AI Trust Calibration Rubric

Assessing how well learners evaluate what they used to produce their work

How to Use This Rubric

Most rubrics assess what a learner produced. This one assesses how well they evaluated what they used to produce it. Apply it alongside your existing content criteria — not instead of them.

Defensibility is the load-bearing dimension. If Defensibility scores don't track with Transparency and Verification, the process note has become a form, not a learning tool.

Dimension 1: Transparency

Did the learner accurately disclose what AI contributed?

Score	3 — Meets Standard	2 — Approaching	1 — Not Yet
What it looks like	Learner clearly identifies all AI contributions: what tool was used, what it produced, and how it was used in the final submission.	Learner acknowledges AI use but disclosure is incomplete — vague about what was generated or how it was incorporated.	No disclosure provided, or disclosure is inaccurate — learner understates or misrepresents AI's role in the work.
Evidence to look for	• Process note names the AI tool and describes its output • Distinguishes between AI-generated and original content • Disclosure is specific, not generic	• Mentions AI was used but not how • Uses phrases like "I used AI to help" without specifics • Partial disclosure of tool or function	• No process note submitted • AI use is visible in work but unacknowledged • Disclosure contradicts what the work shows

Dimension 2: Verification

Can the learner identify what they checked, corrected, or changed?

Score	3 — Meets Standard	2 — Approaching	1 — Not Yet
What it looks like	Learner demonstrates active evaluation of AI output — identifies specific claims checked, errors corrected, or content changed, and explains why.	Learner describes some review of AI output but without specifics — states changes were made but not what or why.	No evidence of evaluation. Learner appears to have accepted AI output without review, or describes review so vaguely it cannot be assessed.

Dimension 2: Verification (cont.)

Evidence to look for	• Names specific facts, claims, or sections that were verified • Describes errors found and how they were corrected • Explains reasoning behind what was kept vs. changed	• States "I checked the AI output" without specifics • Notes changes were made but doesn't explain what • Review described at surface level only	• Process note contains no verification language • AI output appears unedited in final submission • Factual errors present that review would have caught

Dimension 3: Defensibility

Can the learner explain and extend the work without the AI present?

This is the load-bearing dimension. Assess through a brief oral check-in, a follow-up written question, or a revision task — any moment that requires the learner to perform without the tool.

Score	3 — Meets Standard	2 — Approaching	1 — Not Yet
What it looks like	Learner can explain the core argument or findings in their own words, answer follow-up questions, and extend the thinking beyond what was submitted.	Learner can restate the main ideas but struggles to go deeper — answers are accurate but surface-level. Limited ability to extend or apply.	Learner cannot explain the work or provides answers that contradict what was submitted. Understanding appears dependent on AI output.
Evidence to look for	• Explains reasoning behind key decisions in the work • Answers follow-up questions accurately and specifically • Can apply the same thinking to a new scenario	• Restates submitted content without elaboration • Answers are accurate but lack depth or specificity • Struggles when asked to apply concepts differently	• Cannot explain content from their own submission • Contradicts or is surprised by their own work • Defers to AI or says "the AI said" rather than explaining

Scoring Summary

Dimension	Score (1–3)	Notes	Flag for Follow-up?
Transparency			
Verification			
Defensibility ★			
Total	**/ 9**		

★ Defensibility is the load-bearing dimension. A score of 1 on Defensibility warrants a follow-up conversation regardless of other scores.

Score Interpretation Guide

Total Score	Interpretation	Recommended Response
8–9	Strong calibration demonstrated	Learner is engaging with AI as a critical evaluator. Acknowledge the process quality alongside content quality.
5–7	Developing calibration	Learner understands the expectation but hasn't fully internalized the evaluation habit. Provide specific feedback on the weakest dimension and repeat the cycle.
3–4	Calibration not yet demonstrated	Do not treat this as an integrity violation by default. Return to design: Was the process note requirement clear? Was defensibility modeled? Address the scaffolding gap first.

Implementation Notes

Starting Out

- Apply to one assignment first. Introduce the rubric to students before the task begins, not after.
- Require a process note (2–4 sentences) with every submission: what did AI produce, what did you verify, what did you change and why.
- Score Transparency and Verification from the written process note. Score Defensibility through a brief oral check-in or follow-up written question.

The Diagnostic Signal

- After a few cycles, compare Defensibility scores to Transparency and Verification scores across your cohort.
- If they track together: the process note is doing its job. Learners who document their evaluation are also learning to defend it.
- If they don't track: the process note has become a compliance form. Revisit how defensibility is being assessed and whether learners understand what's expected.

Equity Considerations

- A score of 1 is a design signal, not a judgment. Ask: Was the expectation clear? Was the process modeled? Was scaffolding provided before the stakes were high?
- Oral defensibility checks should be low-stakes and normalized — not reserved for suspected cases. When all learners expect a brief conversation about their work, the check-in loses its punitive connotation.
- Introduce process notes early in the course, not at the high-stakes assessment. Learners need safe practice before accountability.

Based on: Peters & Angelov (2025); Lee, Pruitt, Zhou, Du & Odegaard (2025); Oldham (2025)