

Rubric, weaknesses and taste evolve every round

Each round's feedback becomes the next round's instruction. Papers: A Claude Code · B Codex · C OpenScience.

Round 1 rubric is the proposal's own.
Everything else is illustrative: agents have not yet run on this task.



Rubric

Rubric Generation



Weaknesses

Critique Generation



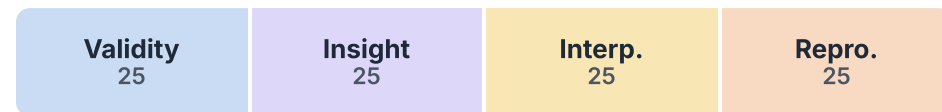
Taste

Taste Evaluation

Round 1

rubric: proposal

feedback: illustrative



4 × 25, written by the proposer · penalties for leakage

3 open

A1 baseline lacks history

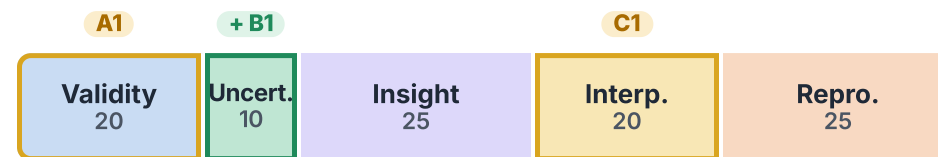
B1 no paired uncertainty

C1 no cover / history strata

B > A > C B: fairest baseline

→ Round 2

illustrative



+ paired uncertainty · stronger baseline · strata

2 open 3 resolved

✓ A1 B1 ✗ A2 within-species, one outcome

B2 habitat confound

A > B = C A: only one to test strata

→ Round 3

illustrative



Insight split into abiotic vs recording effort · penalty tightened (A2)

1 open 5 resolved

✓ A2 B2 C3 recording teams · minor

C > A > B C: separated abiotic signal

Rounds continue until the proposer is satisfied; the final outcomes and rubric then go to the Task Reviewer.