Piloting AI in oral assessment: 5 practical lessons for education providers

Article #10 of AI in Education Article Series: June 2025

In early 2025, Scarlatti developed and piloted an Artificial Intelligence (AI) agent for oral assessment. We hypothesised that this technology could improve outcomes for students with learning difficulties, who are neurodiverse, or who speak English as a second language. As of June 2025, pilots are complete and an evaluation underway. This article shares the **5 practical lessons** education providers can take when piloting their own similar agent.

This article is the **tenth 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 has been 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. We invite you to **get in touch** to see a demo or explore how you could use AI agents in your own context.

Overview of our pilots

From March to May 2025, Scarlatti undertook two pilots of our AI agent for oral assessment – one with Fruition Horticulture Limited, and one with Dairy Training Limited. These are summarised below.

	Fruition Horticulture Ltd	Dairy Training Ltd
Course	Hei Whanaake	Contract Milking 101
Assessment	Health and Safety Assessment (level 2)	Contract Milking Assessment (level 5)
No. of students	14	11

Student demographics	Predominantly Māori, but also Samoan and Tongan Aged ~16	Mixed ethnicities Aged ~22 to ~40
Aims of provider	Inclusive for young Māori learners, illiterate learners + saves tutor time	Inclusive for remote and / or neurodiverse learners + saves tutor time
Number of questions	19	5
Number of learner responses	284 unique learner answers	55 unique learner answers

Below, we share five key learnings from this project.

Lesson 1: Moderation and system integration are key

The first lesson is the importance of assessing all aspects of feasibility before starting. In our pilots, two important factors were whether moderation would allow the AI agent to ask follow-up questions, to provide positive reinforcement, and feedback; and whether the agent could be easily integrated into the education provider's learner management system (LMS).

Possible solutions

- Examine moderation requirements – These requirements may influence the design of the assessment agent. For example, if they restrict the agent from asking follow-up questions during the conversation or from providing feedback, it may not justify further investment.
- Discuss integration with your LMS provider – Your LMS may not integrate easily with new AI products. It will be important to assess how challenging this process will be. If there are few ways forward, you may decide not to go ahead with using an AI agent.

Lesson 2: Involve teaching and QA staff

In your scoping phase, it is also important to bring the relevant staff onboard. We found that teaching staff and those with expertise in Quality Assurance were valuable in ensuring the AI agent met real needs and met moderation requirements. Teaching staff whose questions had been well addressed were more comfortable piloting the agent with their learners.

Possible solutions

- Include teaching and QA staff in early planning discussions – For example, include them in initial development discussions, and assessment redesign discussions.
- Provide clear information on the AI product – This means briefing staff and doing a demo of the agent. Staff should be allowed to try the agent themselves and ask questions. They should also be provided with more detailed information on why the agent has been built and how it works.
- Involve staff in interpreting findings – After collecting learner feedback and assessing how accurately the agent graded answers, we recommend sharing this back with staff and deciding next steps together.

Lesson 3: Reimagine your assessment ground up

Our design phase highlighted how easy it is to 'reformat' your existing assessment questions to be oral. However, this does not unlock the real capabilities of AI, which could enable styles of assessment that were previously economically infeasible, or inconceivable. For example, it could conduct the assessment as a roleplay where the AI plays an employer, colleague, customer or client; it could conduct a debate with the learner; or it could ask the learner to swap between languages.

Possible solutions

- **Consider redefinition deeply** – As mentioned, AI could be used to enable new styles of assessment that were previously economically infeasible, or inconceivable (see the ISAR model for a framework). This requires both creative thinking to come up with the ideas. We suggest bringing in the right people, and thinking without limits.
- **Consider redefinition critically –** Next, weigh up the costs of changing your assessment dramatically with the perceived benefits. Such dramatic changes may only be worth it in capstone assessments, or assessments where the current format causes significant issues or limitations.

Lesson 4: Balance your agent priorities

During development, we found two balances had to be struck. **Accuracy vs flexibility**: a 'highly accurate' agent can produce a fail grade for a learner missing a certain word despite them otherwise understanding the material, whereas a 'highly flexible' agent may be so open to interpreting a learner's answer that they may be overly generous with grades. **Follow-up ability vs answer security**: An AI agent that has complete access to assessment answers would consistently ask follow-up questions to incorrect and vague answers, but may 'give away' the answer in that follow-up. In contrast, a secure agent may have no access to answers, making it incapable of asking follow-ups.

Possible solutions

- **Consider splitting AI into examiner and assessor** – By having one agent examine (i.e., run the conversation) and one that assesses (i.e., grades answers), you can prevent the examiner from accidentally giving the learner the answer, but, you will hinder follow-up ability.
- **Trial different ways to achieve consistent follow-ups** – This might look like providing explicit instructions to the agent, or having the examiner call an intermediate assessor agent after each response and having that agent tell the examiner whether and how to follow up.
- **Be ready to edit your training content** – Our pilots showed that most of the times that staff disagreed with the AI's provisional grade, it was because the training content (e.g., assessment rubric, exemplars) contained elements that the AI over-relied on (e.g., a keyword). You can address this by editing the training content post-pilot.

Lesson 5: Test in low-stakes environments

The pilot phase revealed the value in testing the agent in low-stakes environments. This is important because things can still go wrong in a pilot and students will be more anxious if the assessment is important to their final grade/results.

Possible solutions

- **Give staff multiple pilot options** – Options could range from demonstrating the agent to learners; having students use it for practice; or having them use it for their actual assessment, either with or without supervision. This will depend on staff's comfort with AI.
- **Be available to answer questions –** Questions could come from staff or learners. For example, one learner was concerned about how the agent would use their data, but the tutor was unsure how to navigate this. Being there to support can help answer these questions.

Next steps for providers

This series is coming to an end in July 2025. Our final articles will share results from our pilots – including grading accuracy, estimated time savings and student/staff feedback.

As we wrap up our pilot, we invite you to:

- Reflect on how AI could support you or your learners
- Contact Scarlatti for a demo or an obligation-free chat on AI agents.

References

Bauer, E., Greiff, S., Graesser, A., Scheiter, K., & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review, 37*, 1-27. 10.1007/s10648-025-10020-8.