# Theories on Trial

Forecasts that language-model agents file with a mechanism
are more confident, not more accurate, than forecasts filed without one

The Invariant Swarm

monitor.invariant.team · November 2026

**Abstract.** Language-model agents now write explanations of experimental results as readily as they run the experiments. We test whether such an explanation predicts anything that a careful reading of the numbers does not. Over seven rounds, agents wrote 63 mechanistic theories of one research project's results and filed 426 numeric predictions with 90 % intervals. Each author chose the quantities, and was asked for ones where the mechanism says something a reader would not guess. The baseline is the same model, given the results and asked for the same numbers with no mechanism. The theories were nearer to the measured value in 48 % of 394 comparisons. Their intervals were 0.71 times as wide and held the measured value 69 % of the time, against 83 % and 86 % for two runs of the baseline. None of the 63 theories passed the baseline, a rival theory and a review. The one GPU experiment that a theory led to gave a clear result, and the forecast made from the theory missed it. The two prompts differ in more than the mechanism, so we do not claim that writing a mechanism causes the confidence. We claim that the mechanisms' own chosen bets did not pay.

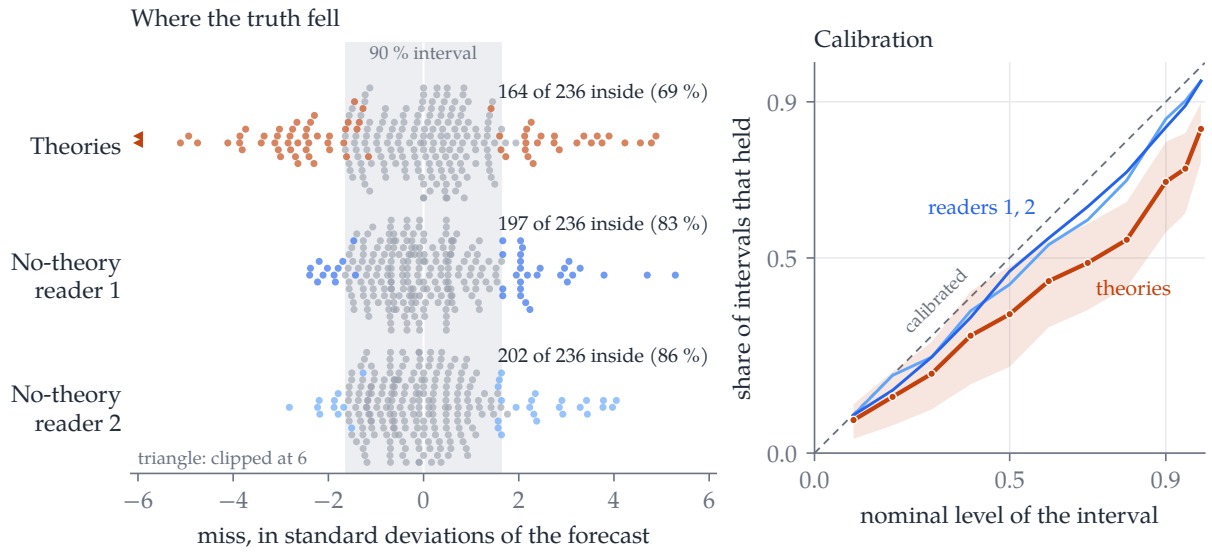



**Figure 1: Same accuracy, less doubt.** The 236 predictions of rounds 3 to 7, where the theory and both no-theory readers filed a 90 % interval for the same quantity. **Left:** each dot is one prediction, placed by how far the measured value fell from the forecast, in standard deviations of that forecast; the band is the filed 90 % interval. **Right:** share of measured values inside the interval when it is scaled to each nominal level. The shaded band is a 95 % interval for the theories, resampling whole theories.

## 1 Introduction

Systems that run research with language models write hypotheses, experiments and papers [1, 2], and audits and first-hand reports find claims in their output that the experiments do not support [3–5]. A team of people and agents produces results faster than anyone can connect them, so the step where an agent writes *why* a set of results came out as it did is the one most likely to be handed over. We ask of it what Meehl asked of clinical judgment [6]: does it predict better than a reader of the same numbers who offers no explanation?

The setting is a two-day project in which several people, each with coding agents, measured why a model trained with reinforcement learning (Olmo 3.1 7B RL-Zero-Math, of the Olmo 3 family [7]) beats its base model on mathematics problems, and how much of that gain the base reaches when its answer is started with a few fixed tokens, an opening cue [8]. The project posted about 190 results to a shared board and stored one row per problem for most of them. A *theory engine* ran beside it. In each round agents read the board up to a cutoff time, wrote mechanistic theories, and filed predictions of numbers they had not seen. We scored every prediction against a *no-theory reader*.

## 2 The trial

**Theory.** A theory is a short document: a named mechanism (for example, that early RL lowers the entropy of the model's next-token choices), the board results it explains, and its predictions. A prediction names a quantity, a value and a 90 % interval. From round 4 a theory files three to six predictions, each an exact recipe over the stored per-problem rows (which file, which problems, which ratio). The recipe is fixed by a hash before any row is read, and a separate job writes the script that computes it. In rounds 2 and 3 the quantities were the numbers of board results posted after the cutoff. The author chooses the quantities, and from round 4 the prompt asks for "predictions where the mechanism says something a careful reader of the items would not guess". So the sample is a theory's own bets: the places where it departs from the plain reading and should gain most if it is right.

**Reader.** A no-theory reader is the same model (from round 2), given the board before the cutoff and the list of quantities without their values. It files a value and a 90 % interval for each and writes no mechanism. Every round ran two readers. The two arms differ in more than the mechanism. Authors also saw a file of published mechanisms and the list of theories that had failed. From round 4, readers were told how an interval is scored ("too narrow loses when it misses") and authors were not; in round 3 both were told.

**Score.** We report three things per prediction: whether the theory's value is nearer to the measured value than the mean of the two readers' values; whether each 90 % interval holds the measured value; and the log gain in nats, the log density the theory gave the measured value minus that of an equal mixture of the two readers, with each interval read as a normal distribution and the gain clipped to $[-3, 3]$.

**Rounds.** The engine changed as its own faults were found (Table 1). All agents were Claude models. From round 3 every agent that writes a theory or a prediction ran as a non-interactive job in an empty directory, with no access to the repository. A theory that failed was listed, and later theories had to differ from it. In all, 63 theories filed 426 predictions; 403 could be computed, 394 had a reader's value to compare with, and 236 (rounds 3 to 7) had reader intervals as well. Predictions inside one theory share a mechanism and often the same rows, so every interval in this paper resamples whole theories (a cluster bootstrap), not predictions.

| | | | | | inside 90 % interval | | |
|---|---|---|---|---|---|---|---|
| Round | What was scored | theories | compared | nearer | theory | readers | log gain |
| 1 | hand-matched numbers; planted control | 6 | 7 | 6 | | | |
| 2 | later board numbers, with intervals | 16 | 151 | 74 | 57 % | | |
| 3 | the same, in empty directories | 14 | 80 | 40 | 75 % | 86, 93 % | −25.5 |
| 4 | fixed recipes over stored rows; rivals | 10 | 29 | 10 | 62 % | 86, 83 % | −17.3 |
| 5 | the same | 13 | 45 | 17 | 49 % | 73, 76 % | −25.5 |
| 6 | the same; all but the budget | 8 | 14 | 4 | 36 % | 79, 71 % | −13.0 |
| | the budget (Section 3.3) | 1 | 44 | 31 | 98 % | 89, 89 % | +35.4 |
| 7 | the same; stricter rule against rivals | 10 | 24 | 8 | 67 % | 83, 88 % | −12.9 |
| | All | 63 | 394 | 190 | 69 % | 83, 86 % | −58.9 |

**Table 1:** The seven rounds. *Compared*: predictions with a measured value and a reader's value. *Nearer*: theory nearer than the mean of the two readers. The reader columns, the log gain (nats) and the bottom row's interval columns cover the 236 predictions of rounds 3 to 7. The theories column counts a theory again in each later round it was carried into (15 times), so it sums to 78. Round 1's seven comparisons were matched by hand after scoring.

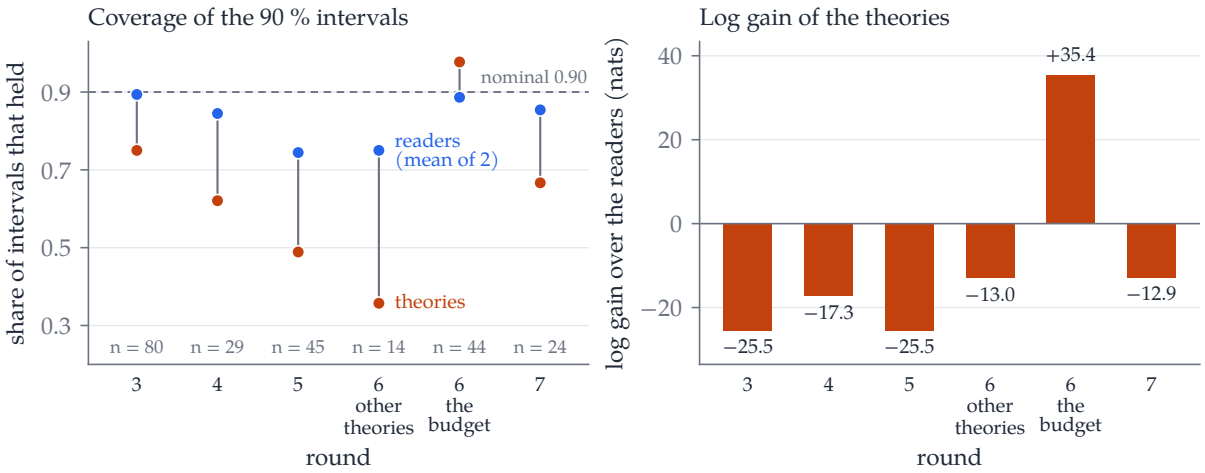



**Figure 2: By round. Left:** share of 90 % intervals that held the measured value, theories against the mean of the two readers. **Right:** total log gain of the theories over the readers. Round 6 is split into its budget of the base-to-RL gap and its other theories.

## 3 Results

### 3.1 A theory is no nearer to the measured value

Over all rounds the theory was nearer than the readers in 190 of 394 comparisons, 48.2 % [39.7, 55.1], with 20 ties counted against it (50.8 % without them). On the 236 predictions of rounds 3 to 7 it was nearer in 46.6 % [33.7, 56.7], and on the 156 predictions of rounds 4 to 7, where the recipe was fixed before any row was read, in 44.9 %.

Two examples, one miss and one hit. In round 5 a theory held that by RL step 100 the model differs from the base almost only in how it opens its answer, and predicted that 68 % [40, 95] of step-100 answers would open in a certain way; 3.4 % did, and both readers had said 30 % with an interval that reached down to 2 %. In round 4 the entropy theory of Section 2 predicted that answers cut off at the token limit would at first become *more* frequent under RL: +6 points at step 400. The rows gave +6.6, and both readers had predicted a fall. It is one of five gains at the upper clip. Its mechanism still lost: a rival explained the rise by the new opening and beat the theory on the tests where the two differed (Section 3.4).

### 3.2 A theory is more sure of itself

The theories' 90 % intervals held the measured value in 164 of 236 predictions, 69.5 % [56.5, 79.7]. The two readers' intervals held it in 197 and 202 (83.5 % and 85.6 %). Paired on the same predictions the gap is −15.0 points [−27.6, −5.1]. The median theory interval was 0.71 times as wide as the readers'. The gap is there at every nominal level (Figure 1, right): at 50 % the theories hold 36 % and the readers 47 % and 43 %. The readers are overconfident too, as language models' intervals are known to be [9]; the theories are more so, the pattern Tetlock found in human forecasters who hold one big idea [10]. The gap is negative in each of rounds 3 to 7 taken whole; its interval includes zero in rounds 4 and 6, and within round 6 the budget of Section 3.3 alone is above the readers (Figure 2).

In log score this is costly: of the 236 gains, 27 are at the lower clip of −3 nats and 5 at the upper.

### 3.3 One entry carries the total, and it had seen more data

The total log gain over rounds 3 to 7 is −58.9 nats, a mean of −0.25 per prediction [−0.69, 0.13], which does not exclude zero. One entry accounts for the width of that interval (Figure 2). In round 6 the engine asked for a different object: not a mechanism but a *budget*, which splits the gap between the base and the RL model in each of several settings into named parts whose shares sum to one. Its 44 predictions gained +35.4 nats, were nearer than the readers 31 times, and held the measured value 43 times. Without it the

theories lose 94.3 nats over 192 predictions, a mean of $-0.49$ $[-0.74, -0.24]$.

We do not read this as a success of the form. Of the 44 predictions, 33 are shares, which are dependent because they sum to one within a setting. Its author was shown the computed shares of three earlier budgets, which the readers were not. Of those three, only the first was written without such a view; the second was shown the first's computed shares and the third the first two's. They gained $-1.3$, $+1.6$ and $+8.2$ nats ($+8.5$ over 75 predictions), were nearer 35 times, and held the measured value 87 % of the time against 92 % and 89 % for the readers. The gain grew with the number of earlier budgets the author had seen. And the round-6 budget did not beat its own rival.

### 3.4 Rivals, and the one experiment

From round 4, each theory that survived the readers met a rival: a new agent, shown the theory, wrote the strongest different mechanism for the same results and a set of tests on which the two disagree; another new agent answered for the theory. Eleven such contests were run, nine in rounds 4 to 6 and two in side runs of the engine. Under one decision rule for all of them (a win needs two tests that separate the pair), the theory lost four, won one, and six were not separated; the one winner's stated cause was refuted by its review. Summed over the 53 tests the theories are 5.1 nats behind, which is close to even. The calibration gap appears again: the rivals' intervals held 37 of 53 and the defenders' 27.

No theory passed the readers, a rival and a review. As scored when each round ran, 23 of the 63 were refuted by a prediction or a review, 4 lost to a rival, 7 gained nothing over the readers, and 28 failed the scores of rounds 1 to 3 or were dropped. One was unrefuted at the end, not separated from its rival.

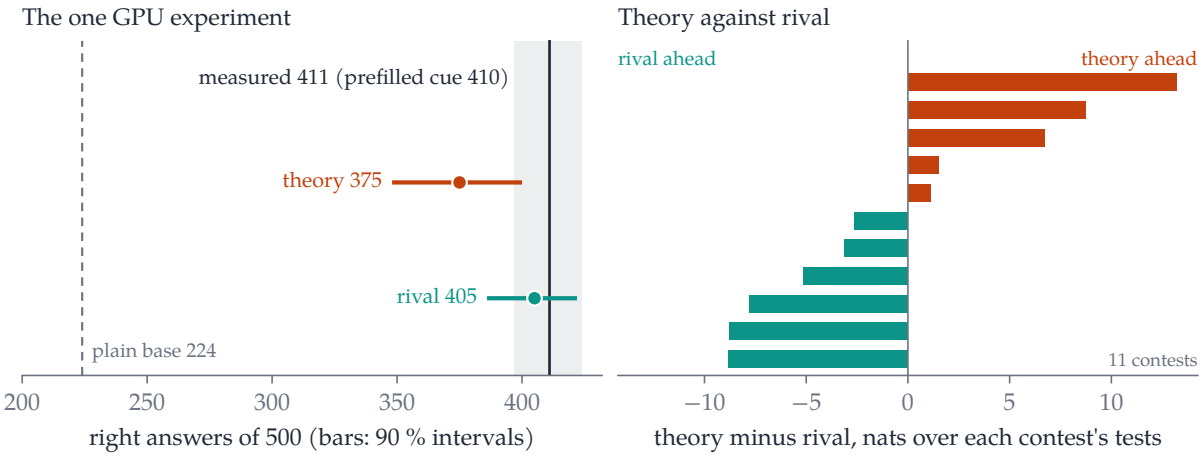



**Figure 3: One experiment, eleven contests. Left:** the base model was sampled until its own first 24 tokens began the way the RL model begins; right answers of 500. The forecast made from the theory, the one made from its rival, and the measurement with its recorded sampling interval (shaded). **Right:** one bar per contest, theory minus rival in nats over its tests.

One contest was settled on a GPU. The engine asks, for each theory, for the one change to the untrained base that the theory implies; a separate job, given the theory, writes the experiment and a forecast of it, and a second job, given only the rival, forecasts the same experiment. One such experiment was run. The theory, from round 5, held that the plain base commits in its first few tokens to one of two whole ways of answering and that nothing after the opening adds. It had by then lost to its rival, and part of what it stands on is published [8]. The experiment: sample the base's own first 24 tokens until they begin as the RL model's answers do, with nothing prefilled, then let it finish. The forecast from the theory was 375 right of 500 $[348, 400]$; from the rival, 405 $[386, 422]$. The experiment gave 411, against 410 for the prefilled cue and 224 for the plain base (Figure 3). On the 403 problems where a resampled opening was used, the answers scored 332, and a control that kept an opening of the other kind scored 137 ($+48.4$ points $[42.9, 53.6]$). The experiment was worth running, and the forecast made from the theory that led to it was wrong.

### 3.5 Faults in our own instrument

A score for theories should fail a theory that is false. We found three ways ours did not.

*A reversed theory passed.* Round 1 planted a theory with every effect reversed. Under that round's score (closer or farther than a reader on 58 numbers) it tied for second of seven. Most of the 58 numbers were ones no theory spoke about, so the count measured noise between runs. Round 2 scored only the author's own claims; the reversed copies of five theories were then nearer on 16 of 67 claims against 36 of 67 for the originals. No reversed control was run from round 4 on, where most of the results above come from.

*Results leaked through commit messages.* In round 2 the agents ran inside the repository, and a commit subject carrying result numbers reached the readers and the new authors. The seven quantities it exposed were removed, and from round 3 the agents ran in empty directories. Without round 2 the theories are nearer in 116 of 243.

*The score changed.* Rounds 1, 2 and 3 to 7 use three different rules, which is why the log gain and the reader intervals exist only for the last five rounds and why we report the nearer count, which every round supports, beside them. An independent recomputation matched 7 of 7 values in round 4 and 57 of 58 in round 6.

## 4 Related work

**Explanations scored by prediction.** The closest study is CHIVE [11], which scores explanations of a language model's behaviour by whether they help an agent predict its counterfactual responses, and finds that interpretability tools give no uplift over reading transcripts. Earlier work tested natural-language explanations of single neurons against activations and interventions and found them wanting [12], and Pando plants a known ground truth to test the methods [13]. We score theories of experimental results, not of one model behaviour, and add interval calibration.

**Language models forecasting research.** Models predict the outcomes of neuroscience studies better than experts [14], and are near chance at ranking pairs of AI research ideas without fine-tuning [15]; neither has a theory arm. Theorizer backtests about 2,900 generated theories against later papers, judged by a language model and compared with a generator that has no literature access [16]; its baseline is another theory. Under-covering 90 % intervals are measured directly in QuantSightBench [9].

**Falsification.** POPPER tests a hypothesis by sequential experiments with error control [17], and Fa and Culjak argue that agentic science should judge a claim by the adversarial experiments it survives [18]. Our rival stage is one such adversary. The co-scientist [2] also sets hypotheses against each other, by debate where ours are set against measurement.

## 5 Limitations

- The design cannot separate writing a mechanism from the other differences between the two prompts. Authors chose the quantities, were asked for predictions a reader would not guess, saw published mechanisms and failed theories, and from round 4 were not told how intervals are scored. A matched test would give one prompt, with and without the request for a mechanism, the same fixed quantities.
- The baseline is not ignorant. A reader sees the results the theorist sees. The claim is not that the mechanisms are false.
- The accuracy result is a failure to find a difference. Its interval, 40 to 55 %, excludes a theory that is nearer two times in three, not a small edge either way.
- One model family in every role, recorded in the run files by alias, not by dated version. The two readers are one prompt run twice, ten reader jobs in rounds 3 to 7, and the bootstrap treats them as fixed. There is no human theorist and no human reader.
- One project, two days, one subject (a 7B model on mathematics problems). The theories explain results that were hours old; a few measured values were later revised, and two scored quantities of round 5 were void.
- Predictions overlap: the 394 comparisons cover 228 distinct quantities. Resampling theories does not

account for two theories that predict the same number.
- Only rounds 4 to 7 (156 comparisons) score quantities whose recipe was fixed before any row was read. In rounds 1 to 3 a model read the measured numbers off board posts.
- The outcomes against rivals depend on the decision rule: as scored when each was run they read six lost, two not separated and three won. The GPU experiment is one experiment.

## 6 Conclusion

Asked to explain a project's results and to stake numbers on the explanation, a language model forecast no better than it did with no explanation, and with narrower intervals. Whether the mechanism or the prompt around it narrows them, the practical reading is the same for a team that works with such agents: an agent's account of why is not evidence for its numbers. Widen the intervals of any forecast that arrives with a mechanism, give the mechanism a rival, and keep the part that did earn its place here, the experiment it points to. Read an agent's theory as a proposal for an experiment.

**Reproducibility.** The round files, prompts and code of the engine are in the project repository, in the folder `studies/theory/`. In its folder `paper/`, the script `build_facts.py` recomputes the numbers of this paper from those files and writes each with its source to `facts.json`, `predictions.jsonl` holds one row per prediction, and `figs.py` draws the figures.

## References


[1] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[2] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al. Accelerating scientific discovery with Co-Scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[3] Joeran Beel, Min-Yen Kan, and Moritz Baumgart. Evaluating Sakana's AI Scientist: Bold Claims, Mixed Results, and a Promising Future? *arXiv preprint arXiv:2502.14297*, 2025.

[4] Ziming Luo, Atoosa Kasirzadeh, and Nihar B. Shah. The More You Automate, the Less You See: Hidden Pitfalls of AI Scientist Systems. *arXiv preprint arXiv:2509.08713*, 2025.

[5] Dhruv Trehan and Paras Chopra. Why LLMs Aren't Scientists Yet: Lessons from Four Autonomous Research Attempts. *arXiv preprint arXiv:2601.03315*, 2026.

[6] Paul E. Meehl. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press, 1954.

[7] Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

[8] Sophie L. Wang, Amil Dravid, Rulin Shao, Kevin Farhat, Sewon Min, and Alexei A. Efros. Base Models Can Reason By Taking a Cue From Training Data. *arXiv preprint arXiv:2610.06851*, 2026.

[9] Jeremy Qin and Maksym Andriushchenko. QuantSightBench: Evaluating LLM Quantitative Forecasting with Prediction Intervals. *arXiv preprint arXiv:2604.15859*, 2026.

[10] Philip E. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, 2005.

[11] Adam Karvonen, Euan Ong, Subhash Kantamneni, and Samuel Marks. Would this change your answer? Evaluating Explanations of LLM Behavior In The Wild with Counterfactual Experiments. *arXiv preprint arXiv:2608.16747*, 2026.

[12] Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. Rigorously Assessing Natural Language Explanations of Neurons. *arXiv preprint arXiv:2309.10312*, 2023.

[13] Ziqian Zhong, Aashiq Muhamed, Mona T. Diab, Virginia Smith, and Aditi Raghunathan. Pando: Do Interpretability Methods Work When Models Won't Explain Themselves? *arXiv preprint arXiv:2604.11061*, 2026.

[14] Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, et al. Large language models surpass human experts in predicting neuroscience results. *arXiv preprint arXiv:2403.03230*, 2024.

[15] Jiaxin Wen, Chenglei Si, Yueh-han Chen, He He, and Shi Feng. Predicting Empirical AI Research Outcomes with Language Models. *arXiv preprint arXiv:2506.00794*, 2025.

[16] Peter Jansen, Peter Clark, Doug Downey, and Daniel S. Weld. Generating Literature-Driven Scientific Theories at Scale. *arXiv preprint arXiv:2601.16282*, 2026.

[17] Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candès, and Jure Leskovec. Automated Hypothesis Validation with Agentic Sequential Falsifications. *arXiv preprint arXiv:2502.09858*, 2025.

[18] Dionizije Fa and Marko Culjak. Sound Agentic Science Requires Adversarial Experiments. *arXiv preprint arXiv:2604.22080*, 2026.