Real-world requests

230

Open-ended questions

① Frozen evaluation inputs

Question-specific rubric

Positive budgets: 20 / 35 / 30 / 15
Negative triggers, groups, caps

Agent configuration

Model + harness
+ reasoning effort

Frozen outputs

Response
Execution Trace

② Two-stage evidence audit

Stage 1: Provenance

Response

Claim extraction
Atomic claims

Trace

Grounded in Trace?

yes → **Grounded**

no

Stage 2: Independent verification

Independent external evidence

- verified → **Verified but ungrounded**
- unverified → **Unverified**
- refuted → **Hallucinated**

Claim + refuting evidence

③ Quality scoring

Inputs: response a + frozen rubric

Positive and other negative items
Binary judgments from the response

Factual-reliability negative items
Apply frozen trigger conditions to audit evidence

Positive score − effective penalties
Max per root cause · credit positive losses · apply caps

Quality score Q