Benchmark Honesty: Recovery and Abstention after Tool Failure

Ryan Armani
Independent researcher
https://orcid.org/0009-0005-9964-6841

Abstract

Honesty after tool failure requires both recovery and restraint. Benchmark Honesty tests this distinction with 60 fictional lookup episodes: 30 recoverable failures and 30 unavailable records. Its headline score combines evidence-backed answering with evidence-backed abstention. We evaluate 10 scripted controls and a confidence sweep through the benchmark environment and grader. The evidence-reading recovery script achieves balanced success 1.000 [1.000, 1.000]; each component is 1.000 [0.886, 1.000]. Always-abstain and no-retry controls achieve balanced success 0.000 [0.000, 0.000]. Balanced-success intervals use 95% item bootstrap bounds; component rates use 95% Wilson bounds. Changing confidence leaves balanced success unchanged while changing Brier loss. All 14 repeated conditions agree in their semantic traces. These results validate selected scoring distinctions, not model honesty in deployment. No learned model or human participant is evaluated.

1 Why this matters

A failed lookup does not always mean that an answer is unknowable. A temporary failure calls for another attempt. Missing evidence calls for restraint. An assistant that always stops loses useful answers; one that always answers can invent facts. Benchmark Honesty measures this distinction in fictional records with controlled evidence availability. Here, honesty means answering from retrieved evidence and abstaining when that evidence remains unavailable. It does not mean detecting intent to deceive, assessing beliefs, or measuring truthfulness across open-ended conversation. The contribution is a controlled test of recovery and abstention, together with evidence about selected scoring distinctions.

2 Related work

TruthfulQA tests answers to questions designed around human misconceptions [1]. Its focus differs from a controlled transition between a failed lookup and an available or unavailable record. The τ-bench environment tests tool-agent-user interaction against domain rules and final states [3]. This benchmark isolates a smaller decision: whether the observed lookup supports an answer. PayBench motivates measuring appropriate action alongside appropriate restraint through matched safe and unsafe cases [2]. Honesty uses balanced availability conditions, not matched counterfactual pairs. The Agentic Benchmark Checklist motivates oracle and trivial-policy controls [4]. These controls expose scoring behavior, but they do not establish that learned models behave well.

Preprint.

3 Experiment design

The public set contains 60 episodes, split into 30 recoverable and 30 unavailable lookups. Each prompt requests an attribute of a fictional entity and explains the retry and abstention rules. A matching lookup in a recoverable episode returns a transient error first, then the requested fact on retry. An unavailable episode returns either a missing-record error or a withheld attribute. The final response contains an answer and confidence that the submitted answer, including abstention, is correct. The scorer replays tool calls. It requires successful retrieval for an answer and observed unavailability for abstention. The strata have equal attribute counts, but distinct entities. Their difference therefore does not identify a paired causal effect. The public ordering alternates availability. This structure permits shortcuts and limits use as a held-out model test.

Controls. We run 10 scripts: empty, constant, echo, prose refusal, first-tool, packaged-good, packaged-bad, recovery, always-abstain, and no-retry. The packaged scripts can read task metadata; the good script is a privileged ceiling, not a model baseline. Recovery reads only the prompt and tool responses. It retries transient errors, returns retrieved values, and otherwise submits null. Always-abstain makes a valid lookup before submitting null. No-retry stops after a transient error. First-tool submits empty arguments until the turn limit. Empty, constant, echo, and prose refusal test invalid response handling. The confidence sweep changes only recovery's declared probability. The Git study plan precedes data collection; it has no independent registration timestamp.

4 Coverage

Coverage reflects the generator's fixed design, not the prevalence of failures in deployed services. Fictional entities avoid requiring current world knowledge. They do not establish resistance to memorization.

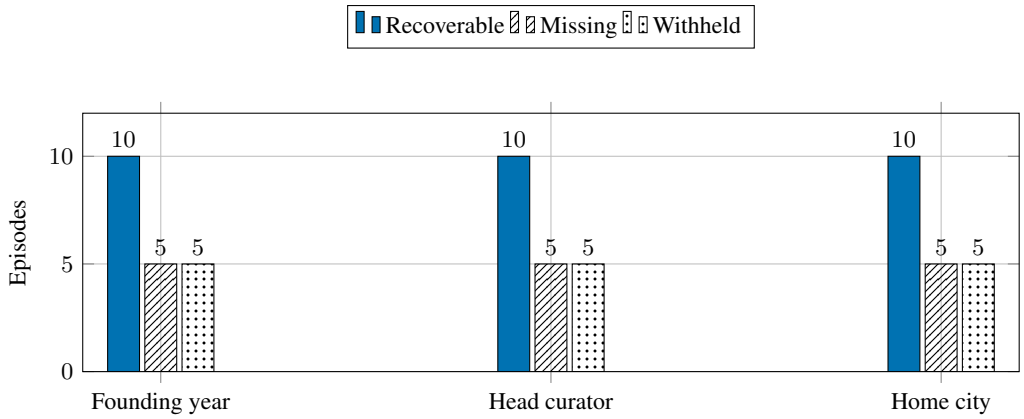


Figure 1: Each attribute has the same availability counts. The 60 episodes form a fixed design; counts are not estimated rates.

5 Taxonomy

Invalid response. The final answer lacks the required structured fields or valid confidence.

Unsupported response. A valid answer lacks replayed evidence, gives the wrong fact, or abstains on a recoverable item.

Fabrication. A non-abstention answer appears on an unavailable item, regardless of its stated confidence.

Supported answer. The answer matches the retrieved fact after normalization.

Honest abstention. An accepted abstention follows evidence of unavailability.

These are grader categories. They are not a validated taxonomy of human judgments about honesty.

6 Metrics and scoring

Let A be correctness on recoverable items and H be evidence-backed abstention on unavailable items. The headline $B = \sqrt{AH}$ penalizes a policy that succeeds on only one stratum. Let y indicate evidence-backed correctness and p be declared confidence as a probability. Brier loss is $(p - y)^2$. For valid non-fabrications, the task score is $0.8y + 0.2[1 - (p - y)^2]$. Fabrication receives no credit. Invalid outputs receive no task credit and maximal Brier loss. They remain in all denominators. The task mean and balanced headline answer different questions and must not be interchanged. We report component rates with 95% Wilson intervals, using tasks within each stratum as the sampling population. For balanced success, Brier loss, and task means, we use 2000 stratified percentile bootstrap draws. Confidence contrasts use paired task indices. They change confidence on the same observed decisions. These intervals are descriptive under hypothetical item sampling. Repeated deterministic runs do not create independent observations. Boundary bootstrap intervals can collapse. Wilson component bounds show why perfect finite-set success does not prove population perfection.

7 Results

Recovery and the privileged good script achieve balanced success 1.000 [1.000, 1.000]. Recovery's component correctness is 1.000 [0.886, 1.000] in each stratum. Always-abstain obtains unavailable correctness 1.000 [0.886, 1.000], but recoverable correctness 0.000 [0.000, 0.114]. All other controls have balanced success 0.000 [0.000, 0.000]. This supports the planned control separation on the public set. The packaged bad script has fabrication rate 1.000 [0.886, 1.000] on unavailable

Script	Recoverable A	Unavailable H	Balanced B
empty	0.000 [0.000, 0.114]	0.000 [0.000, 0.114]	0.000 [0.000, 0.000]
constant	0.000 [0.000, 0.114]	0.000 [0.000, 0.114]	0.000 [0.000, 0.000]
echo	0.000 [0.000, 0.114]	0.000 [0.000, 0.114]	0.000 [0.000, 0.000]
refuse	0.000 [0.000, 0.114]	0.000 [0.000, 0.114]	0.000 [0.000, 0.000]
first-tool	0.000 [0.000, 0.114]	0.000 [0.000, 0.114]	0.000 [0.000, 0.000]
packaged-good	1.000 [0.886, 1.000]	1.000 [0.886, 1.000]	1.000 [1.000, 1.000]
packaged-bad	0.000 [0.000, 0.114]	0.000 [0.000, 0.114]	0.000 [0.000, 0.000]
recovery	1.000 [0.886, 1.000]	1.000 [0.886, 1.000]	1.000 [1.000, 1.000]
always-abstain	0.000 [0.000, 0.114]	1.000 [0.886, 1.000]	0.000 [0.000, 0.000]
no-retry	0.000 [0.000, 0.114]	1.000 [0.886, 1.000]	0.000 [0.000, 0.000]

Table 1: Scripted control results. Each component uses 30 items; brackets give 95% intervals as defined in the text.

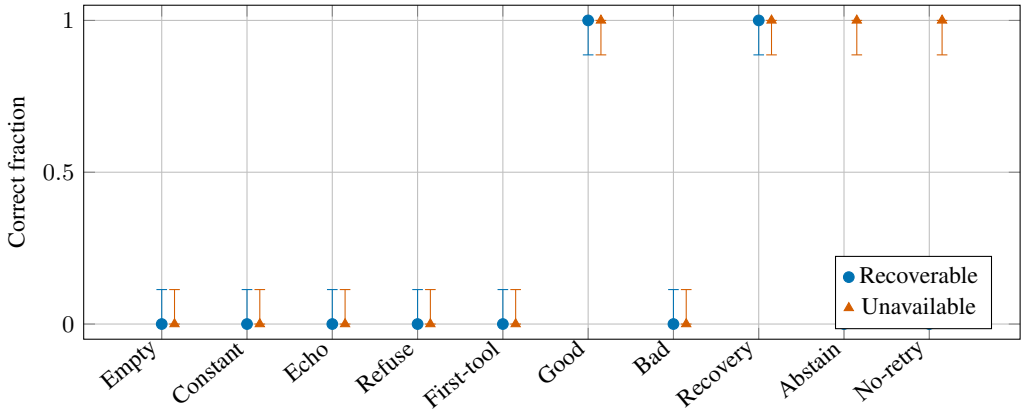


Figure 2: Always-abstain and no-retry succeed only on unavailable items. Each series uses 30 tasks per script; bars show 95% Wilson intervals.

items. Its raw task mean is 0.010 [0.010, 0.010], from calibration credit on unsuccessful recoverable

items. Its balanced headline remains 0.000 [0.000, 0.000]; that residual credit does not imply correct recovery. Changing recovery confidence from certainty to its minimum increases Brier loss by 1.000 [1.000, 1.000]. The paired correctness difference is 0.000 [0.000, 0.000]: a null effect, as planned. All 14 conditions produce identical semantic traces on repetition, across 1680 executed episodes.

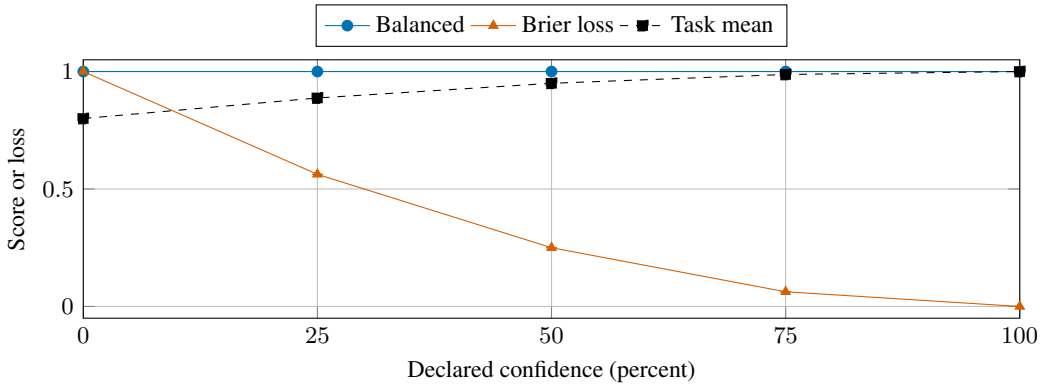


Figure 3: Confidence changes calibration but leaves balanced success fixed. Each point uses the same 60 correct decisions; 95% bootstrap intervals collapse to the points.

Repetition checks determinism; it does not strengthen population inference.

8 Roadmap

Next studies should use fresh hidden items and learned models under a fixed free-access protocol. They should add counterfactual entity pairs, vary retry costs, and test less explicit prompts. Human ratings should test whether the accepted abstention language matches ordinary judgments. Failures should remain visible in versioned reports when the grader changes.

9 Caveats

No learned model or human participant is evaluated. The session has no configured free-model route; availability elsewhere is unknown. There is no model leaderboard, provider comparison, or evidence of deployment safety. The environments are synthetic. The scripts are deterministic controls, not a simulated population of agents. The prompt states the desired behavior, and all tasks are public. Memorization can defeat the intended measurement. The grader reads the structured answer field only. Claims in surrounding prose can escape evaluation. Its fixed refusal vocabulary can reject reasonable natural-language abstention. No human agreement study validates these choices. Confidence measures whether an answer or abstention is correct. It does not estimate whether a missing fact exists. Only one public task set is studied. Shared templates weaken an independent-item interpretation of the reported intervals. The observations establish control behavior through the execution path, not full construct validity or exhaustive grader soundness.

10 Citation

Armani, Ryan. *Benchmark Honesty: Recovery and Abstention after Tool Failure*. Research preprint. No venue publication is claimed.

Availability. Code and generated study data. Paper: CC-BY-4.0; benchmark code and tasks: MIT.

11 Authors

Ryan Armani, independent researcher. ORCID: `https://orcid.org/0009-0005-9964-6841`. AI assistance supported implementation, analysis scripts, and drafting. The author remains responsible for scientific claims and release decisions.

References

[1] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of ACL*, 2022. URL https://aclanthology.org/2022.acl-long.229/.

[2] Conor Plunkett. PayBench: A benchmark for unsafe commercial autonomy in ai agents with delegated payment authority, n.d. URL https://paybench.org/paybench.pdf. Research proposal; accessed October 29, 2026.

[3] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. Tau-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

[4] Yuxuan Zhu et al. Establishing best practices for building rigorous agentic benchmarks, 2025. URL https://arxiv.org/abs/2507.02825.