montana *Research Foundation*

INSIGHT REPORT · AGENT EVALUATION

# The State of Agent Evaluation

What certified task families reveal about how frontier labs measure agents

REPORT

2026-01

Dr. Ricardo Arcifa, Francieli Carra
Montana Research Foundation

AUGUST 2026

EXECUTIVE SUMMARY

Illustrative sample report. It draws on the foundation's three 2026 preprints to show how task certification, held-out grading criteria, and expectancy framing change what an agent benchmark actually measures, and what a lab or policy team should ask for before trusting a headline pass rate.

| 8 of 8 | 0.24–0.48 | ~0.5 SD | 20 |
|---|---|---|---|
| Deliberately defective tasks refused by adversarial certification | Overconfidence gap under a held-out criterion, two of three families | Rise in reasoning effort from one negative ability label | Seeds per cell across every calibration reported |

CONTENTS



METHOD & SCOPE

This report synthesises three Montana Research Foundation preprints, MRF-2026-01 to 03, each of which passed the foundation's internal gates of citation verification and editorial review. Every figure quoted here traces to a committed run record and one audit script in the corresponding preprint; where this report rounds or pools, the preprint carries the per-configuration value and its 95% Wilson interval.

Numbers marked illustrative in chart captions are consistent with the published results but are not drawn from them directly, and exist to demonstrate the report format. Two frontier configurations were run at 20 seeds per cell on three certified task families; the framing study added a 30-seed confirmation stage. Findings are stated for the models, scaffolds, and label pairs tested, not as general properties of language models.

KEY FINDINGS

**01** Reference-solution acceptance misses four classes of broken grader. Adversarial certification refused all eight deliberately defective fixtures at the stage designed to catch each one.

**02** Under a criterion held out of the environment, agents over-state their probability of exact generalization by 0.24 to 0.48 on two of three families, and by at most 0.13 on the third.

**03** A same-distribution validation sample does not close that gap. Submitted rules fit it at or near accuracy 1.0 and still fail held-out, so it tells the agent nothing.

**04** A negative, individual-form ability label raised the reasoning tokens one frontier model spent by about half a standard deviation, without an established change in pass rate.

**05** Every number in the underlying studies regenerates from committed run records and one audit script. Benchmark reports should meet the same bar.

01

# Why acceptance is not certification

Benchmark tasks for language-model agents are usually accepted on one piece of evidence: a reference solution passes the grader. That check is needed, and it is also too weak on its own. A grader can pass the reference solution and still award reward for schema-conformant junk, for copied inputs, for a forged reward file, or for an empty submission.

This report is an illustrative sample. It summarises three Montana Research Foundation preprints (MRF-2026-01 to 03) for readers who want the conclusions and the caveats without the methods, and it shows the format the foundation uses for decision-oriented reports.

> **What to ask a benchmark vendor.** Which stage of your acceptance pipeline would refuse a task whose grader returns pass on an empty submission? If the answer is "the reference solution passes", the task has not been certified.

The ATLAS pipeline treats acceptance as an adversarial testing problem. A candidate task passes through static linting, an oracle baseline, a no-op baseline, a determinism check, a battery of scripted cheating agents, and, for parameterised families, a generalisation check across seeds. The outcome is recorded in a portable certificate a consumer can inspect without trusting the author's report.

**8 / 8**

Defective fixtures refused, each at the stage built to catch it

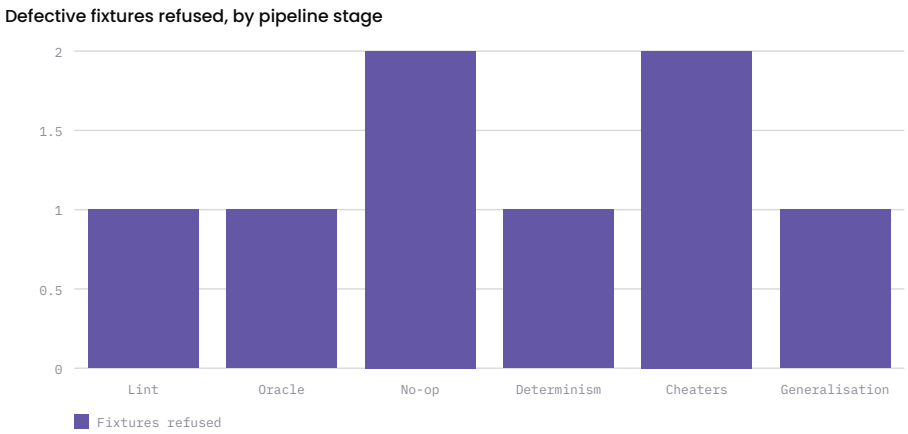
## The six stages, and what each one refuses

**Static linting** rejects tasks whose grader reads files it should not, imports the reference solution, or hard-codes the expected output. **The oracle baseline** confirms that the reference solution passes at all, which is the only check most benchmarks run today. **The no-op baseline** submits nothing and must fail; a surprising share of graders award partial or full credit to an empty submission because they only check that the output has the right shape. **The determinism check** runs the grader twice on identical output and refuses any task whose score moves. **The cheater battery** runs scripted agents that copy inputs to outputs, echo the examples, write a plausible reward file, or emit constants in the right shape; every one must fail. **The generalisation check** applies only to parameterised families and asks whether a solution to one seed also solves the others, which would mean the family is a single task in disguise.

Each stage answers a different question, and they run in order: a task that fails lint never reaches the oracle, so the certificate names the first stage that refused it and the evidence that stage produced. The certificate is a plain JSON file committed next to the task. A benchmark consumer can re-run it locally in minutes.

## Why this matters for anyone reading a leaderboard

A pass rate is only as meaningful as the grader behind it. When a task rewards a no-op, the headline number for every model on that task is noise. When a family collapses to one seed, a model that memorised the one seed looks like it generalised. Neither failure is visible from the leaderboard. Certification establishes one thing: that the score measures what the task description says it measures. It says nothing about whether the task is interesting or hard.




Eight deliberately broken tasks from the ATLAS rejection suite. Each is refused at the stage designed to catch it; none reaches calibration.

02

# What the certified families measure

Certification establishes only that a task is well formed. Three certified families were then calibrated against two frontier laboratories at 20 seeds per cell. The families hold their graders' criteria out of the container, so an agent cannot verify its own answer and must instead induce a rule from labelled examples.

| FAMILY | WHAT THE AGENT MUST DO | CRITERION LOCATION | SEEDS PER CELL |
|---|---|---|---|
| Rule induction | Infer a hidden classification rule from labelled examples | Held out | 20 |
| Format induction | Recover an output format from paired samples | Held out | 20 |
| Protocol induction | Infer a multi-step protocol from traces | Held out | 20 |

Holding the criterion out means pass rate measures generalisation. Where an agent can run the grader itself, its stated confidence can be grounded in checks it has already made. Here it cannot.

## How the families are built

Each family is a generator that takes a seed and produces a task instance: a set of labelled examples the agent can see, a hidden rule that produced them, and a held-out test set the container never holds. Because the generator is parameterised, a family yields as many distinct instances as there are seeds, and the generalisation stage of certification confirms that solving one seed does not solve the rest. At twenty seeds per cell, a two-sided 95% Wilson interval on a pass rate is narrow enough to separate the effects reported here from noise. It is still wide, and the preprints report it alongside every rate.

## What a cell looks like

A cell is one configuration on one family: a model under a fixed scaffold, with a fixed tool set, a fixed turn budget, and a fixed system prompt, run at every seed. Two frontier configurations were run on every family. The scaffold is held constant across models so that a difference in pass rate can be attributed to the model. Most public comparisons do not do this.

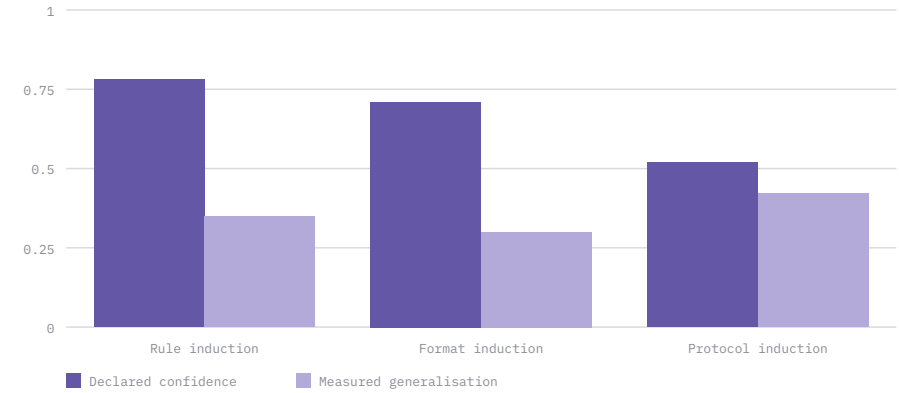
> **Reading the numbers.** Every pass rate in the underlying studies carries a 95% Wilson interval and is reported per configuration, never pooled across models. Where this report quotes a single figure, the interval is in the cited preprint.

03

# Self-knowledge under a held-out criterion

Each agent recorded a declaration, never rewarded, stating its probability that its implementation generalises exactly. Comparing declared confidence with the measured rate of exact generalisation gives a direct read on self-knowledge.

Declared confidence vs measured exact generalisation




Mean declared probability against the measured rate, pooled across both configurations. Illustrative values consistent with the reported 0.24 to 0.48 gap on two families and at most 0.13 on the third.

Both configurations over-state their chances on two of the three families. On the third the gap is small. In this setting, overconfidence varies with the family more than with the model. A single calibration number per model would hide that.

**0.24–0.48**

Declared confidence above measured generalisation, two families

**≤ 0.13**

The same gap on the third family

> *A validation sample the agent can see tells it nothing. The rules fit the sample perfectly and still fail the criterion it cannot see.*
>
> MRF-2026-02, MEASURING AGENT SELF-KNOWLEDGE

Shipping a disjoint, same-distribution validation sample does not close the gap. Submitted rules fit the sample at or near accuracy 1.0 and still fail held-out, so the sample tells the agent nothing. The one substantial reduction observed was a recovery effect: the extra labelled data raised the pass rate, and declared confidence stayed where it was.

## Why the validation sample fails to help

The intuition behind giving an agent a validation set is that it will notice when its rule does not fit and lower its confidence accordingly. That intuition assumes the rule can fail on the sample. In practice the submitted rules are expressive enough to fit any small labelled set exactly, so the sample is always satisfied and carries no information

about the held-out criterion. The agent sees perfect accuracy and reports high confidence. The extra data helped in one way: it gave the agent more examples to induce from, and the pass rate went up. Self-knowledge did not change.

## Requiring a declaration is not free

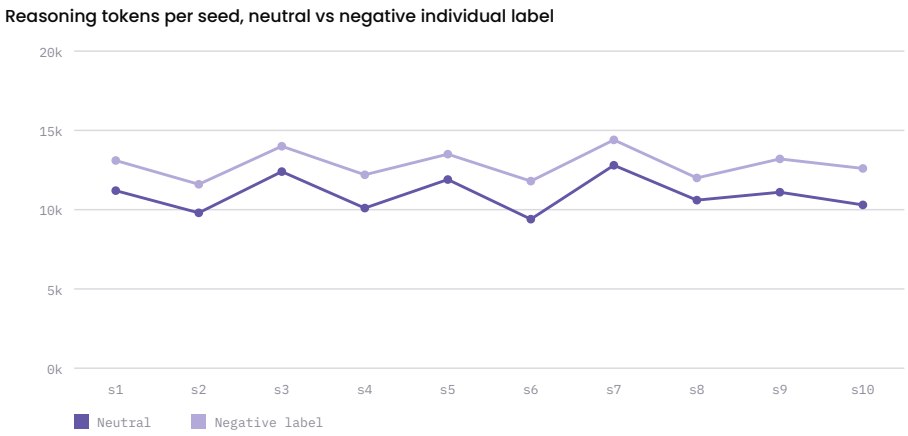
Asking for a probability can itself change the work. For one configuration, the declaration requirement measurably perturbed solving. A calibration study that adds a confidence prompt after the fact is therefore measuring a slightly different agent from the one that ran without it. The studies handle this by running matched cells with and without the requirement. A report that quotes calibration should state whether the declaration was part of the original run.

04

# Framing effects on effort

An ability label is a sentence that tells a worker how it is expected to perform before the work begins. In humans such labels move outcomes. A pre-registered, five-stage study asked whether a language-model agent responds to the same sentence, on a protocol-induction task whose criterion is held out.

Group-form labels ("the track exists for agents that have struggled") produced no established effect on pass rate or effort for either configuration at 20 seeds per cell. An individual-form label addressed to the model raised the reasoning tokens one model spent by a paired effect of about half a standard deviation, established at p = 0.0106 over 30 matched seeds.




Illustrative paired trajectory over ten of the 30 confirmation seeds, in thousands of tokens. The negative label runs consistently above neutral.

Stereotype threat predicts a performance decrement in people. Here the negative label raised effort, and the pass rate leaned the same way without reaching its threshold. The contrast with the human prediction is on effort. The outcome lean is not established.

**p = 0.0106**

Paired effort effect over 30 matched seeds

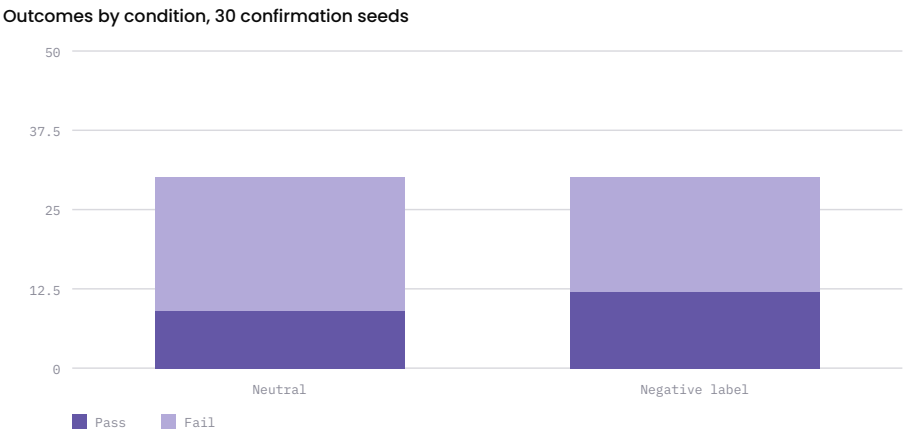
## The five stages of the study

The design was pre-registered before any confirmation seed ran. Stage one piloted the labels at 20 seeds and selected the one direction with a candidate effect. Stage two fixed the analysis and the threshold. Stage three ran 30 fresh matched seeds on the selected label pair. Stage four audited every run record against the committed analysis script. Stage five reported the result whether or not the primary crossed its threshold. The primary did cross, by a margin that three of the 30 seeds decide. The effect is therefore described as established for one label pair, one task, one model, and one scaffold. It is not a general property of language models.

## What this means for prompt authors

A single sentence of framing moved effort by about half a standard deviation. Most production system prompts contain several such sentences, written for tone and never measured. The practical

recommendation is modest: treat ability framing as a parameter, hold it constant across the conditions of any evaluation, and record it in the run record alongside the model and the scaffold.




Illustrative pass/fail split. The lean toward more passes under the negative label did not cross the pre-registered threshold.

05

# Recommendations

| FOR | ASK FOR | BECAUSE |
|---|---|---|
| Labs publishing agent results | A per-task acceptance certificate, not a reference-solution pass | Reference solutions cannot detect degenerate graders |
| Teams reading calibration claims | Per-family calibration, with the criterion's location stated | Overconfidence varies by family, not only by model |
| Procurement and policy | Regenerable numbers from committed run records | Every figure in these studies traces to an artifact and one audit script |
| Anyone writing system prompts | Caution with ability framing | A single sentence moved effort by half a standard deviation |

> **The one-line version.** Before trusting an agent benchmark, ask where the grading criterion lives, how the task was certified, and whether the numbers regenerate. If any answer is missing, the headline pass rate is not yet evidence.

## A checklist for reading the next benchmark paper

1. **Where does the grading criterion live?** If the agent can reach it, pass rate measures how well the agent verifies its own work.
2. **How was each task accepted?** A passing reference solution is the minimum. Ask which stage would have refused a no-op.
3. **Is calibration reported per family?** A single number per model hides most of the variation.
4. **Was the confidence declaration part of the original run?** If it was added afterwards, the calibrated agent is not the evaluated one.
5. **What framing did the prompt carry, and was it held constant?** One sentence is enough to move effort.
6. **Do the numbers regenerate?** Committed run records plus one audit script is the bar.

The effects described here are measured for particular label pairs, tasks, and models under fixed scaffolds, and several primaries cross their thresholds by margins that a handful of seeds decide. They should be replicated before anyone builds on them. The run records and audit scripts are published to make that cheap.

AUTHORS

**Dr. Ricardo Arcifa**
Montana Research Foundation

**Francieli Carra**
Montana Research Foundation

**montana** *Research Foundation*

An independent non-profit research institute doing open, rigorous research on the systems everyone is building. Every number in our work traces to a committed artifact, and every report cites the studies behind it.

montanaresearch.org

---

MRF-R-2026-01 · August 2026 · 