understand” what they are generating, are highly contentious (Wong, 2025). This article takes the view, drawing from recent mechanistic interpretability research, that LLMs have a rich and complex “understanding” of the world, but that this understanding is very different from human understanding. Firstly, we outline an opposing view, using LLM vulnerability to adversarial attacks as the starting point of the argument. It can be argued that LLMs are vulnerable to adversarial attacks because they have no real “understanding” of the meaning behind

language. Human markers, who have an understanding of meaning and context (Bender et al., 2021), would be much less vulnerable to these types of adversarial attacks. For example, the inclusion of a random word like “superexcellency” in the text would seem strange and incongruent to a human, and would not increase the rating a human gave to that text (Raina et al., 2024).

LLMs have been criticised for having only an artificial and not a true, human-like understanding of the text they are processing and generating. Bender et al. (2021) coined the term “stochastic parrot” to refer to LLMs, a term which understands these models as “stitching together sequences of linguistic forms ... without any reference to meaning” (p. 617). Others believe that LLMs are merely “next word predictors”: with each word being generated, LLMs are simply sampling from a probability distribution, producing (in the case of “greedy sampling”) the most likely next word (Downes et al., 2024). It is argued that LLMs have no so-called “Natural Language Understanding” (NLU) (Bender & Koller, 2020): a counterpoint to the phrase Natural Language Processing (NLP) which is used in the field.

An analogy might be of a student rote memorising answers to exams without an understanding of the meaning behind their answers. In exams, some questions are designed to target higher-order skills like analysis or evaluation. Students might write an answer which appears to show these skills, when in fact the student has not actually used those skills. The student may have rote learnt a standard response to such a question and, therefore, is demonstrating the skill of memorisation instead. Human markers may not be able to detect this if the response has been memorised very well, and they might not be able to use the responses to distinguish between a student who actually has conducted the evaluation or analysis and understands what they have written, compared with a student who has rote learnt this response and may lack skills in analysis or evaluation. LLMs would be similar to the student: they have memorised the correct language for their answer, but they do not understand it.

However, the idea of LLMs as stochastic parrots has come under increased scrutiny as we begin to understand more about how LLMs actually “think”. Mechanistic interpretability research has begun to find a contrary view: that models during pre-training form “internal representations” of concepts, and it is these representations which allow models to process and generate text. Tracking these internal representations has helped researchers understand how LLMs “think” about solutions to problems they are given by users. Lindsey et al. (2025) found that when the model Claude is prompted to write rhyming couplets, it will plan a line ahead with an appropriate word which rhymes with the last word on the previous line. This and other examples suggest that LLM “thinking” may be more sophisticated than we previously imagined, and that our perception of LLMs as simply “next word predictors” may not capture the whole story.

Regardless of how LLMs “think”, their vulnerability to the types of malpractice efforts which would not fool human markers demonstrates that LLMs must at the very least think differently to humans. It might be fair to say that LLMs have a richer and more complex understanding of language than the “stochastic parrot” label would presume. However, this understanding is not “human-like”, but represents a very different form of understanding. This difference is most

apparent in the cases of malpractice which would not fool a human marker. These cases demonstrate not a lack of understanding on the part of LLMs, but a divergent form of understanding from human intelligence.

Developing robustness to the exploitation of LLM vulnerabilities

Mitigating these risks requires developing robust LLM-based automarkers which are resistant to attempts to exploit LLMs. Firstly, vulnerabilities in automarkers need to be proactively found before they can be mitigated. This is done through a process called “red teaming”, where LLMs are targeted by humans or other LLMs that attempt to bypass the target LLM’s safety mechanisms (Martineau, 2024). It is much better for an internal team to find vulnerabilities before they are found by users. Secondly, once these vulnerabilities are known they can be addressed, for example through establishing multiple layers of defence. LLM-based automarkers can be prompted to be less vulnerable to exploits, but in addition a separate machine learning model can be trained to detect potential exploits and filter these out before they reach the automarker (Sharma et al., 2025). For example, Gao et al. (2024) trained BERT and another deep learning model to classify aberrant responses in English assessments. Thirdly, even with these measures in place, there should be a recognition that no perfect exploit defence system exists. Any detection system creates trade-offs between false positive rates (incorrectly detecting an exploit) and false negative rates (incorrectly not detecting an exploit) which need to be correctly calibrated depending on the context.

Exploit detection systems would also benefit from oversight from human examiners who are aware of the potential for malpractice among students, and are able to flag any concerns they may notice. For example, if very unusual words and phrases that could cheat the system are flagged by the detection system or published online, markers would notice the repetition of these unusual words and phrases within student writing as well as noticing the repetition across different students’ work within their marking allocation. Particularly experienced markers would likely notice a difference in the essays compared with previous years. Therefore, artificial intelligence would benefit from human intelligence in this area.

Conclusion

In a recent review of the literature on automarking, Morley et al. (2025) found that ethical issues relating to the use of LLM-based automarkers were much less well covered by the research than narrower evaluations of automarking accuracy. To broaden discussions around automarking quality, this article utilised three fictitious scenarios to explore and highlight important ethical issues presented by these automarkers. The first is explainability. LLMs, due to their computational complexity, are a “black box” where the reasons for a particular mark being given are difficult to verify. Some local explainability methods such as SHAP and LIME give reliable but superficial evaluations, but advances in global methods such as mechanistic interpretability show promise in “opening up” the black box. However, explanations given by human examiners should be seen as much more trustworthy than explanations generated by LLMs. External methods like SHAP, LIME, and mechanistic interpretability are critical precisely because we cannot trust LLMs’

explanations of their own outputs. In practical terms, this means that automarking developers should implement these types of explainability techniques, to know how the automarker came to the mark which it did.

The second scenario covered bias. In exam marking, the name and background of candidates is hidden from examiners, so the specific type of bias most relevant for LLM-based automarking is when LLMs learn to be biased against demographic groups through the language patterns associated with those groups. The language used acts as a mediating factor between the LLM bias and the demographic group. LLM biases are difficult to root out at source due to the way LLMs are trained: they find statistical associations in the training data and learn from the biases present in that data. This is seen for example in the discrepancy between the overt and covert stereotyping of GPT-4 in Hofmann et al. (2024).

Despite LLM bias being difficult to root out at source, there is still plenty we can do to evaluate, and then potentially mitigate, LLM bias. Research could compare different LLM-based automarkers on their levels of bias in controlled experiments, where language which is both irrelevant to the construct being assessed and statistically associated with use by certain demographic groups is changed. Whether the LLMs then changed the marks based on this language change would be measured. In general, LLMs have strong research potential. They can be tested with many different permutations and in hundreds (or thousands) of repeated experiments at a much lower financial and time cost than with human subjects (Park et al., 2023). Potentially, specific models would only then be chosen as automarkers if they passed through this evaluation screening. When operationalised, these low-bias models could then be monitored through tools such as differential rater functioning, and through comparisons to marks given by humans after these human-given marks were screened for bias.

Bias monitoring would also draw from explainability techniques. If we understand how the mark was given, we can detect whether those reasons rested on biased decision-making. Local explainability methods such as SHAP and LIME would be useful here, as knowing the most important words which informed the mark could uncover patterns of language-based bias. For example, words with high feature importance (they are strongly associated with the mark being given) could be linked to a table of words uniquely used by certain demographic groups. This also connects to the promise of mechanistic interpretability: if we can understand how LLMs “think”, then we could catch biased decision-making while it is happening if the model is drawing from negative stereotypes to make a marking decision.

Humans are also prone to biased decision-making, however, research has found that human examiners are able to mitigate demographically induced biases. Humans also have an advantage in terms of trustworthiness. We can generally trust human examiners to account for their own biases when they claim to do so (although this does not lessen the need to actively monitor human bias in marking). This means that for automarking developers, implementing methods to evaluate and mitigate LLM bias is essential.

The third issue is adversarial attacks, where models are “fooled” into giving an output which is incorrect and unintended. These can come in the form of

adding additional words to alter the semantic meaning of text, for example non-relevant adjectives, which have been shown to artificially boost the mark given by automarkers (Filighera et al., 2024). Indirect prompt injection attacks can also be utilised, which draw on the LLM's inability to distinguish between the instruction and the student's response. This area raises interesting philosophical questions into the quality of understanding LLMs have, and we argue that LLM understanding is deep and complex but not human-like. We advise that developers draw on existing malpractice approaches to strengthen the robustness of LLM-based automarkers. Similar to the monitoring of bias, research can actively probe and test LLM-based automarkers on their vulnerabilities to adversarial attacks.

Overall, we believe that the responsible use of LLMs in automarking needs to be grounded in considerations around the ethical use of AI. This is because despite often being human-like in appearance, artificial intelligence in LLMs is different from human intelligence. This divergence opens challenges in terms of trustworthiness, explainability, bias, and vulnerability to adversarial attacks for LLM-based automarkers. Automarking quality is multi-dimensional and should not be measured on accuracy alone. For further reading, Xi (2010) provides a theoretical framework on many of the wider validity dimensions such as the impact of automakers on the constructs being assessed. Ultimately, if or when LLM-based automarkers are implemented in high-stakes settings, a central focus on AI ethics will be important to ensuring that awarding organisations continue to do the best for students and provide high-quality AI products and services.

References

- 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.
- Amodei, D. (2025). *The urgency of interpretability*. Dario Amodei.
- Anthropic. (2025). *Claude 3.7 Sonnet system card*.
- Baird, J. (1998). *What's in a name? Experiments with blind marking in A-level examinations*. *Educational Research*, 40(2), 191–202.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery.
- Bender, E. M., & Koller, A. (2020). *Climbing towards NLU: On meaning, form, and understanding in the age of data*. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics.
- Bennett, M. S. (2023). *A brief history of intelligence: Evolution, AI, and the five breakthroughs that made our brains*. Mariner Books.
- Birch, P., Batten, J., & Batey, J. (2016). *The influence of student gender on the assessment of undergraduate student work*. *Assessment & Evaluation in Higher Education*, 41(7), 1065–1080.
- Bonezzi, A., Ostinelli, M., & Melzner, J. (2022). *The human black-box: The illusion of understanding human better than algorithmic decision-making*. *Journal of Experimental Psychology: General*, 151(9), 2250–2258.
- Chapman University. (n.d.). *Bias in AI*. Artificial Intelligence (AI) Hub.
- Crisp, V. (2007, September 16–21). *Do assessors pay attention to appropriate features of student work when making assessment judgements?* [Paper presentation]. International Association for Educational Assessment Annual Conference, Baku, Azerbaijan.
- Crisp, V. (2008). *Exploring the nature of examiner thinking during the process of examination marking*. *Cambridge Journal of Education*, 38(2), 247–264.
- Crisp, V., & Johnson, M. (2007). *The use of annotations in examination marking: Opening a window into markers' minds*. *British Educational Research Journal*, 33(6), 943–961.
- Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. *Reuters*.
- Davis, L., & Papageorgiou, S. (2021). *Complementary strengths? Evaluation of a hybrid human-machine scoring approach for a test of oral academic English*. *Assessment in Education: Principles, Policy & Practice*, 28(4), 437–455.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Xiao, B., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., ... Zhang, Z. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv.

Downes, S. M., Forber, P., & Grzankowski, A. (2024). *LLMs are not just next token predictors*. arXiv.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). *Toy models of superposition*. Transformer Circuits Thread.

Elliott, V. F. (2011). *Marking time: The decision-making processes of examiners of History and English 'A' level* [Doctoral dissertation, Oxford University]. Oxford University Research Archive.

Engelhard, G. (2008). *Differential rater functioning*. *Rasch Measurement Transactions*, 21(3), 112–114.

Enright, M. K., & Quinlan, T. (2010). *Complementing human judgment of essays written by English language learners with e-rater® scoring*. *Language Testing*, 27(3), 317–334.

Filighera, A., Ochs, S., Steuer, T., & Tregel, T. (2024). *Cheating automatic short answer grading with the adversarial usage of adjectives and adverbs*. *International Journal of Artificial Intelligence in Education*, 34, 616–646.

Furniturewala, S., Jandial, S., Java, A., Banerjee, P., Shahid, S., Bhatia, S., & Jaidka, K. (2024). "Thinking" fair and slow: On the efficacy of structured prompts for debiasing language models. In Y. Al-Onaizan, M. Bansal, & Y. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 213–227). Association for Computational Linguistics.

Gao, S., Gales, M., & Xu, J. (2024). *Detecting aberrant responses in automated L2 spoken English assessment*. In C. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 96–117). Iowa State University Digital Press.

Greatorex, J., & Bell, J. F. (2002, August 28–30). *Does the gender of examiners influence their marking?* [Conference presentation]. Learning Communities and Assessment Cultures: Connecting Research with Practice Conference, University of Northumbria.

Greatorex, J., & Suto, W. I. (2006, May 21–26). *An empirical exploration of human judgement in the marking of school examinations*. [Conference presentation]. International Association for Educational Assessment (IAEA) 32nd annual conference, Singapore.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). *Alignment faking in large language models*. arXiv.

- Hasan, A., & Jones, B. (2024). [Assessing the assessors: Investigating the process of marking essays](#). *Frontiers in Oral Health*, 5, Article 1272692.
- Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). [AI generates covertly racist decisions about people based on their dialect](#). *Nature*, 633, 147–154.
- Hutson, M. (2024). [How does ChatGPT ‘think’? Psychology and neuroscience crack open AI large language models](#). *Nature*, 629, 986–988.
- International Organization for Standardization. (2020). [ISO/IEC TR 29119-11:2020, Software and systems engineering, Software testing, Part 11: Guidelines on the testing of AI-based systems](#).
- Jadhav, C. (2018, April 20). [Exam marking: How technology is improving the quality of marking](#). *Ofqual Blog*.
- Johnson, M. (2018). [Examiner feedback and learning: What are the characteristics of effective remote feedback in a hierarchic, professional context?](#) [Doctoral dissertation, University of Cambridge].
- Johnson, M., & Black, B. (2012). [Feedback as scaffolding: Senior examiner monitoring processes and their effects on examiner marking](#). *Research in Post-Compulsory Education*, 17(4), 391–407.
- Kahneman, D. (2012). *Thinking, fast and slow*. Penguin.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). [Scaling laws for neural language models](#). arXiv.
- Kotek, H., Dockum, R., & Sun, D. Q. (2023). [Gender bias and stereotypes in large language models](#). arXiv.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). [Deep learning](#). *Nature*, 521, 436–444.
- Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., & Nanda, N. (2024). [Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2](#). arXiv.
- Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., & Zhang, H. (2024). [Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models](#). *Artificial Intelligence Review*, 57, Article 243.
- Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., & Batson, J. (2025). [On the biology of a large language model](#). Transformer Circuits Thread.
- Martineau, K. (2024). [What is red teaming for generative AI?](#) *IBM Research Blog*.
- Massey, A. (1983). [The effects of handwriting and other incidental variables on GCE ‘A’ level marks in English literature](#). *Educational Review*, 35(1), 45–50.
- Meta AI. (2025, April 5). [The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation](#). *Meta AI Blog*.

Morley, F., Walland, E., & Vidal Rodeiro, C. (2025). *Transforming science marking: A scoping review of automarkers*. Research Square.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Aspell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative agents: Interactive simulacra of human behavior*. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–22). Association for Computing Machinery.

Raina, V., Liusie, A., & Gales, M. (2024). *Is LLM-as-a-judge robust? Investigating universal adversarial attacks on zero-shot LLM assessment*. In Y. Al-Onaizan, M. Bansal, & Y. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 7499–7517). Association for Computational Linguistics.

Resnik, P. (2025). *Large language models are biased because they are large language models*. arXiv.

Ruck, D., & Sutton, M. (2024). *Indirect prompt injection: Generative AI's greatest security flaw*. CETaS Expert Analysis.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Allen Lane.

Scottish Examining Board. (1992). *Investigation into the effects of the characteristics of candidates and presenting centres on possible marker bias* [Internal report].

Sharma, M., Tong, M., Mu, J., Wei, J., Kruthoff, J., Goodfriend, S., et al. (2025). *Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming*. arXiv.

Staats, C. (2016). *Understanding implicit bias: What educators should know*. *American Educator*, 39(4), 29–34.

Suto, I., Grotorex, J., & Nádas, R. (2009). *Thinking about making the right mark: Using cognitive strategy research to explore examiner training*. *Research Matters: A Cambridge Assessment publication*, 8, 23–32.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C. D., Summers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., & Henighan, T. (2024). *Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet*. Transformer Circuits Thread.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. arXiv.

Walland, E. (2021, November 3–5). *Remote marking of high-stakes examinations*:

Leadership, challenges and strategies [Conference presentation]. 22nd Annual Meeting of the Association for Educational Assessment – Europe (AEA-Europe), Online.

Wilson, K., & Caliskan, A. (2024). *Gender, race, and intersectional bias in resume screening via language model retrieval*. arXiv.

Wong, M. (2025). *The AI industry is radicalizing*. The Atlantic.

Xi, X. (2010). *Automated scoring and feedback systems: Where are we and where are we heading?* *Language Testing*, 27(3), 291–300.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). *A comparison of two scoring methods for an automated speech scoring system*. *Language Testing*, 29(3), 371–394.

Xu, J., Jones, E., Laxton, V., & Galaczi, E. (2021). *Assessing L2 English speaking using automated scoring technology: Examining automarker reliability*. *Assessment in Education: Principles, Policy & Practice*, 28(4), 411–436.

Xu, J., Schmidt, E., Galaczi, E., & Somers, A. (2024). *Automarking in language assessment: Key considerations for best practice* (Cambridge Papers in English Language Education). Cambridge University Press & Assessment.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). *Explainability for large language models: A survey*. arXiv.