ThinkActuary



Who marks the markers?

Developing Automated Marking

2017–2026: making expert-informed evaluation and feedback more accessible.

Burt Verster / Matthew Smith

ACTUARIAL SOCIETY OF SOUTH AFRICA
2026 CONVENTION

ThinkActuary

Expert feedback on students' practice exam answers is scarce, slow and expensive.

Student answer

Practising is possible

→

Awaiting expert review

Access is uneven

→

Independent feedback

What earned credit? How can I improve?

ThinkActuary

2017: a first attempt at automated marking

The first Mark prototype tested whether technology could help.

Candidate answer

Represent the wording

→

Memorandum points

Compare using cosine similarity

→

Suggested marks

Apply a threshold and add credit

78%

of marks within one standardised mark

23

pilot attempts

Promising credit estimates, but similar wording can express incorrect reasoning.

ThinkActuary

ThinkActuary connected students with actuaries

Students gained access to independent feedback. We began building a valuable dataset of actuarial judgement, applied through marking.

Student → **Actuary** → **Student**

Attempts a past-paper question → Marks the answer → Receives a mark and feedback

>73k

cumulative question submissions

>6k

actuary reviews

>110

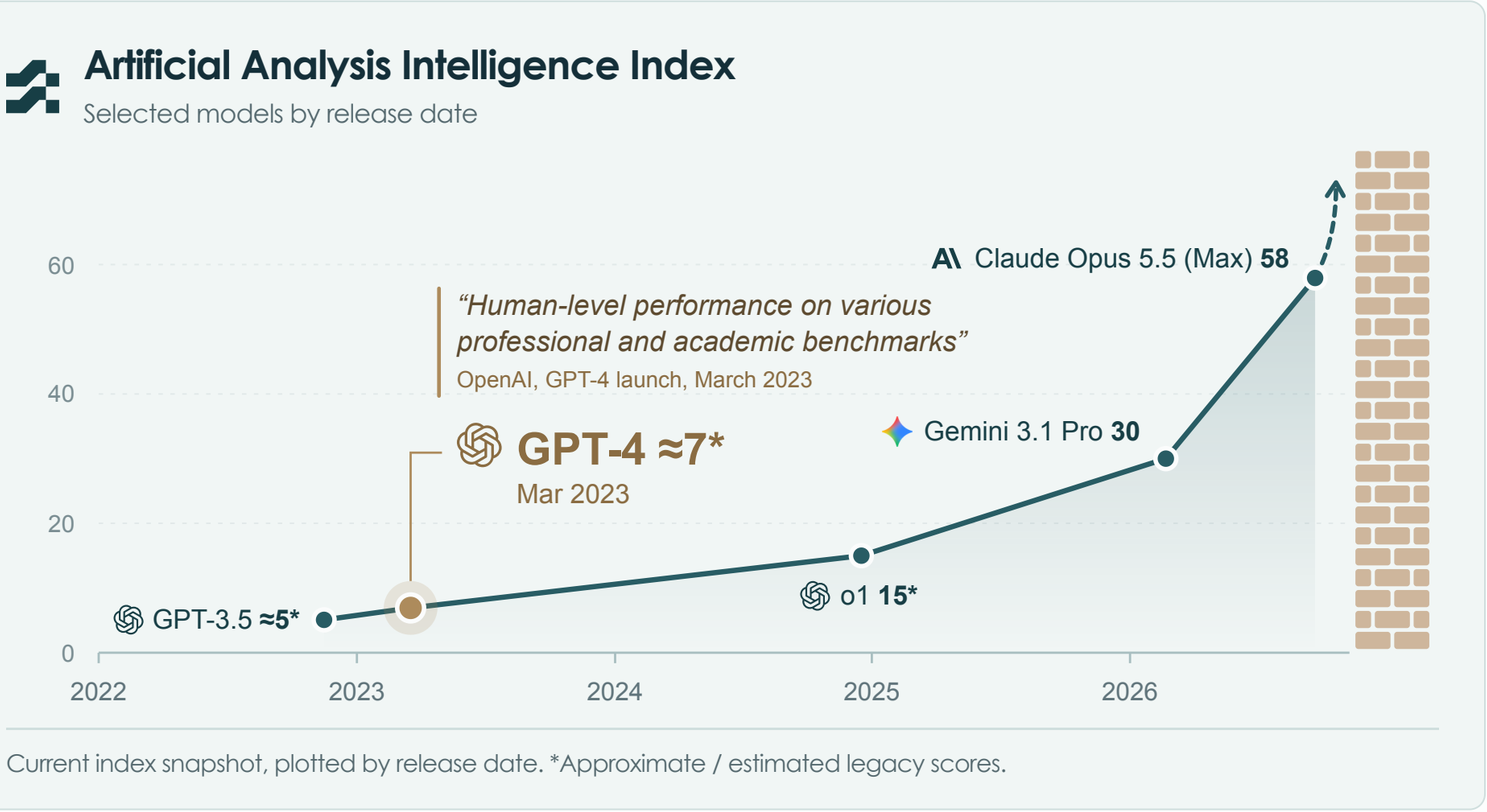
actuary markers

Through 30 June 2026.

ThinkActuary

In 2023, GPT-4 caught our ‘attention’

LLMs might offer a path beyond text matching.



Current index snapshot, plotted by release date. *Approximate / estimated legacy scores.

01 **Could it understand meaning?**

02 **Could it apply context?**

03 **Could it explain its decisions?**

Reliable marking still needed data, structure and governance.

ThinkActuary

Hi, I'm Mark · the new marker
New starter

Meet Mark, a promising new marker

Bright and well-read, but had never marked an exam. We needed to figure out how to train him.

Brain THE MODEL

OpenAI
Anthropic
Gemini
Grok
Z.ai

Method THE HARNESS



▲ VERCEL WORKFLOW

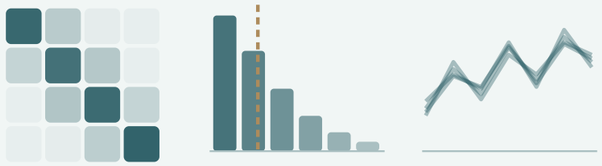


AGENT SKILLS

Textbooks THE EVALUATION

Actuary marking

CONFUSION
MAE
STABILITY



ACTUARIAL SOCIETY OF SOUTH AFRICA
2026 CONVENTION

ThinkActuary

Hi, I'm Mark · the new marker
In training

Learning from senior markers

Every reviewed script is feedback. Every change is checked before it sticks.

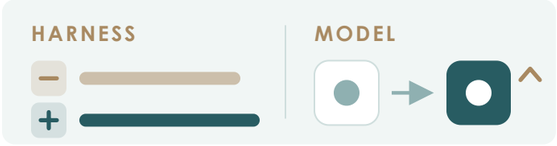
01
Learn from reviews

Understand marking doctrine across exam levels and subjects.



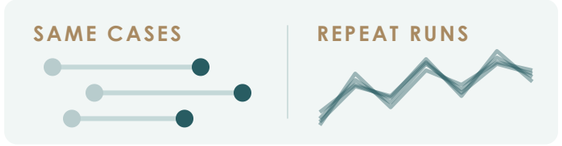
02
Improve the engine

Harness: refine skills and workflow.
Model: try a stronger model when guidance is not enough.



03
Implement and evaluate

Compare on the same cases.
Repeat runs to check stability.



Repeat

Change one thing at a time.

Our dataset powers the loop. People approve changes.

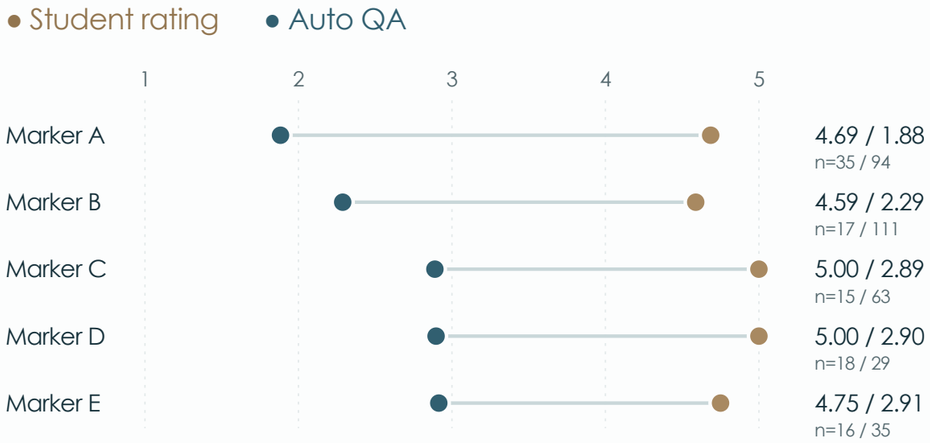
ThinkActuary

Hi, I'm Mark · the new marker
In training

Even senior markers disagree

Our eval set was valuable, but platform reviews showed the human reference wasn't perfect either.

Mean student rating and auto QA / 5

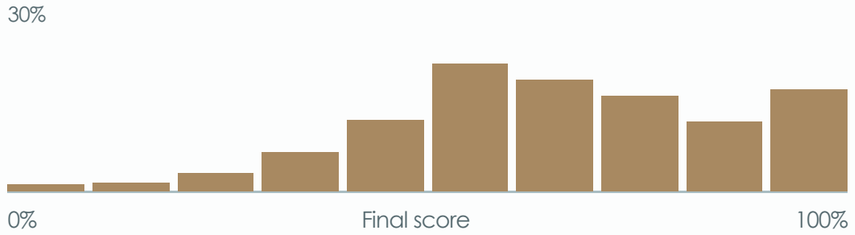


2025 Session 2 · Markers with ≥15 student ratings
Auto QA range: **1.88–2.91/5**, despite high student ratings.

Final-score distribution

2025 S2

Mean 65.1% · n=685



Slightly left-skewed, with a longer tail towards lower scores.

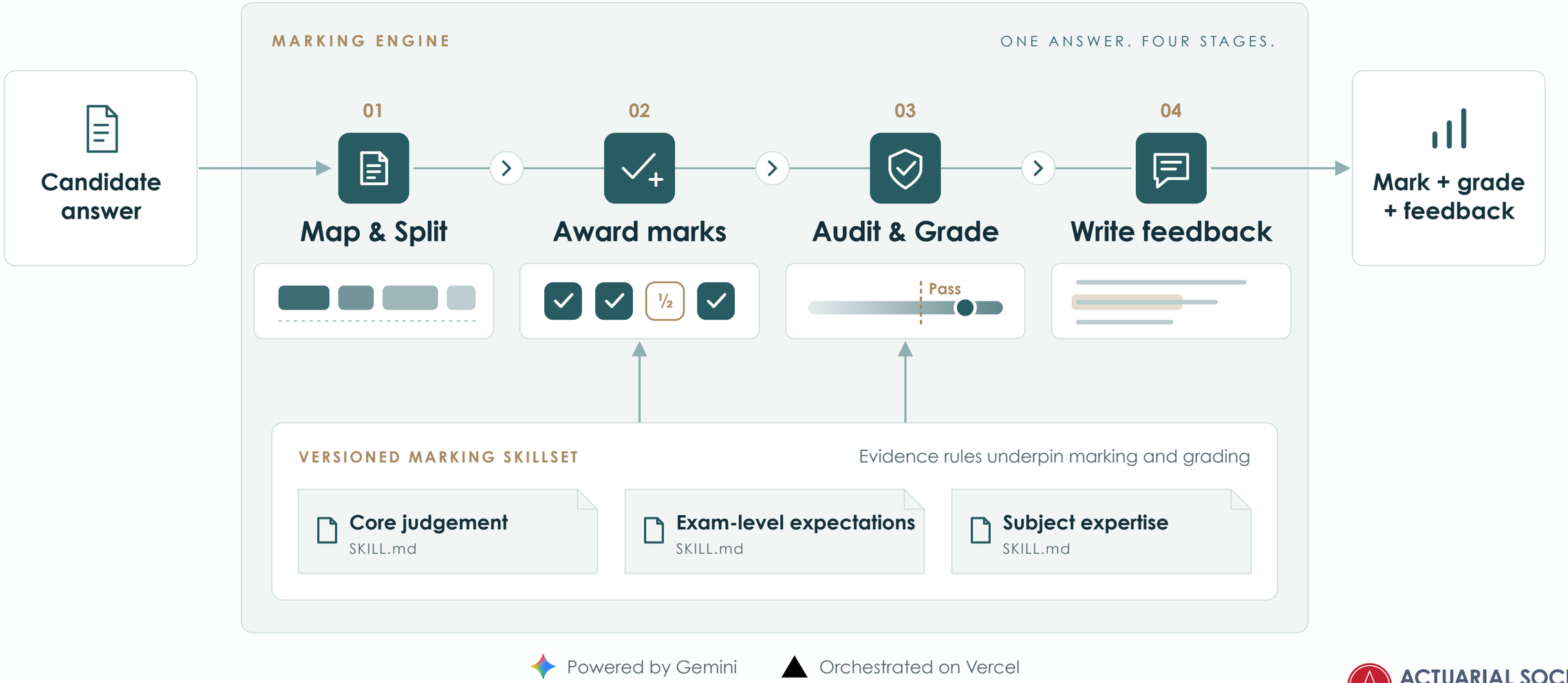
Full session cohort; shape alone does not establish bias.

We had to keep these shortcomings in mind when measuring accuracy. But the noticeable variability between markers also got us thinking.

ThinkActuary

Hi, I'm Mark · the new marker
In training

We taught Mark a method



Powered by Gemini Orchestrated on Vercel

ACTUARIAL SOCIETY OF SOUTH AFRICA
2026 CONVENTION

ThinkActuary



Hi, I'm Mark · the new marker
Trainee (under supervision)

By Feb 2026, Mark was ready for supervised marking

Gemini 3.1 Pro powered all workflow steps. Marking doctrine stabilised, with decent accuracy and consistency.

71.3%

Pass/fail agreement

72 agreements from 101 scripts.
29 disagreements.

Enough promise to explore

Inspectable proposed marks and feedback supported supervised marking.

Still important errors

Average scores hid individual differences. Context and duplicate credit needed work.

Promising support for actuaries, with human review still essential.

ACTUARIAL SOCIETY OF SOUTH AFRICA
2026 CONVENTION

ThinkActuary

First, help the actuary

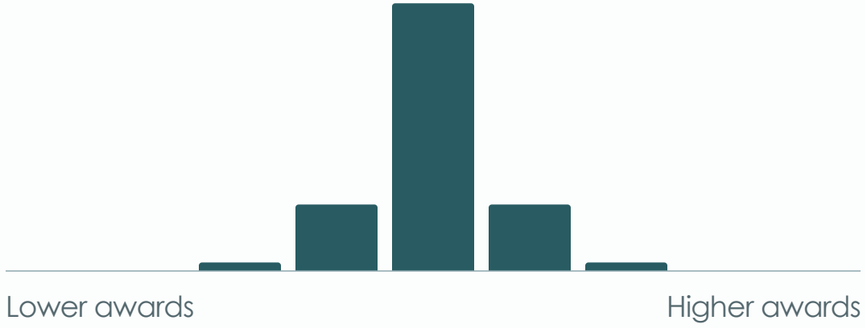
Shared guidance and in-line recommendations could help structure marking and feedback.

Existing guidance



Expected improvement

In-line recommendations



Conceptual illustration, not measured results. No same-script, multiple-marker experiment is shown.

The hypothesis: more structured marking and feedback, with the actuary retaining the final decision.

ThinkActuary

Grader brought the engine into the workflow

2026 S1. The engine proposes. The actuary reviews, changes and approves.

Student attempt

Answer and question

→

Marking Engine

→

Actuary review

Segment decisions and final sign-off

→

Student feedback

Approved mark and guidance

Actuaries stay in control

Inspect each segment's rationale and confidence. Edit awards, review the overall feedback and approve the result.

A real review decision

"Consider the tax implications if the structure were to change"

Engine: **0 marks** → Actuary: **0.5 marks**

Anonymised F103 review, 2026 S1. Record excerpt, not a UI screenshot.

Engine-supported marking.
Actuary-approved feedback.

- grader.thinkactuary.com ↗

ACTUARIAL SOCIETY OF SOUTH AFRICA
2026 CONVENTION

ThinkActuary

Early signs were encouraging

The session before Grader (2025 S2) compared with the first session using it (2026 S1).

Final scores

Before Grader

Mean 65.1% · n=685



With Grader

Mean 53.7% · n=292



More concentrated with Grader; different cohorts, not causal evidence.

Written-feedback QA

Mean score / 5

Before Grader · 30 reviews

3.03

With Grader · 30 reviews

4.20

+1.17 points in mean QA.
Same prompt, balanced samples.

Student feedback

Mean voluntary rating / 5

Before Grader · 1,169 ratings

4.5

With Grader · 101 ratings

4.76

Zero-to-five scale.
Before Grader: 2020–25.

Review timing

Median elapsed time

Before Grader **With Grader**

Final ownership interval

31 min
2025 S2 · n=682

23 min
Early S1 · n=66

Early estimate, not active work.

Overall turnaround

5.3 h
2025 S2 · n=635

2.0 h
2026 S1 · n=276

62% shorter, including queue time.
Post-review QA proxy.

Higher feedback QA is encouraging, but these comparisons do not establish marking accuracy or a causal improvement.

ACTUARIAL SOCIETY OF SOUTH AFRICA
2026 CONVENTION

ThinkActuary

Actuaries accepted nine in ten segment awards unchanged

2026 S1 Grader: paired segment awards, not whole-script decisions.

91.2%

8,683 of 9,517
paired segment awards unchanged

834 changes mattered

Overrides exposed missed credit, over-credit and borderline decisions.

Acceptance needs interpretation

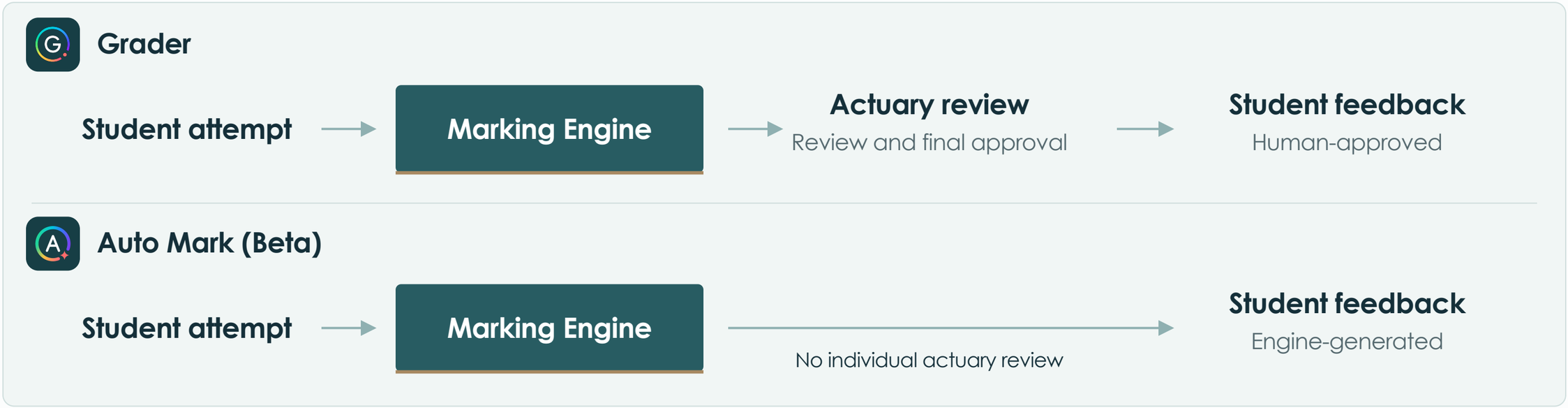
It may reflect good recommendations, human anchoring, or both.

Could the engine already provide useful feedback before an actuary reviewed it?

ThinkActuary

Auto Mark removed the individual review step

Students get feedback in a matter of minutes, at a quarter of the cost. A311 and selected F100 subjects only.



Student price, relative to actuary marking



75%
cheaper with Auto Mark

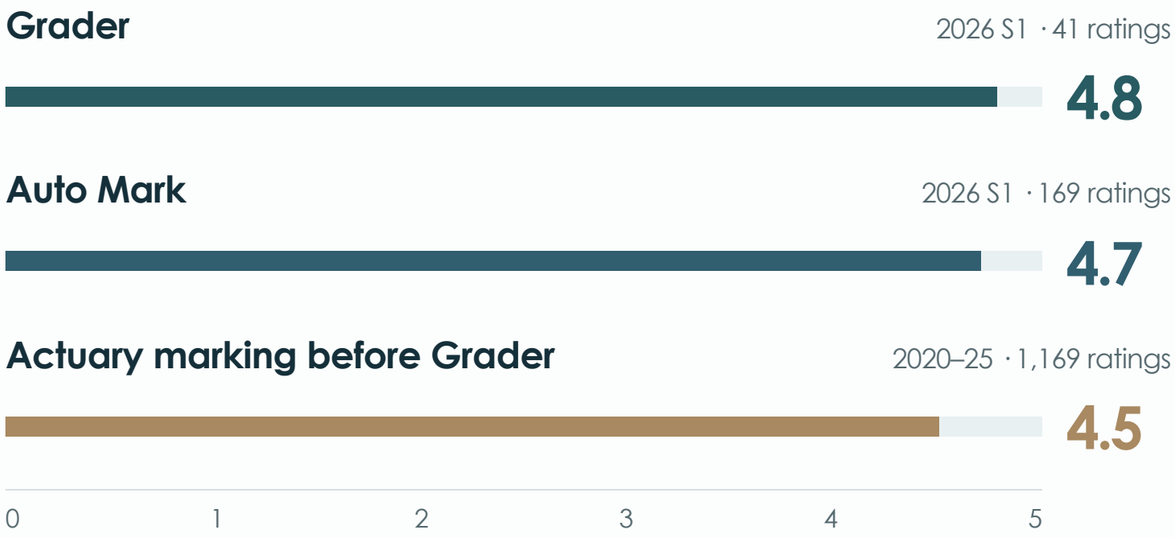
ACTUARIAL SOCIETY OF SOUTH AFRICA
2026 CONVENTION

ThinkActuary

Students rated Auto Mark highly

Voluntary ratings in the same six supported subjects, across different cohorts.

Average student feedback rating / 5



2026 S1 beta use

10.8% of submitted marks in supported subjects, per the session report.

Limited scope

A311, F102, F103, F105, F106 and F108. No F200 subjects in this beta cohort.

What still needs proof

Monitor failures and feedback. Compare matched scripts with human review and test repeatability.

Encouraging student-value evidence. Different cohorts and subject weights prevent a superiority claim.

ThinkActuary

NINE YEARS OF PROGRESS

With the release of Auto Mark (Beta), we made meaningful progress on the original objective.

Scarce →

Accessible

Students can access engine-generated feedback without individual actuary review.

Slow →

Quick

Human sign-off is no longer a prerequisite for returning feedback.

Expensive →

Affordable

A lower-priced route: 25% of the standard credit rate before rounding.

ThinkActuary

Beyond marking: helping students learn

Better feedback can support better learning. The next steps are independent evaluation and stronger feedback.

01

Evaluate with ASSA

Proposed retrospective collaboration using scripts already marked by ASSA examiners. We are also working to understand the impact on pass rates.

02

Improve the engine with Grader

Use actuary decisions and overrides to refine skills and model choices. Include difficult and borderline cases, and test changes before widening use.

An independent comparison is the next evidence step, not a study we have already completed.

ThinkActuary

The people who made it possible

IN MEMORY OF

Wim Els

Where it all began

His contribution started the journey that brought us here.

OUR SINCERE THANKS TO

Paul Lewis

Imraan Mohamed

Siobhain Mahoney

Arne Verster

Kavi Pather

Instrumental in helping us get here.

ThinkActuary

Questions

2017

Expert evaluation and feedback were difficult to access.

Now

Auto Mark (Beta) makes useful feedback more accessible.

Future

Exciting possibilities for personalised learning and the rise of the “Agentic Actuary”.