

November 2025



SCARLATTI
Evaluation | Analytics | Insights

AI agents for oral assessment

Research findings and ideas for the future





Today's agenda

1

**Intro
to project**

(Adam)

2

**Demo and
mechanics**

(Phoebe)

3

**Pilot
findings**

(Adam)

4

**Progress
since then**

(Adam)



AI agent for oral assessment



Problem:

Written assessments are a barrier for many learners



Purpose:

Test whether an AI agent could conduct and grade an oral assessment

Meet the (extended!) team



Adam Barker
Director
Scarlatti



Phoebe Gill
Project manager
Scarlatti



Sam Cormack
Lead developer
Scarlatti



Stella Sutherland Stapleton
Associate researcher
Scarlatti



Tiffany Andrews
Academic manager
Fruition



Jenny Sinclair
Operations manager
Dairy Training



Jackie Lynch
Portfolio Manager
Food and Fibre CoVE



Lester Hoare
GM quality assurance
Muka Tangata



Contract Milking Assessment

This assessment is for Dairy Training Ltd. It will cover:

- Typical contract issues
- Guidelines for issues
- Who to go to for help



DairyTraining
Dairynz



Tutor reviews outputs

- Answer with provisional grade
- Reasoning for grade
- Transcript and audio.

DTL_Contract milking 101

AGENTS RESULTS STUDENTS

< DTL_Contract milking 101 >

Name

sam.cormack@gmail.com

Vaughan Wilson

library@scarlatti.co.nz

phoebe.gill@scarlatti.co.nz

kayla.jordan@scarlatti.co.nz

Assessments for Vaughan Wilson

Assessment:

31/10/2025 31/10/2025 21/10/2025 21/10/2025 10/10/2025 7/10/2025 7/10/2025 7/10/2025 2/10/2025 2/10/2025

Question	Student Response	AI Grade	AI Feedback	Tutor Grade	Tutor Comment	Assigned by	Updated	Actions
What do you consider to be one of the most likely issues to arise during a contract milking term?	A disagreement with the farm owner around milk production targets and whether they had an idea of what that should be, and we didn't meet that.	C	Competent. Divergent expectations around milk production targets are one of the most common contract-milking flashpoints, so your answer demonstrates awareness of a realistic issue.	NYC	Reason for disagreement	Sam Cormack	7/10/2025, 9:23:54 am	

HIDE TRANSCRIPT

Audio Recording: 0:00 / 3:43

Transcript

User: Hello

Assistant: Kia ora and welcome to your Contract Milking Assessment! This assessment measures your ability to understand and respond to contract requirements and issues, preparing you for success in the New Zealand Dairy Industry. Before we begin, please repeat after me, "This assessment is entirely my own work."

User: This assessment is entirely my own work.

Assistant: Great, thank you. Let's get started with the first question: What do you consider to be one of the most likely issues to arise during a contract milking term?

User: Some sort of disagreement with the farm owner maybe around like milk production targets and sort of whether they, you know, had an idea of what that should be and we sort of didn't meet that?

Assistant: Thanks for that answer. Next question: What are the contract guidelines or rules for this issue?

User: Sorry, can I just go back to that question? I didn't quite finish my train of thought there. So I guess sort of we may have maybe they hadn't their milk production target in their head, but we didn't sort of agree on it.

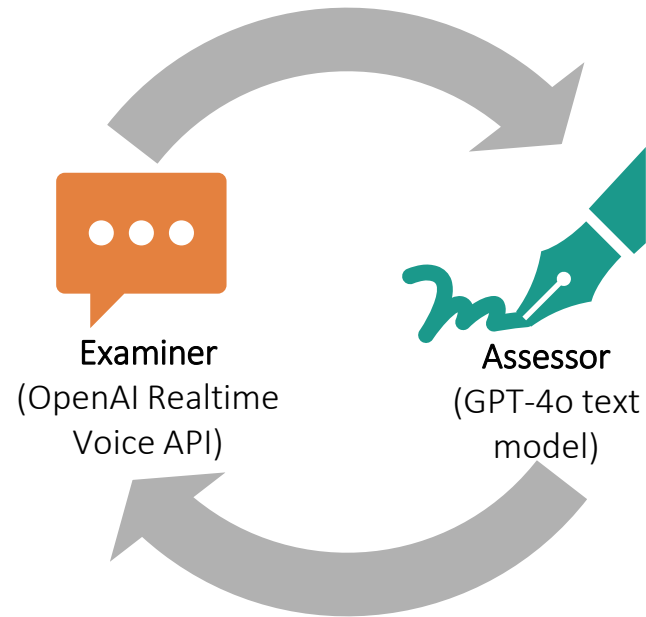
Assistant: Thanks for adding that. Let's move forward. What are the contract guidelines or rules for this issue?

User: So in the contract you might have a sort of plan for the milk production targets and a schedule for that?



A double agent

- There are two agents used
- We do this for specialisation and answer protection
- Note: Fully text options are also possible.





Set assessment questions

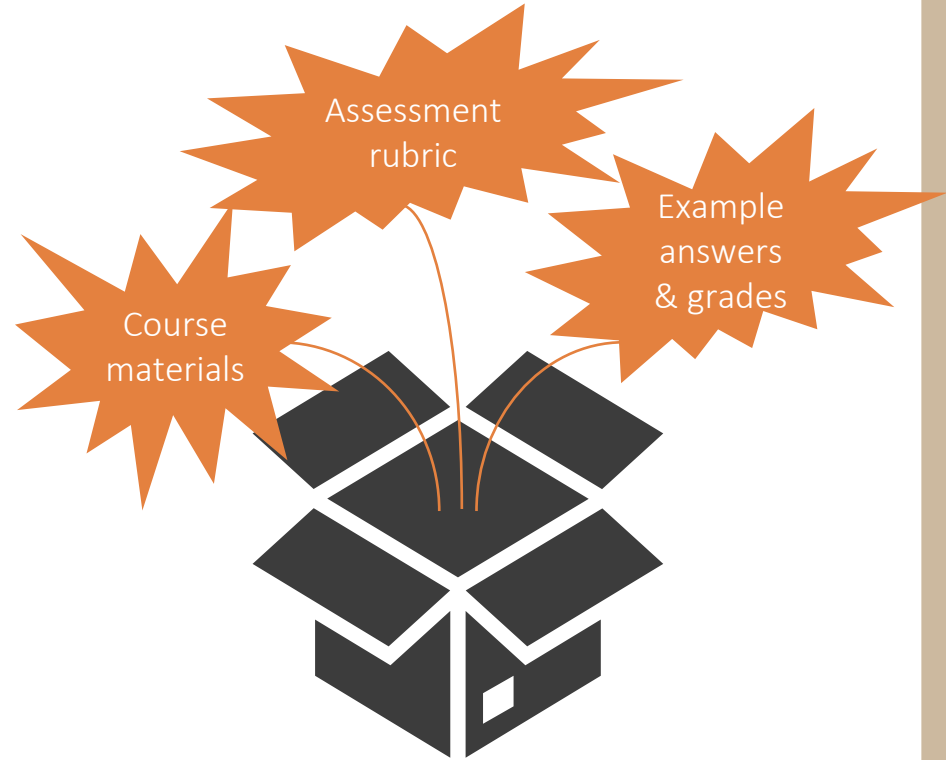
- The educator provides the questions
- This controls the conversation.

Q1: What is a common contract issue...



Prompting content

- We give the agent information for it to decide how to respond/grade
- This avoids it relying only on generic information from online





Principles

1. Privacy
2. Security and resilience
3. Academic integrity and fairness
4. Content safety
5. Transparency and consent
6. Human oversight
7. Continuous improvement



How it works



Platform set up and hosting

We choose the strongest models and settings to build from.

- Google Cloud Platform hosting (Sydney)
- Encryption at rest and in transit
- OpenAI LLM, with own safety mechanisms
- Exclusion of personal identifiers



Agent configuration

We configure each agent and test robustly before release.

- Prompt engineering
- OpenAI Output Guardrails
- Human red-teaming



User interaction

We combine measures to verify user identity.

- LMS-linked login
- Information sheet & explicit learner consent
- Authenticity statement
- Real-time guardrail checks
- Timing controls
- Question design



Post-conversation

We keep humans in the loop to ensure accountability & trust.

- Human grade oversight (by ed provider)
- Audio and transcript storage
- Automatic evaluation suite
- Deletion and retention policy

Continuous feedback

AI evaluation detects fixes needed

How to cheat (and how to stop it)

Cheats

- Access ChatGPT during chat
- Have a friend answer
- Use Google glasses or hidden earpiece
- Use Comet or similar
- Use voice clone
- ...

Mitigations

- No questions in advance
- Authenticity statement
- Audio recording & transcript
- Conversation and lag timer
- Question type (follow up, process, personal)
- Randomized questions
- Video and screen grab
- Combine with in-person
- ...



Contract Milking 101 (Level 5)

Provider:
Dairy Training Ltd

Format change:
Oral tutor to Oral AI

Aim:
Inclusive for remote and / or neurodiverse
students and saves tutor time

Demographics



11 students
Mixed ethnicities
Aged ~22 to ~40

AI Agent



Follow-up questions
during
Feedback at the end



Hei Whanake H&S (Level 2)

Provider:
Fruition

Format change:
Written to Oral AI

Aim:
Inclusive for young Māori learners, illiterate learners, and save tutor time

Demographics



14 students
Predominantly Māori
and Pasifika
Aged ~16

AI Agent

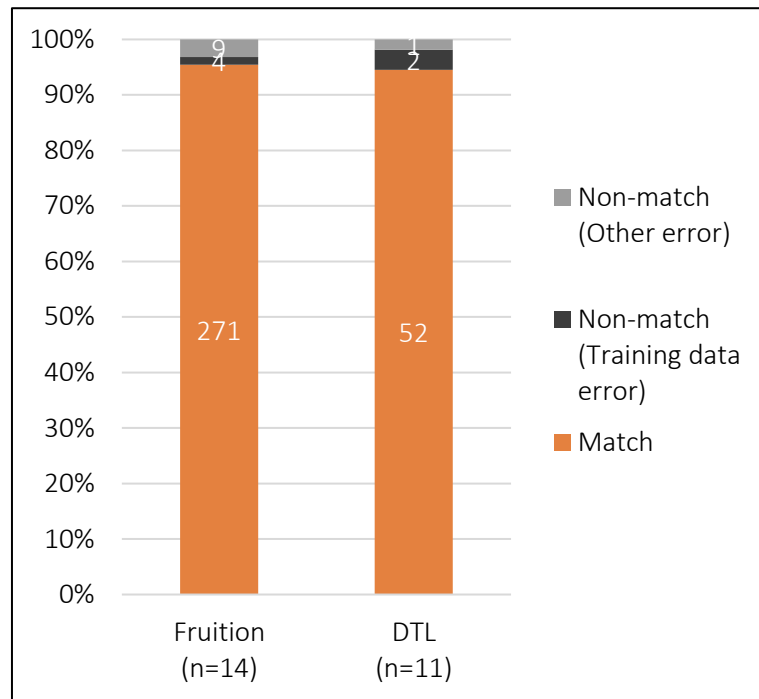


Follow-up question at
the end
No feedback



It appears accurate

- 95% match between AI and human
- Even tutors can disagree
- Many non-matches are addressable.



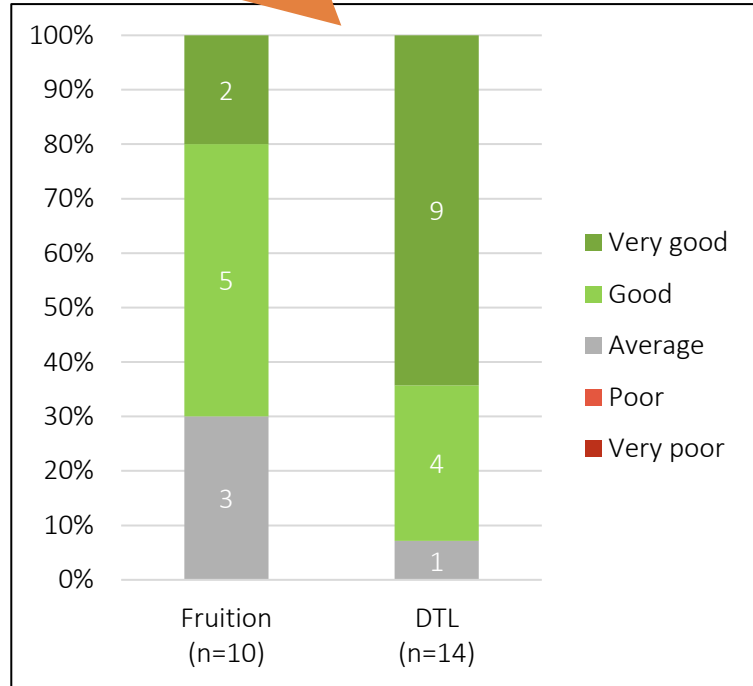
Pilot findings



“Pretty straight forward” (Learner)
“Odd experience at first” (Learner)
“Very clear and understands you well” (Tutor)

It appears easy to use

- 20/25 said “easy” or “very easy”
- Some technical challenges
- To be fair, learning to use for first time
- Some *Fruition* learners prefer writing.



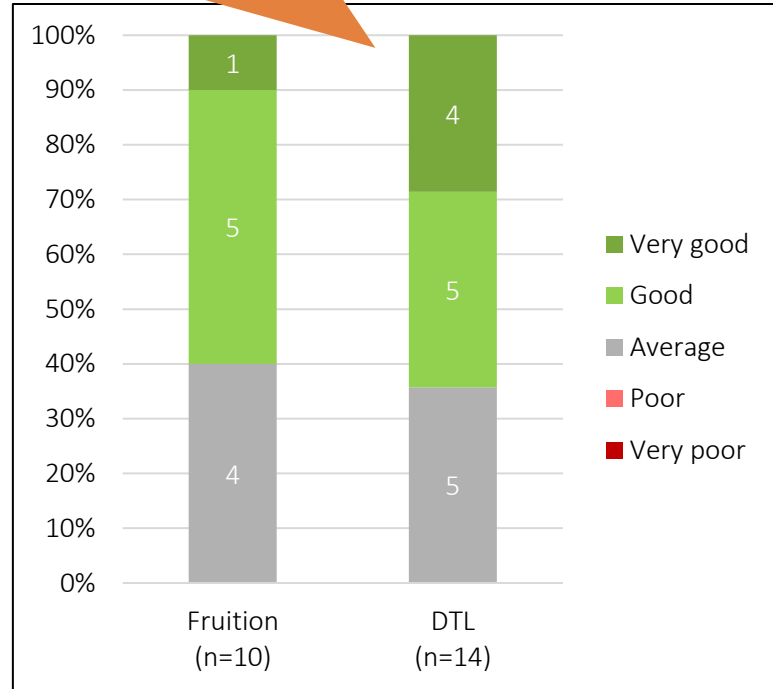
Pilot findings



“You can do it anytime you want” (Learner)
“The instant feedback was great” (Learner)
“We need to talk to people... not computers” (Learner)

It appears quite enjoyable

- 15/24 said “enjoyable” or “very enjoyable”
- Particularly appreciated instant feedback, flexible timing, ability to ask questions, & easier to talk than write
- But *still need human contact*.





It appears to empower

- Seen most in Fruition learners
- Ability to ask questions and not be judged, assessment felt easier/faster
- Built belief in ability to do well.

*“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)*



It will likely save time

DairyTraining
Dairynz

4 to 6 tutor hours saved per assessment

0 learner hours saved per assessment

FRUITION
Horticulture

15 to 20 tutor hours saved per assessment

28 to 33 learner hours saved per assessment

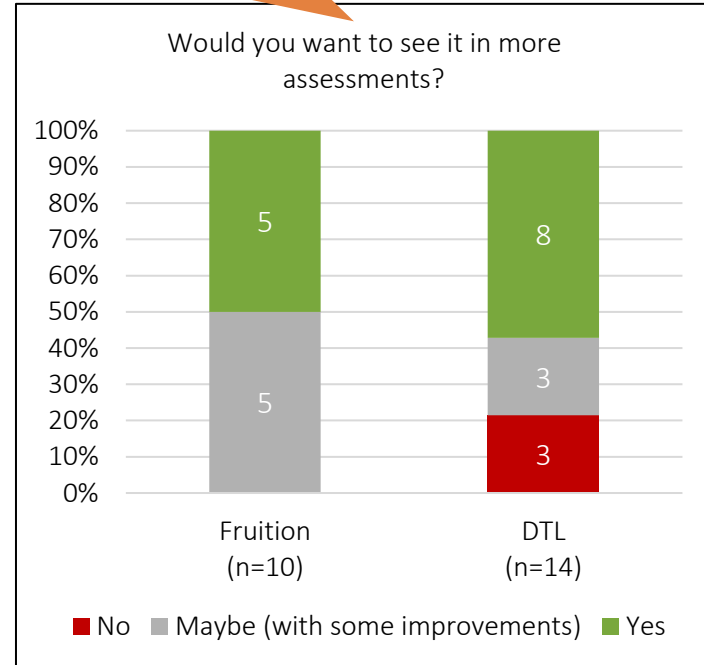
Note that Dairy Training's hours saved are less given they shifted from a tutor-run oral assessment to an AI-run oral assessment. Fruition shifted from a written assessment to an AI-run oral assessment.



"I prefer writing, just personal preference'
(Learner)

Do learners want to see more of it?

- 21/24 either interested to see more of it OR open to it with improvements
- Technical fixes and showing learners how to use it will help
- But – also *evidence that some learners still prefer writing.*

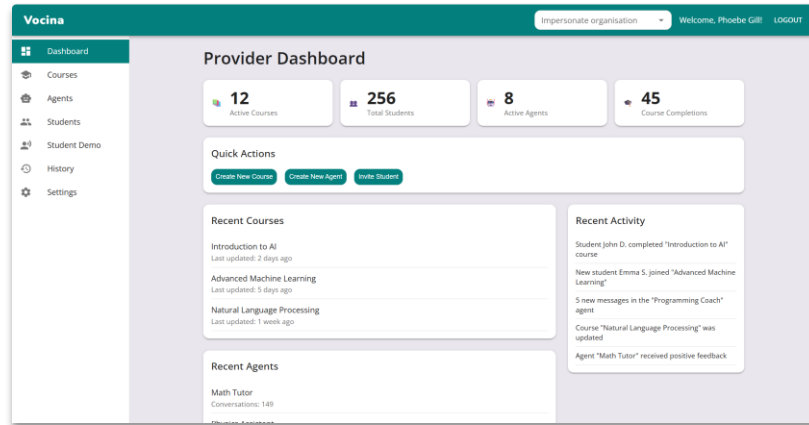


Progress since then



Building Vocina

Create and deploy an AI agent to run an oral assessment, within hours.





Exploring ideas

We have ~8 other pilots starting, and are looking for more.





Exploring ideas

Within those pilots, there have been two new use-cases

Formative



Essay verifier





Exploring ideas

We are looking for other organisations to pilot with.

Summative



Formative



Essay verifier



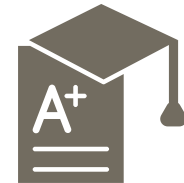
Employer verifier



Pre-screening



RPL



Questions & answers



Come talk to us!



Adam Barker
Director
Scarlatti



Phoebe Gill
Project manager
Scarlatti



SCARLATTI
Turning research into action

www.scarlatti.co.nz