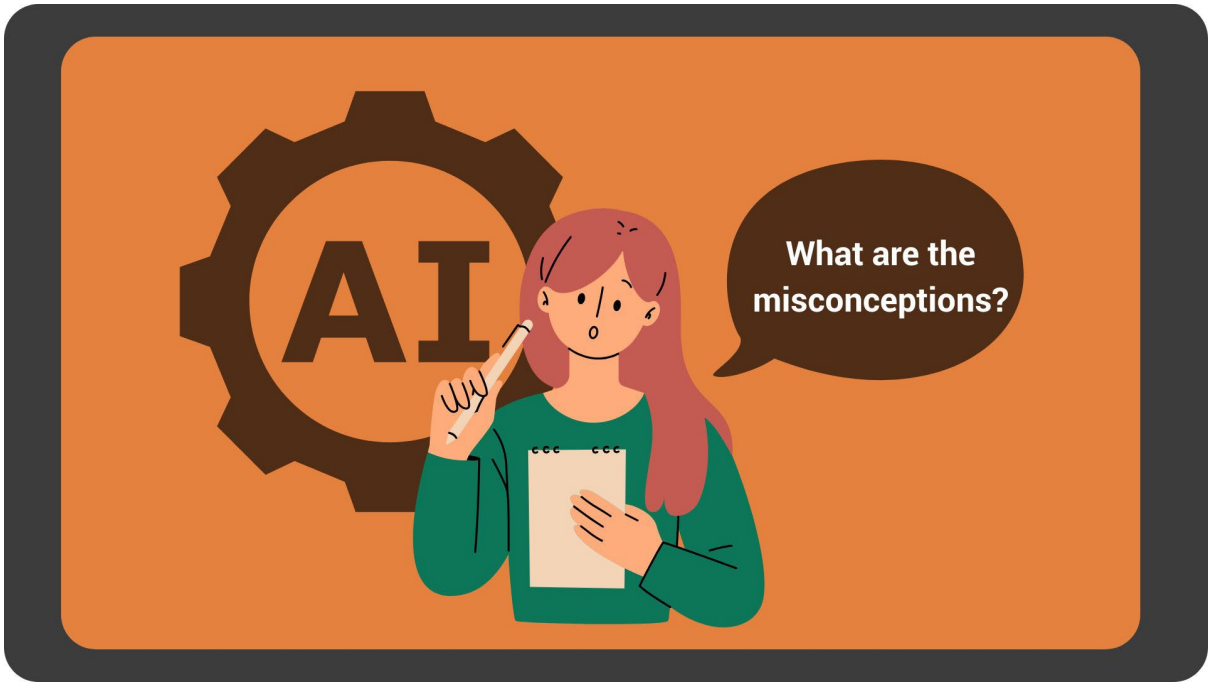
Unpacking AI assessments: Managing common misconceptions

Article #9 of AI in Education Article Series: May 2025



Artificial Intelligence (AI) opens up exciting opportunities in assessment and learner support. However, its successful use depends on understanding and addressing common misconceptions – both those held by learners and educators. This article explores **four common misconceptions relating to the use of AI for assessments**, as well as what education providers can do to manage them.

This article is the **ninth in a series titled "AI in Education"**, aimed at education providers interested in AI. The intention is for this series to act as a beginner's guide to the use of AI in education, with a particular focus on AI agents. This series is being developed as part of a project to develop an AI agent for learner oral assessment, funded by the Food and Fibre Centre of Vocational Excellence (FFCoVE). We invite you to follow along as we (Scarlatti) document our learnings about this exciting space.

Note that this article reflects the views of Scarlatti as of May 2025. They do not necessarily represent the views of the Food and Fibre Centre of Vocational Excellence.

Misconception 1: AI is unbiased OR AI is highly biased

The misconception

Some think that AI systems are neutral and objective, assuming they do not carry bias. Others think that AI is highly biased.

The reality

In practice, the level of bias is somewhere in the middle, but improving. This is because AI systems reflect the data that they are trained on - including any historical biases within that data (Wellner, 2020) and how they are aligned. These systems can also lack understanding of Indigenous knowledge such as Mātauranga Māori, as they're usually trained on data from dominant Western contexts (Ministry of

Education, 2024). This means that the way that individual AI systems are designed and trained directly impacts whether its models produce biased or non-biased outputs.

As of May 2025, audits show that the newest large language models display smaller but still measurable demographic biases. Armstrong et al. (2024) documented sizable gender and race gaps in GPT-3.5. A more recent multi-modal audit by Gaebler et al. (2025) finds that overt hiring-style gender preferences are now only a few percentage points, indicating progress, though not full elimination. Meanwhile, subtler stereotype and intersectional effect persist, as shown by Bai et al. (2025) and Salinas et al. (2025). We suggest that it remains essential to consult the newest evaluations for each model and to run context-specific audits before deploying an agent.

Possible mitigations

- **Run a model evaluation specific to your context** – These are evaluations done to test and refine a system (such as an AI model). This could be done before deciding to use a model, including to see how appropriate it is for your cultural context. It could also be done when any updates are made, to detect improvements or regressions.
- **Identify the level of human oversight needed** – Consider the likelihood for instances of bias with your learners and assessment, and balance this with the time required for human checks.
- **Decide whether AI is suitable** – Using the above, you will need to decide how suitable the assessment is for AI.
- **Put human checks in place in accordance with risk –** If going ahead with AI assessment, human checks on assessments could range from encouraging tutors to check and correct grading, to enforcing randomised checks, or enforcing checks of every grade.

Misconception 2: AI is always correct OR AI is useless

The misconception

Some believe that AI outputs are always accurate and reliable. While others believe that AI is useless for assessments.

The reality

Like misconception 1, the reality is somewhere in the middle of these misconceptions. AI's accuracy (and therefore its useability) largely depends on the quality of the training data and the design of the model. Over the last year, AI models have only gotten more accurate, with notable improvements seen in the quality of training data and algorithmic efficiency.

During our oral assessment agent pilots, we found that tutors agreed with the AI grade at least 94% of the time. The most common cause of AI giving incorrect grades or feedback has been when there is vagueness or missing information in the *original* content used to prompt the AI (in our case, grading rubrics, course content and past answers). However, there are likely still differences between an AI grader and a human grader. For example, a human teacher can adjust for context, learning needs, or misunderstandings, and reward points for originality, insight or learner improvement over time.

Possible mitigations

- **Identify the level of human oversight needed** – Consider the ease of assessment (i.e., whether answers are clearly right or wrong); the risk of incorrect assessment (i.e., how many credits the assessment is worth, and the level); and balance this with the time required for human checks.
- **Decide whether AI is suitable** – Based on the above, you will need to decide how suitable the assessment is for AI.
- **Put human checks in place in accordance with risk –** If going ahead with AI assessment, human checks could range from encouraging tutors to check and correct grading, to enforcing randomised checks, or enforcing checks of every grade.

Misconception 3: AI companies misuse your data

The misconception

Users worry that AI systems will automatically collect and misuse their personal data during their assessment.

The reality

As of May 2025, most AI models now offer control over data sharing. For example, Claude does not use user data for training by default. ChatGPT allows users to opt-out in its settings. Scarlatti's agent has been built using a secure API connection to OpenAI models, which do not use user data for training by default. Despite this, we acknowledge that there are misconceptions (and therefore concerns) about OpenAI's use of data.

Possible mitigations

- **Choose a model service that does not train with your data** – Use a model or subscription that does not use your inputs for training data. For example, any of the options mentioned above.
- **Consider other options to make users comfortable** – As mentioned, certain models do not use your input for training. However, to further ensure user comfort, consider mitigations like decoupling the AI agent from where learners' names are stored; not having it ask for personal information (e.g., location, employer); and avoiding assessments that are personal in nature (e.g., a reflections log).
- **Teach educators to understand and explain data protections** – Make sure that the team piloting the agent understands how it manages data and could answer questions learners may have about data privacy. They are critical to rolling out any innovation to learners.
- **Consider the pros and cons of giving the option of *not* using AI –** Allowing learners to not use AI (e.g., complete the assessment in writing instead) may create additional complexity for learning providers. However, it may comfort learners, especially if these misconceptions continue.
- **Embed resources for learners on data protection** – Consider embedding a high-level video and detailed information sheet. This could include how to use the agent for assessment, but also how data is protected (with links to the AI model's data policies). Ideally, this video would be made by someone trusted by the learner, with the support of a technical expert.

Misconception 4: AI in assessments is purely a cost-cutting tool

The misconception

AI is only used to convert existing assessments into a new format (e.g., written to oral) to reduce costs.

The reality

While AI can reformat content (e.g., turning a written quiz into a voice-based quiz) to reduce costs, we think its greatest potential is in reimagining assessment. Oral AI allows for assessments that are more dynamic, conversational, and aligned to the real-world skills we want learners to build — such as verbal reasoning, negotiation, and critical reflection.

Possible mitigations

- **Undertake a 'reimagining' session** – Encourage educators to start from the learning outcomes and ask, "what would a more authentic and practical assessment look like?". Alternatively, you could prompt AI to reimagine the existing assessment.
- **Balance the pros and cons in your final decision** – This is not to say every assessment should become an AI-powered role play. Consider your reasoning for using AI, your learners, and the type of assessment. Ensure the pros outweigh the cons enough for change to be worthwhile.

Misconception and mitigation summary

Misconception	Mitigations
AI is unbiased OR AI is highly biased	• Run a model evaluation specific to your context • Identify the level of human oversight needed • Decide whether AI is suitable Put human checks in place in accordance with risk
AI is always correct OR AI is useless	• Identify the level of human oversight needed • Decide whether AI is suitable • Put human checks in place in accordance with risk
AI companies misuse your data	• Choose a model service that does not train with your data • Consider other options to make users comfortable • Teach educators to understand and explain data protections • Consider the pros and cons of giving the option of *not* using AI • Embed resources for learners on data protection
AI in assessments is purely a cost-cutting tool	• Undertake a 'reimagining' session • Balance the pros and cons in your final decision

Scarlatti's take

There are a number of misconceptions around AI in education, which may influence whether you want to adopt AI or not. Importantly, these misconceptions are evolving, as a result of the AI models themselves changing, but also our shared understanding of them – whether fuelled by evidence or not. It is highly likely that shortly after publishing this article, the situation will change all over again.

Questions that we are asking for our own AI agent:

- How can we continue to scan for, understand and critically evaluate misconceptions?
- What should be done in future pilots to manage these misconceptions?
- How can we continue to facilitate open discussions on misconceptions with providers?

Interested in following our journey into AI?

- Sign up to receive our next article directly to your inbox.
- Contact the Scarlatti team to share your thoughts or questions.

References

Armstrong, L., MacNeil, S., Liu, A., & Metaxa, D. (2024). The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring. Arxiv, Cornell University. Retrieved 24 April 2025, from https://arxiv.org/pdf/2405.04412

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. (2025). Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Science of the United States of America 122* (8). Retrieved 24 April 2025, from https://pubmed.ncbi.nlm.nih.gov/39977313/#:~:text=independently%2C%20are%20more%20diagnostic%20of,unbiased%20according%20to%20standard%20benchmarks

Department of Internal Affairs, National Cyber Security Centre & Stats NZ. (2023a). Initial Advice on Generative Artificial Intelligence in the Public Service. Retrieved January 8, 2025, from https://www.digital.govt.nz/assets/Standards-guidance/Technology-and-architecture/Generative-AI/Joint-System-Leads-tactical-guidance-on-public-service-use-of-GenAI-September-2023.pdf

Gaebler, J., Goel, S., Huq, A., & Tambe, P. (2025). Auditing large language models for race and gender disparities: Implications for artificial intelligence-based hiring. *Behavioural Science and Policy 11* (1), 1-10.

Ministry of Education. (2024, November 25). *Generative AI: Guidance and resources for education professionals on the use of artificial intelligence in schools.* Retrieved 24 April 2025, from https://www.education.govt.nz/school/digital-technology/generative-ai

Office of the Privacy Commissioner. (2023). Artificial intelligence and the information privacy principles. Retrieved January 8, 2025, from https://privacy.org.nz/assets/New-order/Resources-/Publications/Guidance-resources/AI-Guidance-Resources-/AI-and-the-Information-Privacy-Principles.pdf

Salinas, A., Haim, A., & Nyarko, J. (2025). What's in a Name? Auditing Large Language Models for Race and Gender Bias. Stanford University Human Centred Artificial Intelligence. Retrieved 24 April 2025, from https://hai-production.s3.amazonaws.com/files/2024-06/(Nyarko,%20Julian)%20Audit%20LLMs.pdf

Wellner, G. P. (2020). When AI is gender-biased: The effects of biased AI on the everyday experiences of women. *Humana Mente, 13*(37), 127-150. Retrieved 24 April 2025, from https://www.humanamente.eu/index.php/HM/article/view/307/273