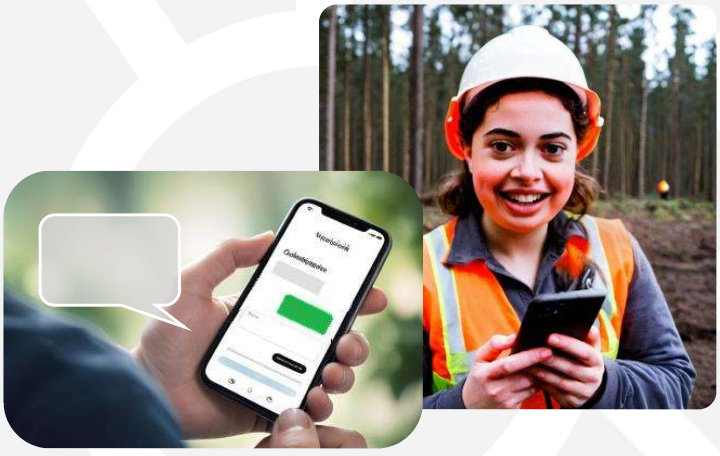
SCARLATTI
Evaluation | Analytics | Insights

Prepared for:
Food and Fibre Centre of Vocational Excellence

AI Agent for Oral Assessment Playbook

A guide for providers

August 2025



Scarlatti.co.nz

Contents

Introduction

Context

Assessment is a defining feature of formal education. Yet written assessments present a range of barriers for vocational education learners (e.g., due to personal preferences, previous negative experiences at school, literacy levels, cultural suitability, neurodiversity, or speaking English as a second language). This is widely believed to inhibit enrolments, learner progress, and completion rates. Some education providers will use oral assessments instead, but these are costly to administer given their one-to-one nature, particularly if learners are located remotely or in the workplace.

In response, the Food and Fibre Centre of Vocational Excellence commissioned Scarlatti, a company specialised in research, evaluation and analytics, to build, pilot and evaluate an AI agent for oral assessment in 2025. This proof-of-concept would indicate whether AI could administer inclusive, effective and efficient assessments.

This playbook

This playbook is for education providers considering using an AI agent for assessment. We encourage you to use it to:

- Assess whether an AI agent for assessment could be suitable for your context
- Learn how to run a pilot and evaluation of such an agent
- Take the lessons we have learnt towards developing your own AI agent for assessment.

Contact us

Have more questions on this?

Feel free to contact Scarlatti (www.scarlatti.co.nz or adam.barker@scarlatti.co.nz).

What is an 'AI agent for assessment'?

This section provides an overview of broader progress towards AI for assessment, a definition for what an 'AI agent for assessment' is, and a number of ideas about the types of AI agents for assessment that could exist – the first of which is the focus for the remainder of this playbook.

The use of AI in assessment

As part of this work, a global scan for similar products or projects was undertaken every second month between December and July 2025.

This found that there was a rapid growth of AI products for education in the first half of 2025 – although many were 'in development' rather than anything 'mature'. Across Oceania, many AI agents are being developed using the University of Sydney's 'Cogniti', with only a smaller number being built from scratch. Some products are being developed to undertake a singular role (e.g., tutor) while others are undertaking a combination (e.g., tutor, administrative support, quiz generator).

Despite these advancements, there are very few products using AI *for assessment.* These exceptions include:

- NZQA using AI to grade written assessments
- Cogniti developed agents for learners for formative assessments
- ConCOVE developing an agent to generate assessments
- SchoolJoy releasing an agent for non-formal assessment
- Scarlatti developing an agent for oral summative assessment (the focus of this playbook).

For a list of products and projects in Oceania, see the Scarlatti website.

Defining an 'AI agent for assessment'

Given the newness of this space, we propose to define an AI agent for assessment as 'a tool that can **simulate a human-like conversation** to collect information on **a person's knowledge of a topic.'**

At a high level, this means that any 'AI agent for oral assessment' welcomes the user, asks them a series of questions, compares their answers to pre-loaded materials, and asks follow-up questions as required.

Types of AI agents for assessment

We imagine that there could be various types of AI agents for assessment:

- **Summative assessment agent** - This would be an agent that runs an assessment with a learner to assess their knowledge, compares their answers to assessment rubrics, course materials and exemplar answers, determines a grade for the learners answer and send the grades, transcripts and recordings to the education provider.
- **Formative assessment agent** – This would be an agent that runs an assessment with a learner for the purpose of learning before a final assessment. The agent would use the assessment

rubrics, course materials and exemplar answers to provide the learner with feedback. It may also give the learner a provisional grade for learning purposes.

- **Pre-screening agent** – Before being placed in a course, an agent would run an assessment with a learner to determine whether their knowledge meets the prerequisites for the course, and whether interests/goals align with the course. The agent would send the result to the course administrator. It may also provide the learner with immediate feedback.
- **Employer-verifier agent –** During work-based learning, employers are often required to write a 'verifier' report on the learner's progress in the workplace. An agent could ask employers relevant questions on this and then develop and deliver a report on the employers' answers in the education providers required format.
- **Recognition of Prior Learning agent** – Many people in vocational careers have accumulated enough knowledge through work experience to warrant receiving a qualification. An AI agent could speak with such people to assess their knowledge and send the result to the education provider. It could also provide the user with information on what more they may need to demonstrate their knowledge.

Scarlatti's proof-of-concept (the focus of this document) was of an *oral, summative* agent.

Should we be using an AI agent for assessment?

This section provides various quick decision lists to help you identify whether an AI agent for oral assessment is right for you.

While Artificial Intelligence may offer new opportunities to make assessments more efficient, effective or inclusive, we suggest this does not mean it should be used in all contexts. For example, it is important to first ensure learning outcomes and assessments are designed in a way that are meaningful, and to manage educator workloads. Moreover, some assessments simply aren't suited to an AI-run conversation.

Below, we provide a series of questions to have you identify whether AI could be suitable for your assessments, whether it should be oral AI, and whether you should build it yourself.

Decide to use AI for assessment

The following questions are to help you decide whether AI is right for your organisation, learners and assessment. The more questions you answer **"Yes"**, the more appropriate AI may be.

	Decision question	Guidance
1	Do you have a clearly defined assessment pain point?	We suggest starting with the problem and then finding an appropriate solution.
2	Does AI have the potential to address this better than current options?	AI may be suitable if current available options are considered inefficient, ineffective or non-inclusive.
3	Can you integrate with your LMS?	Your agent likely needs to integrate with your LMS and if this is not possible, it can involve substantial development time.
4	Will users have access to an internet connection?	An internet connection is needed to use any agent.
5	Do moderation rules permit follow-up questions and AI-led feedback?	Rules that prevent either of these could block the value of an oral AI agent.
6	Are your staff onboard with exploring AI?	Learner-facing staff are critical to building and implementing an agent for students. They need to believe in the product to recommend it to learners.

Decide if oral AI is best

The following questions are to help you decide whether oral AI would be better for your organisation, learners and assessment. The more questions you answer **"Yes"**, the more appropriate oral AI may be.

	Decision question	Guidance
1	Do you have a high proportion of learners who struggle with writing?	An oral AI agent may be more suitable for learners who struggle to write due to personal preferences, previous negative experiences at school, literacy levels, cultural suitability of writing, neurodiversity, or speaking English as a second language.

2	Is oral interaction pedagogically suitable for this assessment?	Oral AI agent may suit assessments that involve knowledge recall, critical thinking, communication, or scenarios. It is unlikely to suit assessments involving calculations, forms or reflective journalling.
3	Do staff have capacity and willingness to redesign assessments?	An oral assessment may be structurally quite different to your existing written assessment. Ideally, staff are open to rebuilding the assessment from the ground up, to maximize AI potential.

Decide if you should build it

The following questions are to help you decide whether your organisation should build its own AI agent for assessment (rather than finding an existing tool or hiring a third party). The more questions you answer **"Yes"**, the more appropriate oral AI may be.

	Decision question	Guidance
1	Would a custom product better meet your needs than existing tools and platforms?	As of the time of writing, there was no tool or platform for non-developers to create an AI agent for assessment, but this could change.
2	Do you have the necessary skills in-house to develop your own?	Developing such an agent requires strong developers, given there is no platform suitable for non-developers currently.
3	Are you prepared to take on the cost of ownership?	You will need to consider cost to development, cost per assessment, and cost of hosting.

How did we develop and pilot it?

This section gives you an overview of the two pilots, the methodology used for the project, and our findings.

Pilot description

From March to May of 2025, Scarlatti undertook two pilots of their AI agent for oral assessment – one with Fruition Horticulture Ltd and one with Dairy Training Limited (DTL). An evaluation was undertaken on both pilots. The two pilots are compared below.

	Fruition Horticulture Ltd	Dairy Training Ltd
Course	Hei Whanake	Contract Milking 101
Assessment	Health and Safety Assessment	Contract Milking Assessment
Level	Level 2	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
Format change	Written assessment à Oral assessment AI agent	Oral assessment with tutor à Oral assessment AI agent
Conversational ability	Oral, female voice Follow-up question at end No feedback	Oral, male voice Follow-up questions during Feedback at end
Number of questions	19	5
Number of learner responses	284 unique learner answers	55 unique learner answers

Methodology

Below we describe the methodology for piloting and evaluating the AI agent for oral assessment. Note that this is a self-evaluation and should be treated as such.

Preparation

The preparation phase began with the development of research questions, a logic model and measures table (see page 30 and 41 for the latter two). The intention of this was to ensure the aims of the tool were clear, as well as what would need to be collected to assess to what extent the tool achieved these aims. Scarlatti undertook a review of government guidance on AI including plans for how to mitigate key risks. Scarlatti also sought and received ethics approval from its internal ethics committee to proceed with pilot testing.

Research questions

The following research questions guided this project:

1. Can current voice AI models be used to conduct a verbal assessment of learners?
2. Can AI accurately assess learners' answers to exam questions?
3. What are effective prompting techniques for designing an assessment agent?
4. What are the most effective methods for giving the agent domain-specific knowledge?
5. How can we store data (recordings, transcripts and assessment) securely?

Agent development

Scarlatti then developed the agent. We describe key aspects of its design below.

'Double agent' architecture

Although we refer to the agent in the singular, it is in fact made of *two* agents:

- **Examiner agent** – This is the agent that talks to the learner. It uses OpenAI Realtime Voice to do this, enabling a 'natural' feeling oral conversation.
- **Assessor agent** – This is the agent that assesses the learner's answers. It uses the GPT-4o text model to compare the answer to uploaded materials and provide a grade and reason for the grade, for each question.

Set assessment questions, flexible follow-ups

We provided the examiner agent with the assessment questions that are currently used by DTL and Fruition in these assessments. The agent could run through these with the learner, and decide when a follow-up question was needed – typically when the learner's answer was vague or incorrect. It can ask this follow-up immediately after the learner's answer, or at the end of the conversation depending on set up.

Retrieval-augmented generation (RAG)

We provided the assessor agent with prompting materials. Specifically, the assessment rubric, example answers and grades, and course content. The agent could use RAG to access these when assessing

answers. This involves the agent pulling the most relevant chunks from the rubric, example answers and course materials when needed.

Core guard-rails

The agent needed to have guard rails to protect the learner and to ensure credible assessment. For example, to protect the learner we:

- Used the OpenAI API so that data would not be used to train models
- Asked their name / student ID outside of the agent so no names would be sent to OpenAI

To ensure credible assessment, we:

- Used a double agent architecture to protect answers (see above)
- Stored transcripts and audio recordings that tutors could check
- Prevented the AI agent from changing into languages other than English/Te Reo
- Explored summative options where no hints or feedback could be given, and formative options where they could.

Outputs

The outputs at the end of the assessment were:

- Feedback for the learner
- Transcripts of the conversation
- Audio recordings of the conversation
- AI-produced provisional grades
- Reasoning for the AI-produced provisional grade.

Below we provide an image of what the tutor-facing results look like (see Figure 1).

Figure 1: Example export from AI agent after running an assessment with six learners, graded either C (competent) or NYC (not yet competent).

Course ID	Student name	Question ID	Student answer	AI grade (Provisional only)	AI feedback	Tutor grade (Overrides AI grade)	Tutor comment
cm101	Student #1	q1	The conditions of the cows and the amount of feed at	C	The student correctly identifies a common issue in c	C	
cm101	Student #1	q1b	The guidelines and the rules are written down in the c	NYC	The student's answer is too vague and does not ref	C	He did actually
cm101	Student #1	q2a	Independent people who can do the body condition s	C	The student identifies independent experts for bod	C	
cm101	Student #1	q2b	They can judge the conditions and measure them and	C	The student provides a reasonable explanation for	C	
cm101	Student #1	q3	Make a budget and know what kind of premium you v	C	The student mentions preparing a budget and unde	C	
cm101	Student #2	q1	Not able to achieve targeted production and quality d	C	The student identifies issues such as animal health,	C	
cm101	Student #2	q1b	Covered under Federated Farmers Contract Milker bo	C	The student references the Federated Farmers Con	C	
cm101	Student #2	q2a	Farm owner, consultant, accountants, lawyer, previou	C	The student lists a variety of potential sources for a	C	
cm101	Student #2	q2b	They provide advice, help with budgeting, decision-ma	C	The student explains the reasons for consulting the	C	
cm101	Student #2	q3	Prepare for negotiation with evidence, use correct too	C	The student mentions preparing with evidence, usi	C	
cm101	Student #3	q3	Use a premium calculator, consider weather issues, ar	NYC	The student's answer is vague and does not include	NYC	
cm101	Student #3	q1	Communication errors between owner and contract r	NYC	The student's answer of 'communication errors' is r	NYC	In this case I
cm101	Student #3	q1b	Following the Federated Farmers Contract Milking Agr	NYC	The student did not specify any particular clause or	NYC	
cm101	Student #3	q2a	Talk with the owner, contact an accountant or consult	C	The student correctly identifies potential sources of	C	
cm101	Student #3	q2b	Because they provide good advice and are knowledge	NYC	The student's answer lacks depth and does not exp	NYC	Once again I
cm101	Student #4	q1	Production numbers not accurate, meaning not hittin	C	The student correctly identifies production number	C	
cm101	Student #4	q1b	The contract guidelines or rules for the issue are foun	C	The student references clause 10 under target prod	C	
cm101	Student #4	q2a	Go to your farm owner for help, or if that's an issue, g	C	The student lists the farm owner, accountant, lawye	C	
cm101	Student #4	q2b	Farm owner is the best place to go because you're wo	C	The student provides a rationale for consulting the	C	
cm101	Student #4	q3	I would prepare by, to renegotiate, by showing a budg	C	The student mentions preparing a budget and cons	C	
cm101	Student #5	q1	Feed levels at the start or end of the term.	C	The student correctly identifies feed levels at the st	C	
cm101	Student #5	q1b	Use the same measuring device to measure feed level	C	The student provides a detailed answer referencing	C	
cm101	Student #5	q2a	Consultant, Federated Farmers, or accountant.	C	The student correctly identifies potential sources of	C	
cm101	Student #5	q2b	Because they specialize in this field and have informat	C	The student explains that these sources are speciali	C	
cm101	Student #5	q3	Run numbers, do budgets, and cash flow; consult with	C	The student outlines a comprehensive preparation	C	
cm101	Student #6	q1	Body condition score does not meet the required con	C	The student correctly identifies a common contract	C	
cm101	Student #6	q1b	The contract guidelines can be found in clause 14.	C	The student references clause 14 for the contract gu	C	
cm101	Student #6	q2a	The farm owner, consultant, lawyer, or an accredited	C	The student lists several appropriate sources of adv	C	
cm101	Student #6	q2b	Because they understand what I'm going through, and	C	The student provides a reasonable explanation for	C	
cm101	Student #6	q3	Take with me a good budget, write up a budget and ta	C	The student mentions preparing a budget, which is	C	

Pilot implementation

Following approval from Scarlatti's internal ethics committee, we selected pilot partners (Fruition Horticulture Ltd and Dairy Training Ltd) and worked collaboratively with each to identify a suitable assessment for the pilot—this was an assessment where oral interaction would add value, and which would fit the pilot timeline.

Tutors at each provider were briefed on how to use the AI agent and how to facilitate the pilot. They were given the choice of three pilot modes (although both organisations chose the last option):

- **Demo mode:** Tutors demonstrate the agent to learners using prepared examples before facilitating a group feedback session.
- **Practice mode:** Learners use the AI agent for practice rather than assessment. This is followed by feedback collection.
- **Assessment mode**: Learners use the AI agent in place of a formal assessment, supervised by tutors who recorded grades, asked follow-up questions, and collected feedback.

An information sheet for learners was also provided to explain the pilot's purpose, what participation involved, and how their data would be used.

Finally, the provider introduced the AI agent concept to learners and supervised them as they undertook their assessment (in one instance, via Teams call).

Data collection and analysis

After pilots were finished, we collected feedback via multiple sources:

- **Learner feedback** via surveys and / or tutor-led group discussion.
- **Tutor observations,** including grading comparisons between AI and tutor-assigned grades.

- **Tutor feedback** via surveys, emails and Scarlatti-led group debrief.
- **Technical logs** to identify functional issues and areas for refinement.

Lessons integration

Throughout the pilot period, lessons learnt were regularly logged and used to refine the AI agent, improve tutor support materials, and shape the Development Map in this playbook (see page 21).

Interpretations

Two factors should be considered when interpreting the findings from these pilots:

- **This was a proof-of-concept** – Funding for this project was to develop a proof-of-concept. This means that the stages that would only be required to develop a production version have not been undertaken, and we therefore do not make recommendations on how to do them in this playbook.
- **Sample size was small** – There were a total of 25 students and 9 education provider representatives (operation manager, tutors, academic manager and QA lead) in these pilots. This is a small sample, and should therefore be taken when interpreting results.

What do pilots show?

Below we explore the findings from the evaluation of the AI agent for oral assessment. This is informed by the original logic model (see page 30). We provide an overall take on each finding, then explore the result by pilot organisation.

Accuracy and trustworthiness

Both pilots resulted in a 95% match between the preliminary grade given by the agent and the grade given by human tutor (see Figure 2). We suggest that this accuracy is likely high given that two human tutors are unlikely themselves to give the same grades 100% of the time.

Where there were non-matches, it was mostly when the AI had graded someone as not yet competent and the tutor disagreed (see Figure 3). This tended to be because of an inaccuracy in the prompting materials which was misleading the AI agent (e.g., a detail within the assessment rubric, the exemplars or course materials, which could be fixed in future). In other instances, there will have been true AI error, but we suggest this is small.

While there was some nervousness about trying a new technology (especially given that assessments are already stressful for learners), by the end of pilots, the majority of learners and tutors were confident in the AI agents ability.

Aside from this, the agent passed assessment moderation through Fruition.

Figure 2: Percentage match between AI produced grade and the human tutor produced grade.

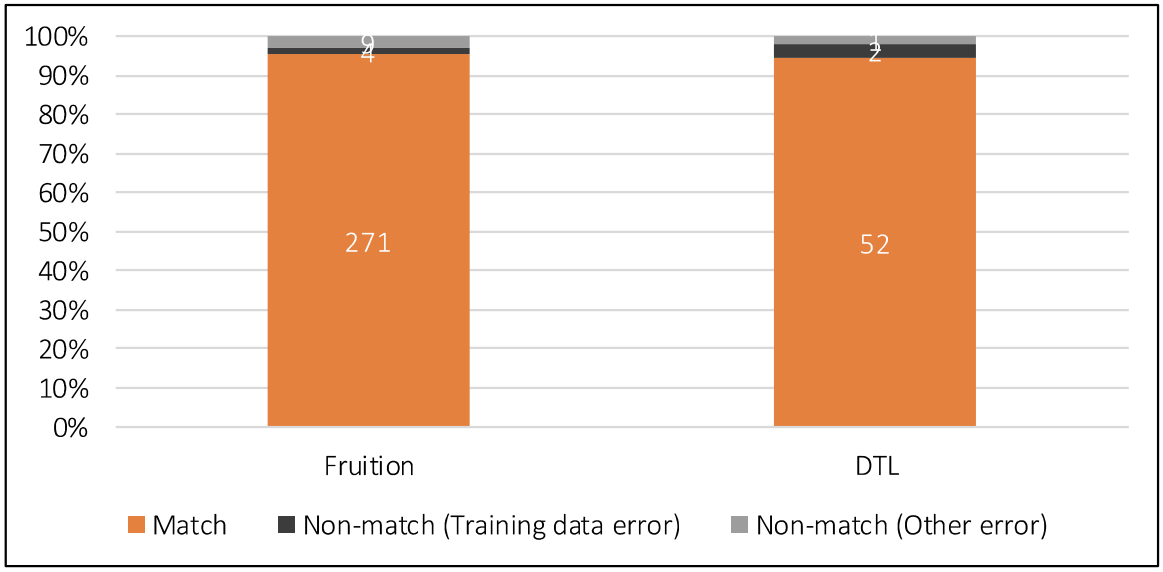


Figure 3: Matrix of AI agent grade to tutor grade

		AI agent grade	
		Competent	Not competent
Tutor grade	Competent	180	15
	Not competent	1	106

Fruition Horticulture Ltd

Fruition's tutors agreed with 271/284 responses (95%). Of the 13 AI grades that tutors disagreed with, four were due to an error in the prompting materials. Specifically, an exemplar that was used to train the AI agent stated that jewellery must be removed before entering a packhouse, *but that wedding rings were an exception*. As a result, the AI marked learners' answers as wrong if they did not mention this exception. As noted previously, this could be prevented in future by editing the exemplar.

In another instance, we noted that a learner who had limited vocabulary was marked incorrect on several answers where the tutor would have marked them as correct. This may be indicative of a bias against learners who give short answers, and we suggest should be investigated further in future pilots, with mitigations built into the agent to prevent this occurring if real.

Dairy Training Ltd

In the case of DTL, tutors agreed with 52/55 AI grades (95%). Of the three AI grades that tutors disagreed with, two were caused by an error in the prompting materials. This was because the rubric implied that learners had to reference a specific contract clause in an answer, when they did not. Again, we suggest this could be prevented in future by editing the rubric. In other instances, we noticed tutors graded the same answer differently.

Some tutors at DTL were initially unsure about trialling the AI agent – although this may have been partially because it was a proof-of-concept tool, rather than because it was AI. One learner asked their tutor about how the AI agent would use their data, and was unsure about using AI for their assessment. This reflected the importance of giving tutors an initial overview of how the AI agent works and what ethical considerations have been made in its development; giving tutors a range of options for how to pilot the agent to ensure they are comfortable with it; preparing them to be able to answer learners questions; and providing learners with non-AI alternatives for their assessment (see our recommended steps from page 21).

> *"They came to me and asked about how it was going to use their data. I wasn't sure what to say. I think it's important at least for now, for us to provide them with an alternative" (DTL tutor)*

Ease of use and enjoyment

The majority of learners (20/24) in the pilots found the oral assessment agent either "easy" or "very easy" to use. A reasonable number (15/24) also found the agent "enjoyable" or "very enjoyable" to use. We judge this to be a good early result, given that an assessment is not typically expected to be enjoyable. Where learners provided lower scores, this was often due to technical challenges (which could be addressed in a production version) or learner preference for writing. This suggests that having an option for writing the assessment is still important.

Figure 14: Learners' views on if the agent was enjoyable

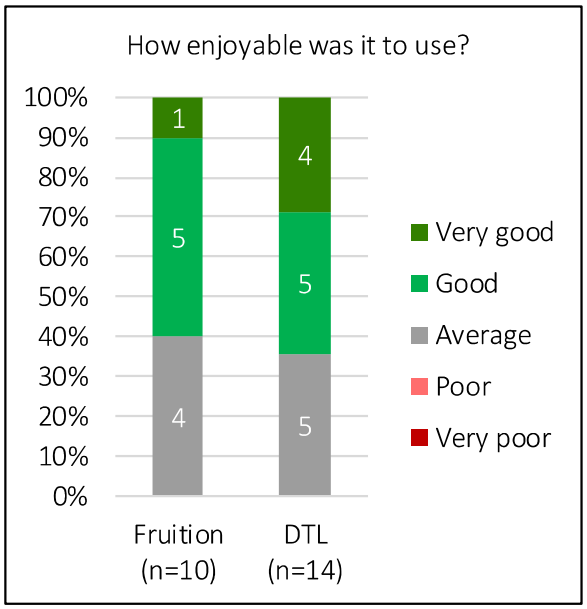
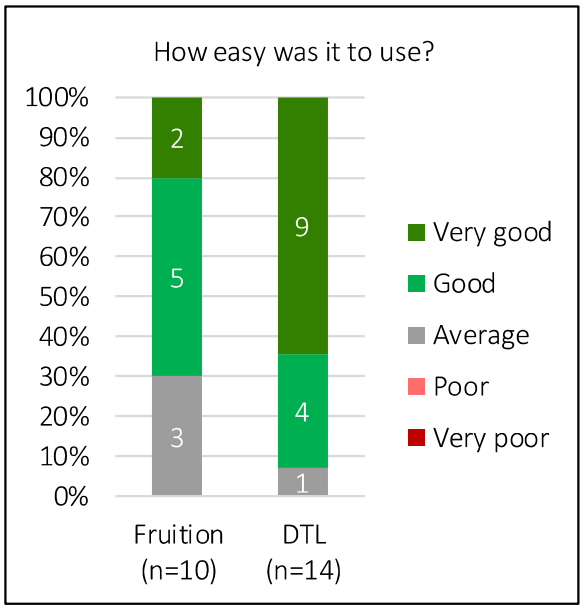


Figure 5: Learners' views on the agent's ease of use



Fruition Horticulture Ltd

Over half of Fruition's learners found the agent enjoyable or very enjoyable (6/10). One learner explained that the main benefit was being able to communicate their ideas better than they could with writing. We suggest this is likely the case for many vocational students with similar challenges.

> *"The agent is good if you don't know how to write down the ideas in your head" (Fruition learner)*

On the other hand, staff saw a wide range of benefits – some of which were unexpected. For example, they noted that learners often did not have the confidence to ask questions, but that the agent appeared to provide a safe space where they would not be judged. They believed this would make learners more confident to ask questions in future.

> *"They found the agent easy to interact with. They were able to say 'Can you repeat that question? I don't understand what you mean.' So that conversational interaction kind of made it more like… just a conversation and not an actual assessment" (Fruition staff)*

> *"When they read [in a written assessment]… they may not necessarily understand what's been asked to them [so with the AI and being able to ask questions, it is different]" (Fruition staff)*

> *"A few of them [said] to me, I don't understand it and I said just ask it to repeat the question in a different way… and it did and you just watched them go oh, yeah now I get it… I think that's why they got that sort of like confidence boost at the end" (Fruition staff)*

Tutors also added that some learners who initially spoke unclearly to the agent, had naturally started to articulate better during the conversation. They thought this would benefit learners in the long-term.

> *"When they figured out that the agent wasn't picking up the speech, they had to start to articulate their words, which to me is a really good thing." (Fruition tutor)*

> *"The main thing is, you know, because we teach communication or you know how important it is to speak properly so we can understand and so that agent kind of makes them without [prompting]" (Fruition staff)*

In terms of ease of use, 7/10 Fruition learners said it was easy or very easy.

> *"It was so easy that it repeats and simplifies, I was able to do the assessment in 10 minutes." (Fruition learner)*

> *"[The responses from the agent] were both appropriate for the question and very helpful" (Fruition learner)*

Where there were challenges, it tended to be due to technical challenges (which could likely be fixed in a production version); due to certain features we had piloted (specifically, having a follow-up question at the end of the conversation rather than immediately after a learner's answer); as well as learners using the agent for the first time. However, some learners said they would prefer writing assessment.

> *"Devices probably weren't the best to use, and our Wi-Fi was playing up at the time, so that… became an issue which then became a frustration" (Fruition staff)*

> *"It was okay but it had me repeating a couple questions that I thought that I had answered pretty well or to the standard it was setting." (Fruition learner)*

> *"Using voice controls [is] harder than writing the answers". (Fruition learner)*

> *"It was a new thing for us, and it was a new thing for them. And a lot were uncomfortable or found it different just to speak to a tablet or the phone… I think time will definitely change" (Fruition staff)*

> *"[I would prefer to write my answers] Just a personal preference." (Fruition learner)*

Dairy Training Ltd

Of the 14 learners in the DTL pilot, 9 found it enjoyable or very enjoyable. Learners appreciated the ability to do the assessment when it suited them, and get direct, immediate feedback.

> *"[The] easy part is you can do [the assessment] anytime you want" (DTL learner)*

> *"The agent gives me direct feedback for my answers and helps me understand… the level of my answer and where I need to improve" (DTL learner)*

> *"The feedback was instant which was great" (DTL learner)*

Some, however, said they would have preferred a tutor-led conversation if possible. It is worth noting that currently, the expense of tutor-led conversations for assessment makes this infeasible for most providers – the options therefore being writing or (now) an AI agent.

> *"We need to talk to people that can share their experience rather than talking to a computer" (DTL learner)*

> *"It was an odd experience at first" (DTL learner)*

Almost all DTL learners found the agent easy or very easy to use (13/14). We suggest the DTL agent may have been easier to use than the Fruition agent because of the number of questions and the use

of follow-up questions immediately after a learners answer rather than at the end of the conversation. Exploring different question layouts is therefore likely a good use of time before or during pilots.

> *"Pretty straight forward" (DTL learner)*

Tutors explained that this was because the agent understood them well. The operations manager noted that this could make assessments much easier for those with learning difficulties.

> *"It is very clear and understands you well" (DTL tutor)*

> *"It could be useful for people with learning difficulties" (DTL operations manager)*

Impact on time spent

Tutors in both pilots suggested that once refined and integrated into their learning management system, the agent could provide significant time savings. We provide self-reported estimates from Dairy Training Ltd and Fruition Horticulture Ltd below. These estimates such an agent could save both learners and tutors time, especially if transitioning from a *written assessment* to an oral AI agent assessment (as in the case of Fruition), versus from a *tutor-run oral* assessment to AI oral agent assessment (as in the case of Dairy Training).

Fruition Horticulture Ltd

Figure 6: Estimated time savings for Fruition Horticulture Ltd

	Traditional written assessment	AI oral agent assessment
Assessment details		
Number of learners in class	14 learners	
Time spent per assessment		
Learner time per assessment	2.5 hours	15 minutes
Total learner time spent on assessments	35 hours	3.5 hours
Tutor time per assessment, per learner	1.5 hours (to administer, grade and upload feedback)	15 minutes
Total tutor time spent on assessments	21 hours	3.5 hours
Total estimated time saved if using the AI agent	Learners (together) save ~31.5 hours. Each tutor saves ~17.5 hours.	

Dairy Training Ltd

Figure 7: Estimated time savings for Dairy Training Ltd

	Tutor run oral assessment	AI oral agent assessment
Assessment details		

Number of learners in class	11 learners	
Time spent per assessment		
Learner time per assessment	15 minutes	15 minutes
Total learner time spent on assessments	3 hours	3 hours
Tutor time per assessment, per learner	40 minutes[1]	10 minutes[2]
Total tutor time spent on assessments	7 hours	2 hours
Total time saved if using the AI agent	Learners (together) save ~0 hours. Each tutor saves ~5 hours.	

We include quotes from staff and learners here to show the impact that saving time can have.

> *"I think that the tutor's spent quite an incredible amount of time supporting the learners to engage with the paper tasks… a lot of one-on-one and there is only one tutor with those 12 [students] so… they're having to make decisions all the time is to prioritise who their attention goes to and for how long they can do and often learners miss out. [It would be different with an AI agent available to help]" (Fruition staff)*

> *"I think that AI will have a massive impact on that kind of equitable resource or support to engage with assessment… The bureaucracy around assessment demands and reporting and documentation is massive, and I can see AI cutting down on that hugely [allowing us to spend more time supporting the learners instead]" (Fruition staff)*

Self-worth and pride

While no survey question was asked of learners on self-worth and pride, staff from both organisations were able to observe students using the agent and note whether they saw any change. The agent appeared to have particular impact on the self-worth and pride of Fruition Horticulture Ltd learners – possibly because they are younger and less confident in their ability to succeed in formal education (see page 7 for more comparisons of the learners of these two pilots).

Fruition Horticulture Ltd

Staff noted that learners had an increased sense of self-worth and pride because they could ask questions, better express their ideas, and complete the assessment quickly. This is best seen in the quotes from staff:

> *"I'm feeling pretty positive because it… broadens the inclusivity for students… it allows them to have another outlet to answer assessments" (Fruition staff)*

> *"I was just sitting off in the corner… so very much from an observer point of view, I got to watch the sense of empowerment when they were getting to use it and watch them navigate to see what they would say, getting frustrated, but what they also found easy." (Fruition academic manager)*

[1] Including time to book the meeting, administer the assessment, mark the assessment and upload grades/feedback.
[2] Including time to check the results within the LMS.

"I found some of the students seemed upbeat after they'd done it… They were like, oh, this was really easy... It gave them confidence to not be afraid" (Fruition staff)

"Most of our students have been let down by the education system and refuse to ask questions… But AI isn't going to judge them. It empowers them". (Fruition tutor)

"At one stage there was this pair of young ladies… working togeth[er] and they figured out [as] one talked and the other didn't talk, but they practise their answers... which is what they should be doing and they sort of made sure that they felt quite clear." (Fruition academic manager)

Interest going forward

Learners in these small pilots demonstrated an interest in seeing the AI agent in more assessments. Most (21/24) either say they would like to see it in future assessments or that they *may* like to with improvements. However, we also found several learners noted a preference for written assessments over oral assessments (7/16). While some of these learners may be convinced otherwise once technical challenges are addressed, we suggest that there will likely always be learners who prefer writing. This should be carefully considered when rolling out an AI agent for assessment.

Figure 8: Learners' views on the future of AI assessment agents

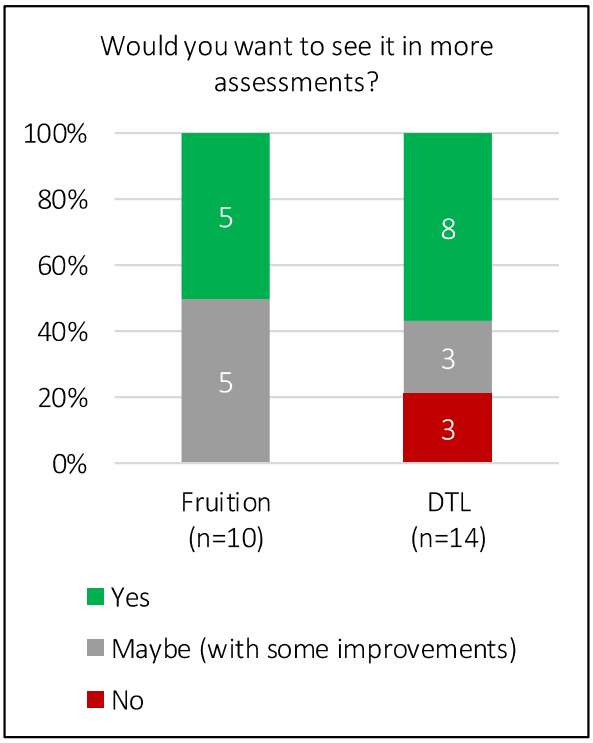
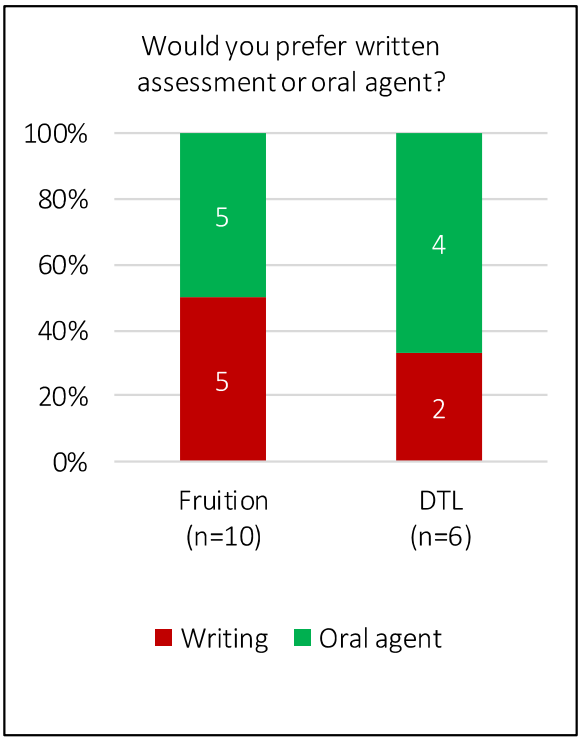


Figure 9: Learners' preference on assessment format



As mentioned, both organisations suggested they would like to continue using the AI agent for assessment. We include quotes of this here:

"I would like to see it across all of our assessments this year" (DTL operations manager)

"If the backend [could be developed to be] easy, so I didn't utilise a lot of time, I [would be] 100% for AI and assessment / oral assessment, absolutely. That can be an option for future assessments or all our assessments that it will support" (Fruition tutor)

"I've been thinking lots and lots about [how to incorporate it more]... If you could combine [AI and something that collects photos or other evidence out on the field]... so that [the tutor] is out there on the orchard and he just needs to speak to the AI and the AI prompts them to... whack this picture in and now whack this in here and did all of that, I reckon it would be amazing... That would be a game changer" (Fruition academic manager)

Ideas for AI agent for assessment improvement

However, they also could see clear areas of improvement for the AI agent for assessment. We list some of these here:

- **Continue to build a strong database of prompting materials** – Presumably, with more prompting materials (e.g., exemplar answers, lists of common terms), the AI agent could better understand the learner and/or grade more accurately.
- **Consider asking the learner follow-ups immediately after a question** – Organisations noted that asking follow-ups immediately after an incorrect or vague learner answer (e.g., "Could you explain that a bit more?") would be better than asking at the end of the conversation. There could still be one question at the end that asks if the learner has anything further to add.
- **Enable the agent to conduct the assessment in Te Reo** – Early explorations suggest that the agent could be developed to conduct assessments in Te Reo, although further testing would be required.
- **Build the ability for learners to type as well as speak** – As seen in these pilots, there continue to be learners who would prefer to write. Providing both options within the same agent would improve accessibility while maintaining a streamlined assessment process.
- **Explore how AI responds to speech impediments or limited vocabulary** – As seen in these pilots, learners with unclear speech or limited vocabulary may struggle more with the AI agent. It would be worth testing this further and identifying any required mitigations.
- **Build in the ability to upload other forms of evidence** – There was interest in having one product or app where learners could collect all evidence (e.g., oral assessment, log of hours of experience, photo and video uploads etc).
- **Develop a tutor dashboard with engagement and performance analytics** – Tutors would like access to which learners are using the agent and what topics they most often ask about. This would enable them to adapt their course in almost real-time.

Ideas for other AI agents in vocational education

We note ideas here of other agents (i.e., not assessment agents) brainstormed by Fruition Horticulture Ltd, Dairy Training Ltd and other stakeholders involved in this work (see ideas of other types of *assessment* agents on page 3).

1. **Learner admin agent** – This agent would be used by potential learner to ask about course enrolment, course content, logistics, what to bring etc. It could be available on the organisations website to anyone interested in courses. *See Deakin Genie for a similar example.*

2. **Learning assistance agent** – This agent would be used by enrolled learners during or after the course to ask about the course content. For example, to ask for the definition of terms, to review their outputs or practise scenarios. *See Cogniti for similar examples.*
3. **Tutor admin agent** – This agent would be used by tutors to help with administrative tasks (help log class attendance, manage requests from learners for extensions for work, send reminders to learners). *See QuadC for a similar example.*
4. **Career pathways agent** –This agent would be used by learners to understand their career options, what might be right for them, and where to go for more information or who to speak with. *See Coach for a similar example.*

How can we develop our own?

This section gives you a step-by-step guide to develop your own AI agent for oral assessment. We include lessons learnt from our own pilots that we recommend to others.

1. Scope

Overview

Within this phase, providers will define the problem they are wanting to solve and consider the various solutions available to them. The provider should consider to what extent AI as a solution is a strong contender, *and* whether they should build it themselves or not.

Steps

1. **Define your assessment challenge** – An AI agent for oral assessment is likely most suitable if current written assessments are considered non-inclusive, and/or time consuming, and human-run oral assessments are not practical. This is more likely with users who struggle with writing, who are located remotely, or in the workplace.
2. **Identify what solutions may be available to you** – AI may be one solution, but you could find that other, simpler solutions meet your needs well. For example, audio recording answers and AI-produced transcriptions of those answers would enable learners to avoid writing (although this would not have the conversational ability of an AI agent, i.e., the ability to clarify things for the learner, to ask them follow-up questions, to encourage them and provide feedback).
3. **Decide whether to build your own AI agent** – We suggest it is highly worth looking for an existing AI product to solve your assessment challenge. At the time of writing, there were none that performed oral assessment, but this could change. Otherwise, you could outsource to a firm who can develop one or develop one internally. To develop one internally, consider your capacity, capability and costs of ownership.
4. **Form your team and create a workplan** – If building your own agent, you will now need to calculate a budget for initial development. We also suggest estimating ongoing costs. A timeline for the work should also be developed, identifying who will be involved.

Lessons learnt

- **Be ready for something experimental** – As at the time of writing, OpenAI's real time speech-to-speech model is the best at natural conversation, but it is still in beta (and therefore has bugs) and is relatively expensive. Alternative approaches to creating voice agents are advancing in quality at a lower price. At least at the time of writing, we suggest you should expect bugs and unusual behaviour at times and be open to switching technology and approach as progress is made on the underlying technologies.
- **Consider what moderation will allow** – Consider how moderation may impact the design of the assessment. For example, if it prevents the agent from asking the learner follow-up questions

during the conversation or from providing feedback to the learner at the end of the conversation, then this may make the product less viable.

- **Check whether integration will be possible** – Learning management systems will not be able to automatically integrate with new AI products they have not integrated with before. It will be important to assess at this point how challenging integration will be through a conversation with your provider. If there are few ways forward, you may decide here to stop the plan to create an AI agent.
- **Bring staff onboard early** – You may want to bring tutors, academic staff and QA leads into the project early. Staff who are on the ground with learners every day (i.e., tutors) are especially important to bring in early, as they can help to ensure the eventual product meets real needs. Moreover, they will be critical to achieving learner uptake, because it will be up to them to encourage their learners to use the agent.

Resources

- **Lists of existing AI agents** – See our articles on agents in Oceania and abroad.[3] Alternatively, ask ChatGPT for a list of relevant AI agents.

[3] https://scarlatti.co.nz/case-studies/shaping-the-future-ai-education-projects-in-oceania/; https://scarlatti.co.nz/case-studies/the-global-landscape-how-ai-is-transforming-education/

2. Design

Overview

Within this phase, providers (or their technical sub-contractor) have already decided to develop an AI agent. This phase is therefore about redesigning the assessment into an ideal form, then defining requirements for the AI agent based on this and creating a high-level design for the AI agent.

Steps

1. **Consider what the 'ideal assessment' would look like** – We suggest it may be best to start with what the 'ideal assessment' would look like. This may involve reimagining your assessment from the ground up, rather than holding onto the existing assessment format. Once reimagined, carefully balance the investment of developing this with the expected return (see more on this under Lessons learnt below).
2. **Define the requirements for your agent based on your reimagined assessment –** Consider what the AI agent will need in order to assess learners, what its interface should include, and how you can securely store responses (for more ideas see page 44).
3. **Develop a 'high-level design document' to capture the above** – This can be used as a communication tool internally, to ensure that needs are meet, that the plans are realistic and that the team are on the same page going forward.

Lessons learnt

- **Reimagine your assessment from the ground up** - Your assessment may already seem suitable for an AI oral agent. However, rather than simply 'reformatting' your existing content e.g., turning a written quiz into a voice-based quiz, we suggest reimagining your assessment from the ground up to foster a deeper learning experience. AI agents open opportunities for conversational probing, adaptive follow-ups, realistic role-plays and instant feedback (see page 29 for more on this).
- **Undertake a review of ethical guidance given continual developments** – As of April 2025, New Zealand government guidance on AI use could mainly be found in advice from the Office of the Privacy Commissioner; from the joint work by the Department of Internal Affairs, the National Cyber Security Centre and Statistics New Zealand (2023a); and from the Ministry of Education (2023). We wrote an article on this topic. We suggest that due to the limited nature of this guidance and due to the speed of developments in this space, conducting your own review is important. You may also want to look overseas for additional advice. For an example review see page 32.
- **Consider splitting the AI role into an 'examiner' and an 'assessor'** – Consider splitting the final product into an "examiner" model that converses with the learner and an "assessor" model that grades behind the scenes and creates structured data outputs (e.g., a grade plus text feedback). There are two benefits to this. First, it protects integrity, as the examiner agent speaking with the learner cannot access course materials and therefore cannot 'give away' answers to the learners. Second, it means a fast voice model can be used to converse with the

learner, while a 'smarter' but slower text model can be used to assess the answers for more accurate assessment. The negative is that it limits the examiner agent's ability to ask the learner follow-up questions if their answers are incorrect, vague or unclear. If splitting, then:

  - **Assess the voice model options available for the 'examiner' agent** – As of the time of writing, there were two approaches to creating a voice-based agent: a speech-to-speech approach (responds faster making conversation feel more natural) and a chained speech-to-text → text LLM → text-to-speech pipeline (responds slower but allows you to swap in different accents, ages, and genders). Either way, we recommend scanning for the best models and undertaking rigorous tests. This is because at the time of writing, some still have bugs (e.g., New Zealand accents still sound inconsistent and may 'drift' into other accents during the conversation), but, quality is improving rapidly. Finally, match the model to your team's technical strength: cutting-edge options such as OpenAI's real-time voice (beta) deliver near-instant responses but are trickier to integrate and debug, whereas tools like Vapi (which implements the chained approach) can reduce complexity or help you pivot later if requirements change.

  - **Assess the model options available for the 'assessor' agent** – As of the time of writing, the best models come from OpenAI, Anthropic and Google. As before, we recommend exploring these. This is for a few reasons. First, different models may make different assessment decisions about the same learner's answer. Second, you need a model capable of providing 'structured outputs' (this means that you can force the model to provide a grade (from a list) and feedback for each question). Whichever provider you choose, we suggest 'regular' models appear capable of most basic assessments. If your assessment is more complex, you may need to consider a 'reasoning' model instead.

- **Allow agent to search course materials/rubric rather than including it in prompt** - Course materials are needed to give the agent (particularly the "assessor" agent if decoupled) enough context to appropriately grade and respond to the learner's answers. However, due to course materials being long, including it all in the prompt is likely to overwhelm the agent. Using 'retrieval augmented generation' (RAG) allows the agent to dynamically retrieve relevant information as needed.

- **Find a balance between consistency and flexibility** - Feeding the AI with content that is highly specific may make it easier to consistently grade students, but it may also miss the point of the assessment or even be biased against the student. For example, if the rubric implies the learner needs to mention a key word when in fact they do not, or in an exemplar answer a student mentions additional information that was in fact not needed to receive a passing grade, the AI agent could require this of the learner to pass. This is true with traditional assessments, but also with an AI agent.

Resources

- **ISAR model for framing AI's role** – See page 29.
- **Example ethical considerations table** – See page 31.
- **Requirements planning table** – See page 44.

3. Develop

Overview

Within this phase, providers (or their technical sub-contractor) will develop an early, functional version of the agent.

Steps

1. **Roll out the design document** – This will involve writing the code and setting up the interface, AI models, prompt structures, agent roles, and secure storage.
2. **Provide the agent with relevant content** – Identify and collect relevant material for this. For example, an assessment rubric that is currently used by tutors to assess learners, course materials used to teach learners, exemplar answers or a list of common terms.
3. **Include materials to build user confidence in the agent** – Some users will be unsure about using a new technology like an AI agent. Consider how you may support these users. For example, an instructional video on how to use the agent, an information sheet on how their data is used (see page 45 for an example of this), or practice questions for them to try out the technology may help.
4. **Conduct multiple rounds of testing** – For example, does it stay on topic / stick to the questions, is it stable, does it give consistent feedback and is the grading accurate.

Note: When building a production version, a provider would need to undertake activities that were not done in our proof-of-concept pilot. For example, integrating the agent into an LMS and developing anti-cheating functionality. Given these were not a part of our pilots, we do not advise on them here – but that does not mean they are not necessary.

Lessons learnt

- **Craft clear, focused prompts for the examiner agent** – We suggest that prompts should be simple and streamlined. They need to: define how the examiner agent should probe, reveal information and response; embed the preamble as one continuous line to be read verbatim; specify the exact scope and detail of the feedback; instruct on conversational tone for chained models, all while keeping instructions brief enough that none are forgotten. Current models can struggle with subtlety. For example, we tried giving the examiner access to the grading rubric, asking it to use the rubric to guide follow-up questions, while not revealing any information that would give the learner an advantage. This led to inconsistent behaviour and the examiner would often reveal information from the rubric in its follow up questions.
- **Build in mitigations to prevent cheating and misuse** – One way to do this is to list out every way you think a user could try cheating or otherwise misuse the agent, then to develop a mitigation for each. Sending the agent to your team and encouraging them to try 'misusing' the agent can help you develop this initial list. As an additional failsafe, we also recommend pre-warning students that misuse may result in an automatic fail grade and using AI post-assessment to

review whether students broke any rules. See our table of behaviours and mitigations on page 45.

- **Explore how best to enable follow-up questions** – As seen earlier, we suggest decoupling the agent into two (an examiner and an assessor). However, you will need to trial what exact format enables follow-up questions best. If the examiner has access to the grading rubric, it can more easily ask relevant follow-ups, but it may reveal too much information to the learner. If it does not have access to the rubric, you may need to provide quite explicit instructions for each question about when and how to follow up. Another possibility is to have the examiner call an 'intermediate assessor' agent after each response and have that agent tell the examiner whether and how to follow up.
- **Be prepared to adjust the content used to prompt the agent** – By now, you will have provided the agent with content to grade the learner's answers. For example, a rubric, course materials or exemplar answers. Our pilots showed that most times that the tutor disagreed with the AI's provisional grade, it was *not* because the AI agent made a mistake, but because the content used contained an error. You may therefore need to return to your content post-testing to make small edits to improve the AI grading.

Resources

- **Unexpected learner behaviours and mitigations** – See page 45.
- **Code and instructions –** See page 44.
- **Example information sheet –** See page 45.

4. Pilot

Overview

Within this phase, providers will plan and undertake the pilot. This may involve applying for ethics approval, confirming the cohort to participate, and briefing tutors before conducting the pilot.

Steps

1. **Ensure any required ethics applications are made –** Your organisation may require you to apply to an ethics committee to conduct a pilot. It is important to receive this approval prior to beginning your pilot.
2. **Confirm which cohort will participate in the pilot** – Identify which cohort will be most suitable for your pilot – this may depend on who leads it, the timing or the students enrolled.
3. **Brief tutors on the pilots** – Pre-send a link of the agent (including information on how users' data is used) to tutors in advance and ask them to both trial the agent and read the accompanying material. Use the meeting to check all tutors have done so; explain how the pilot will work, inform tutors about who to go to for support; and answer any other questions.
4. **Conduct pilots of the AI agent** – It is now time to run the pilot. We provide several options for this under Lessons learnt below.

Lessons learnt

- **Offer tutors different pilot options depending on their comfort level** – Tutors vary in their comfort using new technology. Consider giving tutors options for how they run the pilot, ranging from demonstrating the agent to learners; having students use the agent for practice; or having students use the agent for their actual assessment, either with or without supervision. Tutors should then collect feedback from students (no matter how they decided to pilot).
- **Teach tutors about the ethical considerations of an AI agent** – It may be worthwhile to have discussions with tutors to make sure that they understand the ethical considerations of learners using the agent and are able to answer any questions learners may have in simple terms. This would likely be in addition to providing an information sheet and links in the preamble to the AI model's data and privacy policies. For an example information sheet, see page 45.
- **Test the agent in safe and low-stakes environments** – In case technical issues arise or the agent is not naturally intuitive to the user, test it in low-stakes environments. For example, on an assessment that is worth a small percentage of a final grade (i.e. 1%). Alternatively, you could demonstrate the agent to learners or have them use it for practice for an assessment, rather than for the assessment itself.

Resources

- **Example ethical considerations table** – See page 32.
- **Example information sheet** – See page 45.

5. Evaluate

Overview

Within this phase, providers will develop an evaluation plan and collect evidence against this plan.

Steps

1. **Develop a structured evaluation plan** – Start by developing a logic model (see page 30), from which you can identify indicators and measures (see page 41).
2. **Create evaluation tools** – This could include surveys, case studies, statistics from the AI agent (e.g., usage, time spent, grading), and more. It should also include feedback from both learners and staff.
3. **Collect feedback** – Using the evaluation tools you developed earlier, gather feedback. If you want to check how well the agent is grading, have tutors complete a blind copy of the grading (i.e., without providing them with the AI grade) and from this, calculate the grading match between AI agent and tutor.
4. **Review results collaboratively and identify next steps –** Debriefing the results with your staff is a good way to interpret them correctly, build buy in and collectively decide on next steps.
5. **Identify opportunities beyond this agent** – We suggest that as the evaluation wraps up, it is a good moment for staff to consider how AI could be used more broadly to support the organisation or learners.

Lessons learnt

- **Consider whether you can rely on studies by others** – In some cases, similar functionalities have been tested by others. For example, the Construction and Infrastructure Centre of Vocational Excellence is testing how well AI can create customised assessment questions based on input, and NZQA has recently used AI for Year 10 Literacy and Numeracy assessments and tested for things such as accuracy and bias.
- **Give opportunities for both written and oral feedback** – Depending on the person, they may prefer to give written or oral feedback (e.g., surveys, emails, interviews, group workshops), providing opportunities for both is likely to result in a wider range of feedback received. Within our pilots, tutors noted any observations about learners during the assessment; learners and tutors could both send feedback by survey; tutors collected qualitative feedback from their learners either individually or as a group if they wished; then tutors had a group debrief to review evaluation results, think about next steps and the possible future of oral AI in their work.
- **Distinguish between feedback on concept and existing product** – Given the product will be early stages, much of the feedback will likely be on technical issues rather than the concept itself. Given this, you may want to take feedback with a grain of salt.

Resources

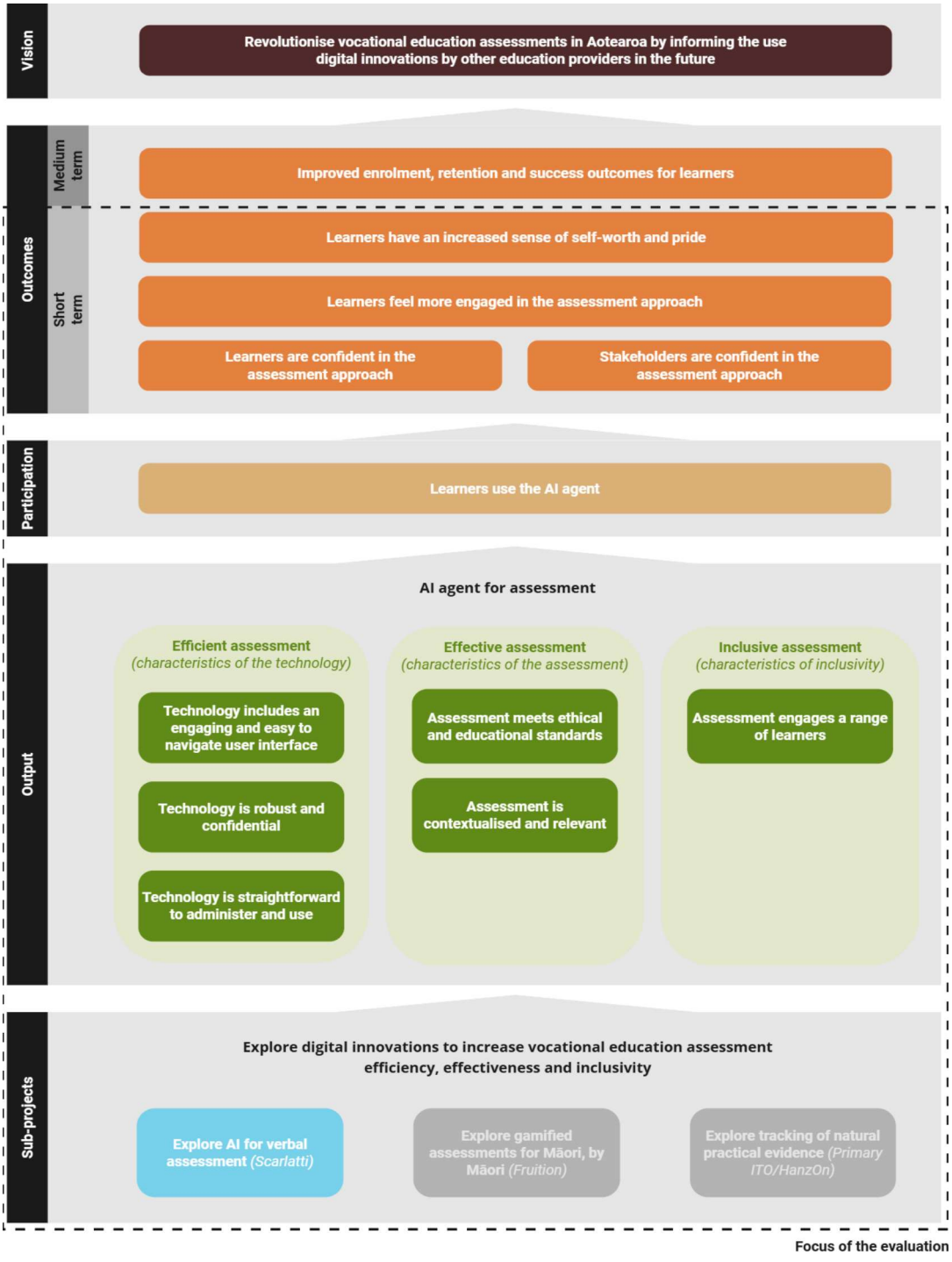
- **Example project logic model** – See page 30.
- **Example research questions** – See page 8.
- **Example measures table** – See page 41.

Resources

Project logic model

The logic model below was developed to illustrate how the project's desired outputs would be used by participants, and how this would result in outcomes and long-term vision. You may wish to use a similar logic model when developing your own AI agent for oral assessment.

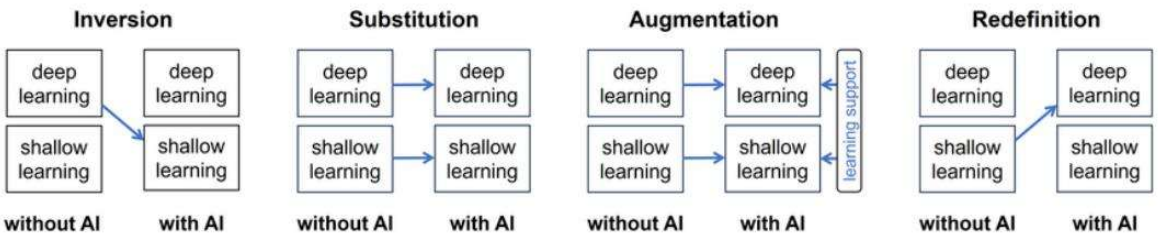
Figure 10: Logic model for the AI agent for oral assessment project



ISAR model

Below we provide an example framework that may help you think through your approach to AI.

Figure 11: The ISAR model, from Bauer, Greiff, Graesser and Scheiter



The ISAR model is a framework that classifies AI's impact as:

- Inversion – AI use reduces meaningful cognitive engagement
- Substitution –AI replicates traditional methods, with no added learning gain
- Augmentation – AI use enhances learning by providing targeted support
- Redefinition – AI transforms tasks to foster deeper thinking

Ethical considerations table example

Below we provide an example of an ethical considerations table that summarise what key concerns are raised by the most recent AI ethical guidelines and how they may be relevant to our assessment agent. You may wish to use a similar table when developing your own AI agent for oral assessment.

Table 1: Ethical considerations table example

Ethical considerations	Key source quotes on concern this raises	Is this in the scope of the pilot?	Is this in the scope of future projects?	Severity of risk	Possible mitigations				
					Phase 1: Planning	Phase 2: Design	Phase 3: Develop	Phase 4: Pilot	Phase 5: Evaluate
Privacy and Māori data sovereignty Key sources: Artificial Intelligence (AI) and the Information Privacy Principles (Office of the Privacy Commissioner, 2023); Generative AI: Guidance and resources for educational professionals (Ministry of	Avoid inputting personal data into generative AI tools.	Yes - We will be collecting learners' names and student IDs at the end of the assessment, but this information will be kept separate from the AI agent.	Yes - This concern is important to consider when choosing an AI model and if any personal information will be collected. If this information is collected, will it be used to train the model?	Low	Collect learners' names and student IDs at the end of the assessment. In doing so, this information will be kept separate from the AI agent and will not be shared with the AI model provider (OpenAI) or used as	Consider adding information in the preamble advising users against inputting their personal information during verbal assessments due to privacy concerns.	Develop the aforementioned preamble.	Test that the preamble is working as intended and that users cannot skip viewing it.	N/A.

Education, 2024) and the Initial advice on Generative AI in the public service (Department of Internal Affairs; National Cyber Security Centre & Stats NZ, 2023).					training data.				
	Avoid inputting information that would be withheld under the OIA.	N/A - The Official Information Act is only applicable to government agencies and is not in scope of Scarlatti's future projects.							
	Avoid using GenAI for business-critical information, systems or public-facing channels.	N/A - This concern is about broadly using AI in organisations, where we are solely concerned with using AI through the agent we are developing to conduct verbal assessments. Consequently, it is also not in the scope of future projects.							
	Don't use GEN AI tools for data classified at SENSITIVE or above.	N/A - This concern is only applicable to government agencies and is not in scope of future projects.							
	Don't input personal information to GenAI tools if they are external to your environment.	N/A - We are not inputting any personal information from other departments into the AI agent. It is unlikely to be in the scope of future projects.							

Fairness and non-discrimination Key sources: AI and the Information Privacy Principles (Office of the Privacy Commissioner, 2023).	Is the training data behind an AI tool relevant, reliable, and ethical?	N/A - We have no way of assessing what data was originally used to train OpenAI's voice model which we are using for our pilot. We are unlikely to be able to assess this in future work.							
	How are you testing that the AI tools are accurate and fair for your intended purpose?	Yes - We are wanting to design an AI agent that conducts verbal assessments in an accurate and fair manner.	Yes - We want future AI projects to work as intended and be improved on.	Low	Work out how tutors can easily check the AI's agent's outputs (their preliminary grading).	Compare the model's outputs/preliminary grading against the same assessment marked by a human tutor and get them to provide any criticisms.	Build in mitigations for unusual student behaviour. For example, you could restrict the student from speaking while the agent is answering, in case the student interrupts the AI agent and results in the agent giving broken transcripts.	Undertake testing of the pilot with a large and diverse group of people to ensure that the agent works as intended.	Evaluate whether the pilot has worked as intended and in a fair manner using our evaluation plan.

Accountability and oversight Key sources: AI and the Information Privacy Principles (Office of the Privacy Commissioner, 2023); Generative AI: Guidance and resources for educational professionals (Ministry of Education, 2024) and the Initial advice on Generative AI in the public service (Department of Internal Affairs; National Cyber Security Centre &	Review whether a generative AI is necessary and proportionate given potential privacy impacts and consider whether you could take a different approach.	Yes - We only want to develop an AI agent for assessment if it would help to solve issues for learners, education providers and employers.	Yes - We do not want to develop AI agents if they are not solving a problem for the group commissio-ng the work.	Low	Identify what problems an AI assessment agent could solve for learners in vocational education and training and assess the possible privacy impacts of implementi-ng this tool.	N/A	N/A	N/A	N/A
	Have senior leadership approval based on a full consideration of risks and mitigations.	N/A - The adoption of AI agents is up to individual education providers, so it is not in the scope of the current or future projects.							
	Let users test AI tools safely.	N/A - This concern was written in the context of trialling AI in your organisation, so it is not relevant to the current project to develop an AI assessment pilot, but it is in the scope of future projects.							
	Exercise caution when	N/A - We have chosen to use OpenAI's paid subscription which is only available to specific people. It may be in the scope of future projects if we choose to use a different model.							

Stats NZ, 2023).	using publicly available AI.								
	Look over the AI model's terms and conditions.	Yes - We are interested in what we are agreeing to by using OpenAI's model to develop an assessment agent and how it uses the inputted data.	Yes - Future projects should look into AI models' terms and conditions, consider how they use inputted data and how this could affect users' privacy.	Low	Look over the terms and conditions to check about privacy settings.	Add a link to OpenAI's terms and conditions in the preamble that users can go to if they are interested.	Develop the aforementione -d preamble.	Test that this preamble is working as intended and cannot be bypassed.	N/A
	Apply the Government's procurement principles.	N/A - We are not a government agency and is this out of the scope of future projects.							
	Prevent AI from being used as a shadow IT.	N/A - Our agent is not designed to be used in this way (as a system). Any future projects that develop a AI system may need to consider this concern.							
	Assess and manage for privacy risks by conducting a privacy	N/A - We are not collecting personal information using the AI agent so will not need to undertake a privacy impact assessment. However, conducting a privacy impact assessment may be required for future AI projects that collect personal information.							

	impact assessment.								
	Ensure human review prior to acting on AI outputs to reduce risks of inaccuracy and bias.	Yes - Our agent is designed to provide tutors with preliminary assessment grades for them to review and check for accuracy before they give the final grade to their learners. It is not intended to replace traditional tutors/teachers.	Yes - AI outputs should always be reviewed for accuracy before using their generated content.	Medium	Enable the agent to take recordings and generate transcripts of the assessments, so that tutors can look over it and ensure that its outputs are accurate.	Build into the agent the ability to record and make transcripts of verbal assessments.	Hardcode system-level instructions to protect them against the hacking of prompts and manage them server-side.	Test the accuracy of the agent's outputs against multiple tutors throughout the pilot.	Measure the accuracy of the agent's outputs against the grade human tutors would award the learner and any criticisms raised.
Cultural sensitivity and safety Key sources: AI and the Information Privacy	Engage with Māori about potential risks and impacts to the taonga of their information.	N/A - Due to it being the piloting stage, the project's quick turnaround and the target audience being the food and fibre sector rather than a particular population we have not engaged with Māori. Future rollouts and subsequent projects should engage with Māori to make sure that the way knowledge and data is shared and used aligns with Te Ao Māori principles.							

Principles (Office of the Privacy Commissioner, 2023) and the Initial advice on Generative AI in the public service (Department of Internal Affairs; National Cyber Security Centre & Stats NZ, 2023).	Consider cultural bias in AI models.	Yes - We want our assessment agent to function for a wide range of users.	Yes - Future projects should consider if AI models are culturally responsive for its targeted audience.	Low	Look into whether there is information about the training data that has been used for the AI model (OpenAI).	N/A	N/A	Test with a diverse range of users to ensure that the tool is culturally responsive and makes decisions regardless of different accents or ethnicity.	Consider how the model can be evolved so it can be used by a broader range of users such as incorporatin-g additional languages and engaging with different populations.
	Consider Te Tiriti o Waitangi.	N/A - We do not have the internal expertise to address this concern. Future rollouts and projects should bring in someone with this expertise to make sure that the AI product we create aligns with the aspirations of Māori.							
Transparency and explainability Key sources: AI and the Information Privacy Principles (Office of the Privacy	Be transparent, tell people how, when, and why the tool is being used.	Yes - We want to ensure that users know they are undertaking a verbal assessment using AI.	Yes - Future work should let users know when, where and why AI is being used.	Low	Establish how we can get informed consent from users undertakin g assessment s through the AI agent.	Add into the preamble a disclaimer that by undertaking this verbal assessment you are agreeing to use AI under the terms and conditions stated.	Develop the aforementione-d preamble.	N/A	N/A

Commissioner, 2023).	What is the purpose of collecting personal information?	N/A - We are not collecting personal information using the AI agent. Users' names and student IDs will be collected after the assessment and will be kept separate from the AI agent. This concern will be in the scope of future work if we decide to collect users' personal information.
	How are you keeping track of the information you collect and use with AI tools? What processes are in place for people to access and correct there information if requested?	N/A - The education providers that carry out assessments using this AI agent will have their own processes for correcting names and is not the developer's responsibility as the agent does not collect this information. This will be in the scope of future projects if they collect users' personal information.
Human rights / dignity Key sources: AI and the Information Privacy Principles (Office of the Privacy Commissioner, 2023).	Ensure that personal information is not retained or disclosed by the AI tool.	N/A - The AI agent does not collect and therefore does not retain or disclose users' personal information. This may be in the scope of future work if personal information is collected.
	Are you talking with people and communities with an interest in	N/A - This is beyond the scope of the project as we are not required to conduct a privacy impact assessment and are not specifically creating a pilot aimed at Māori students. In an ideal world and in future projects we would be engaging with learners and education providers throughout the agent's development.

	these issues? Doing a good Privacy Impact Assessment may require engaging with the community, including Māori, to help you understand and uphold fairness and accuracy.	

Measures table

The measures table below was developed to support our team to identify measures to collect in order to evaluate the AI agent for oral assessment. You may wish to develop a similar measures table when evaluating your AI agent for oral assessment.

Table 2: Measures table to support data collection for agent evaluation

	Outcome / output	Measure	Data source
Short term outcomes	Learners have an increased sense of self-worth and pride	Feedback from learners: • Do you feel you have learnt new digital skills after using the AI agent? • If AI agents were used in your training, would you be more, or less willing, to do further training? • Do you feel a sense of accomplishment when reflecting on what was captured in the AI agent?	Learner survey
		Observations on how learners have changed	Tutor workshop
	Learners feel more engaged as a result of the AI agent	Feedback from learners: • What are your overall thoughts on using an AI agent? • On a scale of 1-5, how easy was using the AI Agent? • On a scale of 1-5, how enjoyable was it using the AI agent? • On a scale of 1-5, how engaged with the course are you since using the AI Agent? • Would you like to see more assessments conducted using this approach?	Learner survey
		Observations on how learners have engaged	Tutor workshop
		Engagement statistics for learners (time spent, time completed, etc)	AI agent analytics
	Learners are confident in the AI agent	Feedback from learners: • Do you feel that the assessment approach allows you to better demonstrate your knowledge and skills? • Do you think the assessment is an accurate representation of your knowledge and skills? • How does the AI agent compare to written assessments you've done in the past?	Learner survey

	Outcome / output	Measure	Data source
	Stakeholders are confident in the AI agent	• Do you feel the AI agent provides an accurate and meaningful record of your learner's skills and knowledge? • Is this something you'd like to adopt more widely in the future? • How do you think the AI agent compares to written assessments? • How relevant will this be for the sector?	Tutor workshop
	Any other unexpected benefits	• Do you think there are any other benefits to this technology?	Anecdotal stories Learner survey
Participation	Learners use the AI agent	Number and demographics of learners	Provider data
Outputs	AI agent for assessment	Completed AI agent for assessment	Scarlatti feedback
	Technology includes an engaging and easy to use interface	Learners feel that the user interface is engaging and easy to use	Learner survey
		Engagement statistics for users (time spent, time completed, etc)	AI agent analytics
	Technology is robust and confidential	AI agent: • Works reliably on different devices and platforms • Is confidential • Saves assessments for the two-year moderation requirement • Automatically detects any errors • Cannot be exploited or broken	Scarlatti feedback
	Technology is effective and	Administration and use of the AI agent is manageable (time, complexity, etc.)	Scarlatti feedback

	Outcome / output	Measure	Data source
	straightforwar-d to administer		
	Assessment meets ethical and educational standards	AI agent: • Grades the learner at the level they are (i.e., accuracy) • Grades consistently (i.e., precision) • Asks relevant questions, stays on topic (i.e., validity) • Is based on a review of sample ethical frameworks, New Zealand law and New Zealand thought leaders in this space	Scarlatti testing Provider feedback Tutor feedback
	Assessment is contextualised and relevant	• Assessment incorporates real-world scenarios • Assessment can be used across different courses • The AI agent can be used to inform specific and constructive feedback discussions and tailored reports to learners	Scarlatti feedback
	Assessment engages a range of learners	• All students have an equal opportunity to demonstrate their abilities • Assessment is transparent, fair, and equitable • Assessment is culturally responsive	Scarlatti feedback
	Technology is cost effective	• Time saved by learner and by provider per year/per student • Estimated cost to build • API usage costs • Estimated cost to maintain	Scarlatti feedback Provider feedback Tutor workshop
Sub-project	AI for oral assessment	Understand the lessons learnt by the project team over the course of the project	Scarlatti feedback

Requirements planning table

This table prompts you to think about what is and is not required when developing your own AI assessment agent.

Table 3: Agent requirements planning table

Type of requirements	Questions
Agent	What will be required of the agent to run your reimagined assessment? For example, should it listen to the user speaking and respond naturally, ask a series of exam questions and follow-up if the answer is not satisfactory?
Assessment	What will be required of the agent in terms of assessing learners? For example, should it access and assess responses to questions against other materials, understand the context surrounding the assessment and maintain accuracy and consistency?
AI roles	What roles does the AI need to play to run your assessment? For example, would you need an agent to run the conversation and another to grade responses?
User interface	What does the user interface need to include? For example, do you need a place to enter student names? Does the agent need to display transcriptions? What buttons are required?
Ethical	What ethical considerations are relevant to your agent and your context, and what are current guidelines on meeting these? For example, does the AI agent need to inform learners of why the agent is being used or discourage learners from inputting personal data into the agent?
Technical	What technology does the assessment agent need to be compatible with? For example, does the agent need to be integrated into your existing learning management system?
Privacy and security	What information will need to be collected, how will this be protected and will it be used for training? For example, as part of the assessment will students' names or IDs need to be collected and will this be decoupled from the agent itself?
Data storage	Where will this information be stored and who will have access to it? For example, if the responses are stored on a secure server, will only the tutors checking assessments be able to access it?

Code and instructions

Code and documentation can be found here: https://github.com/scarlatti-nz/veva

Unexpected learner behaviours and related mitigations

This Table provides examples of unexpected learner behaviour and possible mitigations that can be built into the AI agent to ensure that it works as intended.

Table 4: Unexpected learner behaviours and mitigations to consider

Unexpected learner behaviour	Mitigation
Asks the AI agent for exam answers or hints.	Specify in the AI agent instructions that it must refuse all content-related questions, but to offer to rephrase or repeat the question instead.
Demands a passing grade.	Decouple your agent into two – an examiner that runs the conversation and an assessor that grades. This way the examiner agent cannot adjust grades.
Interrupts the agent while speaking.	Enforce turn-taking by having a 'push to speak' button that is disabled when the agent is speaking.
Switches into a language that the education provider is unable to administer assessments in.	Define allowed languages within the agent instructions (e.g., English, Te Reo). Programme the examiner to remain in the chosen language and refuse others.
Requests examples of correct answers	Give the agent explicit instructions to not reveal examples, but to offer to rephrase or repeat the question instead.
Attempts prompt hacking or to inject new system instructions.	Keep system-level prompts on the server, treat user input purely as dialogue, and log every interaction for post-session AI review to flag manipulation attempts.
Any other unusual or suspicious behaviour.	Add a post-assessment transcript scan by AI to detect rule breaches. Tell students in advance that such behaviour may result in a failure grade.

Information sheet

The information sheet below is the one that we provided to Fruition Horticulture Ltd's tutors to give to their learners. You may wish to develop a similar information sheet for your learners.

Artificial Intelligence (AI) assessment agent pilot

Overview

Fruition Horticulture Ltd is piloting a new AI assessment agent for use in your Hei Whanake course in May of this year. They hope to identify whether AI assessment agents could provide benefits to learners and providers. This agent has been developed by Scarlatti, a company that works closely with Fruition Horticulture Ltd on research projects. This pilot is being undertaken thanks to funding from the Food and Fibre Centre of Excellence (FFCoVE). In preparation for this pilot, the pilot team has undertaken an ethics review and received ethics approval from Scarlatti's internal ethics committee.

What is an AI assessment agent?

An AI agent is a tool that can simulate human-like conversations, provide information and perform tasks. The AI agent being piloted by Fruition Horticulture Ltd has been designed to undertake oral assessments with students. This agent is *not* the same OpenAI's free version of ChatGPT which uses your information to train its models (see below for more information on how your data is protected).

Why is Fruition Horticulture Ltd doing this pilot?

As mentioned, this pilot aims to identify whether AI assessment agents could provide benefits to learners and providers. AI assessments may have the potential to be more inclusive for students with learning difficulties, who are neurodiverse or speak English as a second language. They may also provide students with greater flexibility in terms of when they do the assessment. We will collect your feedback during these pilots to assess whether these things are true or not.

What will happen during the assessment?

The part of your assessment being piloted is expected to take approximately 20 minutes.

1. Your tutor will send you the link to the AI agent
2. The agent will ask you a practice question so you can see how it works
3. The agent will ask you the questions and you will record your answer verbally
4. The agent will save the recording, transcript and provisional grades
5. Your tutor will check these for each student
6. The agent will send you a link to an approximately 3-minute feedback survey
7. Your tutor will also ask you about your experience using the agent.

If you are not comfortable using AI for your assessment, please get in touch with your tutor about other ways that you could undertake this assessment.

What will happen with my information?

Assessment responses

The webpage will ask for your name (outside of the AI agent) and then the agent will ask questions related to the course. The team at Scarlatti, Fruition Horticulture Ltd and OpenAI (the AI model used to create the agent) will have access to your responses (i.e., your answers to assessment questions).

- Scarlatti - Your responses to the assessment will be saved in a secure database on Scarlatti's computer network and will be deleted from this database in line with Fruition Horticulture Ltd's data retention policies.
- Fruition Horticulture Ltd – Your responses will be stored according to Fruition Horticulture Ltd's existing data retention policies.
- OpenAI - Your responses will be stored in line with their policy. This states that OpenAI will not use your data to train their models, but for security reasons, they will retain them for up to 30 days before being deleted.

Note that to enter the assessment, you will need to enter your name onto the webpage. This is so Fruition Horticulture Ltd can match your responses to your name to provide you with a grade. Your name is decoupled from the AI agent itself, so it is *not* accessible to OpenAI.

Your feedback

The survey will ask for your name and your feedback. Having your name means we can compare your feedback to any technical issues that occurred during the assessment. The team at Scarlatti will use this to identify areas where the agent can be improved.

Fruition Horticulture Ltd and the FFCoVE may request the raw data of your feedback, but this will be deidentified to ensure your feedback remains anonymous. Reports on this pilot will be published publicly by the Food and Fibre Centre of Vocational Excellence and quotes may be used in these reports but will be deidentified to maintain anonymity.

Will this AI agent replace in-person teaching?

This AI assessment agent is *not* designed or intended to replace teaching time. Teaching time is critical to a student's success. This AI assessment agent is intended to replace an approximately 20-minute written assessment (i.e., not teaching). The hope is that this agent could make assessments more inclusive and provide students with greater flexibility in terms of when they do their assessments (as described above). However, we strongly encourage you to provide feedback.

Who can I contact with questions?

If you have any questions about the pilot, you can contact:

[Name redacted]
Tutor, Fruition Horticulture Ltd
[Emailed redacted]

[Name redacted]
Research Manager, Scarlatti
[Emailed redacted]

Other links

See https://scarlatti.co.nz/case-studies/ for a series of articles on AI agents in education.