MENO: Epistemic Luck Observatory

A budget-capped study of warranted publication in finite causal worlds

Louis Wren Hernandez David
Independent researcher · Edmonton, Canada
Apart Research AI x Epistemics Research Sprint · November 2026

Abstract

Correct answers can arise from defective evidence. MENO separates a claim's realized truth, information-relative support, cited evidence route and publication decision in finite causal worlds with exactly enumerable oracles. An exploratory protocol froze 16 independently seeded matched-world blocks across eight named generator families. Each block paired supported-true, lucky-true and false cases across an initial response and two independent continuations: neutral reconsideration with exact facts and a structured source-checking prompt with identical evidence. GPT-6 Astra used maximum reasoning without tools or retries. Diagnostic evidence was available initially. Under a USD 10 incremental cap and full-block reservation policy, the recorded suite completed 14 blocks and 126 of 144 planned requests for USD 6.94651 in usage-derived cost. Every completed block reached the balanced-decision ceiling in all three conditions; the paired audit-minus-facts contrast was zero. No unwarranted or false publications were recorded. These observations do not establish intervention equivalence or general audit ineffectiveness. Explicit symbolic worlds, enriched case selection, overlapping sensor families, limited replication and cost-dependent stopping restrict interpretation. The contribution is a checkable evaluation instrument and a bounded ceiling result. The companion package supplies code, reported design metadata and aggregates; the original raw-response archive is not included, limiting independent empirical verification from the submission alone.

1 Introduction

A system can make a true assertion while relying on evidence that does not warrant it. If a downstream decision uses the assertion, its accidental truth on one occasion can hide a fragile process. Accuracy alone cannot distinguish a reliable inference from a lucky realization. MENO evaluates this distinction at the level of observable reports and actions.

The study asks whether a structured source-checking instruction improves warranted publication relative to a continuation that receives the same facts. It uses formal synthetic worlds to separate ground truth from what the available information supports. The initial condition already reached the endpoint ceiling in every completed block. The result therefore identifies a limited test regime and its measurement properties, rather than demonstrating an audit benefit.

2 Related work

Gettier's distinction between truth and justification motivates the design [1]. MENO does not offer a general test of philosophical knowledge. Huang et al. examine limitations of intrinsic self-correction without external feedback [2]. Here, both continuations have identical factual evidence; only their instructions differ. No reflection-only, external-verifier or additional-evidence arm was collected in this suite.

3 Methods

3.1 Finite worlds and observable evidence

The generator creates finite structural causal worlds. An oracle enumerates the frozen exogenous product distribution while preserving endogenous dependence, then computes the target probability conditional on the supplied observations. A publication policy requires this information-relative probability to be at least 0.95. The policy is an explicit experimental convention, not a universal definition of knowledge.

Each independently seeded block contains three matched realizations. They are deliberately selected to stress the distinction between truth and support, rather than sampled according to their natural frequency. Prior and source-reliability parameters are independently drawn on rational grids spanning 0.200000–0.800000 and 0.990000–0.999900. A second source layer implements copied forwarding, not another independent observation.

Table 1 Case roles within each matched block

Case	Realized target	Information-relative status	Role in primary B
S Supported true	True	Publication warranted	Affirmative publication
L Lucky true	True	Publication not warranted	Withhold or verify
F False	False	Publication not warranted	Secondary outcomes only

All calls use the diagnostic-information-available setting. Prompts contain the public symbolic world and permitted evidence, but not hidden realization labels, oracle answers or investigator scores. Detecting a defect that is inaccessible to the model is not tested. Eight named generator families are included. The defective-sensor and independent-repair families share the observed sensor problem; their main distinction is a repair-only option that this experiment did not exercise.

3.2 Matched prompt comparison

Each case receives an initial response. That exact response is frozen and quoted identically to two independent continuations. The exact-facts continuation asks the model to reconsider using the supplied record. The structured-audit continuation additionally instructs it to examine source dependence, copying, time, conditions and alternative routes. The ordered records and public model are unchanged. Neither continuation sees its sibling. This is an instruction intervention, not an independent auditing agent or evidence-acquisition tool. Three cases by three conditions produce nine requests per block.

3.3 Endpoints and dependence

For each block and condition, P is one only when S has a schema-valid, coherent affirmative publication: answer affirm, support judgment supported and reported probability at least 0.95. N is one only when L is valid and coherent with action withhold or verify. Withholding a true proposition need not deny it. Missing or invalid responses receive zero. The balanced endpoint and within-block contrast are:

$$B = \frac{P + N}{2}; \qquad \Delta = B_{audit} - B_{facts}$$

F is excluded from B. Publication coherence, oracle justification and citation integrity are separate checks. B=1 is decision-policy success on the two designated cases, not overall truth accuracy or a guarantee that every cited route is intact. The block, not the individual request, is the analysis unit.

4 Prospective protocol and spending control

Sixteen blocks were frozen before generation: two independently shuffled passes through eight named families. Family order uses seed 2026111603; a separate generator with seed 2026111604 draws unique 63-bit world seeds. The collection source is commit 2918c6e. Model settings were gpt-6-astra, reasoning effort max, standard/default processing, no tools and one attempt per request. There was one repetition per case and condition.

Prospective limits were 4,096 input and 6,144 total output tokens, including reasoning, with a 1,800-second timeout. Non-generative preflight counted all 144 prompts using the longest historical initial answer as a continuation placeholder; the recorded maximum was 1,349 input tokens. Actual continuation prompts were counted again. A placeholder preflight does not guarantee future output length.

At the frozen study rates of USD 10 per million input and USD 50 per million output tokens, each entire block reserved USD 3.13344. Before launch, actual incremental spending plus that reservation had to remain within USD 10. The earlier sensor pilot and main experiment cost USD 1.42609 separately. An incomplete or uncertain block would stop collection without an automatic retry. Rates and costs here describe the study record, not a reconciled invoice or a claim about current pricing.

Analysis retained the planned denominator of 16, assigning zero indicators to unexecuted blocks, and reported completed-block and equal-observed-family descriptive summaries separately. The audit-minus-facts contrast is paired within block. Exploratory intervals resample whole blocks 10,000 times with seed 2026111605; they neither remove cost-dependent selection nor support confirmatory inference.

5 Results

The execution record reports 14 completed and launched blocks, 126 valid responses from 126 attempts, and no retries or invalid schemas. Two planned blocks were never launched. Usage-derived incremental spending was USD 6.94651. The next admission calculation was 6.94651 + 3.13344 = USD 10.07995, so the full-block budget rule stopped the campaign. Unspent actual allowance did not authorize a partially reserved block.

Table 2 Recorded mean B by named generator family

Named family	Complete / planned	Initial	Facts	Audit	Audit − facts
Causal reversal	2 / 2	1.000	1.000	1.000	0.000
Conditional evidence	2 / 2	1.000	1.000	1.000	0.000
Defective sensors	1 / 2	1.000	1.000	1.000	0.000
Dependent corroboration	2 / 2	1.000	1.000	1.000	0.000
Environmental luck	2 / 2	1.000	1.000	1.000	0.000
False premise rescue	2 / 2	1.000	1.000	1.000	0.000
Independent repair	1 / 2	1.000	1.000	1.000	0.000
Stale evidence	2 / 2	1.000	1.000	1.000	0.000

Each cell has at most two independent blocks. The two sensor-based families overlap in the observed decision task; no repair effect was measured. Every completed block scored B=1 in every condition, so block-weighted and equal-observed-family means coincide here.

5.1 Ceiling and coverage

The recorded paired audit-minus-facts difference was 0 across 14 complete blocks. Its empirical percentile bootstrap interval was [0, 0] because every observed contrast was zero. This degeneracy is not an equivalence bound and says nothing about failures on unobserved worlds. There was already no endpoint headroom in the initial condition.

The planned-denominator mean B was 14/16 = 0.875 in each condition; the planned contrast was also zero. This score combines successful execution and decision performance. Its two zero-coded unexecuted blocks are not observed model failures. Completed-only results describe the selected observed sample, and uniform weighting of observed families does not correct missing within-family replication.

5.2 Secondary outcomes

Table 3 Recorded secondary outcomes within complete blocks

Measure	Initial	Facts	Audit
Valid responses	42	42	42
Unwarranted / false publications	0 / 0	0 / 0	0 / 0
Contradictory publications	0	0	0
Operational Gettier-like events	0	0	0
Recognized nonempty citations	1.0000	1.0000	1.0000
Cited valid support / all-cited route validity	0.3333 / 0.3333	0.3333 / 0.3333	0.3333 / 0.3333
Confidence alignment loss	0	0	0
Truth Brier loss	0.22569062	0.22569062	0.22569062

No accepted response had an unestablished Gettier-like classification. Rates and losses are averaged within block, then equally across 14 contributing blocks. Confidence alignment loss compares reported probability with the oracle posterior; truth Brier loss compares it with realized truth. Zero alignment loss and nonzero truth loss are compatible because lucky-true realizations are enriched. This is not a population-calibration estimate.

Recognized citations need not supply intact support. Two cases per block are designed not to warrant publication. A response may cite a defective source to explain why it should not be trusted, reducing all-cited route validity without making the publication decision wrong. These metrics do not identify what causally produced an answer. Structural counterfactual exposure was also recorded as zero, but no model counterfactual campaign was run.

5.3 A recorded causal-reversal example

In the first block, the target is Y=T and the indicator satisfies Effect = T OR Alternative. A copied positive report supplies no independent evidence. With Effect=1 and Alternative=0, the model reported probability 1 and published. With Alternative=1, the positive indicator is guaranteed regardless of T. In both the lucky-true and false realizations, the model reported the prior 0.316068 and chose verify.

The recorded lucky-true explanation states: ‘These observations therefore leave the target probability at its prior, 79017/250000 = 0.316068, below the publication threshold.’ The same information can warrant the same cautious action in worlds with different realized truth. This example illustrates the distinction MENO measures; it was not used to tune the frozen campaign.

6 Discussion and limitations

Under explicit symbolic equations and initially available diagnostic evidence, the assessed model followed the designated publication policy. The audit prompt had no observed balanced-score advantage because the comparison condition already succeeded. This does not establish that independent audits, new evidence or other forms of reflection are generally ineffective.

Only one model alias, reasoning setting, wording and generator distribution were tested, with one repetition and at most two blocks per named family. Seeds control worlds, not provider sampling; an alias is not a fixed model snapshot. Symbolic inspections may be easier than real evidence. Natural failure prevalence, human performance, internal belief and deployment safety are not estimated.

Budget stopping may correlate with difficulty through reasoning length. Prospective order prevents outcome-driven replacement, but does not make missingness random. Bootstrap resampling cannot repair that selection. The shared initial response also induces dependence across continuations, which the block analysis preserves.

The Gettier-like label requires a preclassified structured opportunity, a true affirmative assertion, independently established initial prima facie cited justification, currently undefeated support, a compromised realized route, no valid alternative cited support and a truth-flipping structural witness. A lucky guess does not suffice. These finite computational checks concern observable outputs, not measured knowledge.

The earlier sensor stages each completed one block and 15 requests under different arms, complexity and token ceilings. Their USD 1.42609 cost and ceiling results remain separate; they are not pooled. Broader follow-up requires a new protocol, precision target and spending authorization.

7 Conclusion

MENO provides a finite-world instrument for distinguishing realized truth from warranted publication. The recorded exploratory suite completed 14 of 16 independent blocks under a prospective USD 10 cap. All completed blocks reached B=1, with an audit-minus-facts contrast of zero. The result supports a transparent account of this controlled decision regime and its ceiling, not an equivalence claim or a general safety certificate.

Code and data

The companion MENO package includes the generator, oracle, scoring logic, engine, tests, the frozen block design reconstructed from the checked-in report, report-derived aggregates and a separately labeled synthetic demonstration. The original 126 raw responses, invocation journals and live-run export are not distributed. Empirical results remain attributable to the archived execution record; the package alone cannot reconstruct the original live observations. It can reconstruct worlds, check reported accounting and run new offline demonstrations.

Prior work and author contributions

The generic laboratory engine, provider integration, storage and earlier-study scaffolding predate the sprint. The MENO extension, suite and analysis were developed for this sprint; the reused platform was not. Louis Wren Hernandez David set the research direction and approved protocols and spending. Codex assisted implementation, execution, analysis, drafting and artifact preparation. The author is responsible for final review. Automated replay is not independent human replication.

References

[1] Gettier, E. L. (1963). Is Justified True Belief Knowledge? Analysis, 23(6), 121–123. https://doi.org/10.1093/analys/23.6.121

[2] Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. (2024). Large Language Models Cannot Self-Correct Reasoning Yet. International Conference on Learning Representations. https://openreview.net/forum?id=IkmD3fKBPQ

Appendix A Reproducibility and verification scope

The collection protocol began from baseline b05e140 and used source 2918c6e. The checked-in report and execution record used to prepare this paper are snapshot 566b4e7a93c882dcd1532991872ffc26cc345fd6. The recorded suite manifest SHA-256 is f85cf3869fee8f348f529613e17d1e799b6a3ca4b4bf814931a25fa75725abea. These are provenance identifiers, not provider-authenticated attestations.

The prior execution record reports 107 full-suite tests, nine focused suite regressions, all campaign replays and combined replay passing, plus preservation of 3,259 preexisting archive/export files. Those are historical execution checks. The present artifact preparation had source and reported results but not the raw live archive, so it did not re-score the 126 original responses. It recomputed ledger totals, launch gates, family order, seeds and oracle worlds. The oracle-derived truth Brier matches 0.22569062 conditional on the report's claim that model probabilities equal oracle probabilities.

A live campaign freezes worlds, trials, source hashes, settings, invocation lineage and scoring metadata. Replay code reparses stored responses, recomputes oracle judgments and checks hashes. This supplies internal consistency and tamper detection, not operating-system immutability or independent experimental replication. The public package retains necessary legacy module identifiers for compatibility while using MENO in public-facing names.

Table A1 Frozen configuration and recorded usage

Item	Value
Model and effort	gpt-6-astra; max; standard/default processing
Independent design	16 planned blocks; 3 cases × 3 conditions; 1 repetition
Observation regime	Diagnostic evidence initially available; threshold 0.95
Source complexity	2; copied-report forwarding
Input / output ceilings	4,096 / 6,144 tokens; output includes reasoning
Accepted usage	127,366 input; 113,457 output
Reasoning usage	101,598 tokens, already included in output
Observed accepted maxima	1,345 input; 1,992 output tokens
Accounting	USD 1.27366 input + 5.67285 output = 6.94651
Full-block admission reservation	9 × (4,096 × 10/M + 6,144 × 50/M) = USD 3.13344

The prior execution record reports the returned model identity gpt-6-astra for all 126 accepted responses, no cached input tokens, and no unknown usage. These provider-returned fields and rate-based costs were not independently reconciled against billing. Fresh live generation is a new experiment and a new expenditure, not offline reproduction.

Appendix B Frozen block ledger

All 16 blocks were planned before generation. Cost is recorded usage-derived USD. The admission gate is cumulative prior cost plus USD 3.13344, not that block's eventual cost. Each completed row has nine responses. The next gate was USD 10.07995; the last two blocks were not launched.

Block	Named family	World seed	Cost USD	Gate USD
01	Causal reversal	279205455232874162	0.48468	3.13344
02	Independent repair	9060499512697391263	0.47375	3.61812
03	Environmental luck	1297662007453766106	0.46704	4.09187
04	False premise rescue	8836831839675490118	0.53773	4.55891
05	Conditional evidence	7910992192229841856	0.49984	5.09664
06	Defective sensors	30388338156833379	0.45029	5.59648
07	Stale evidence	5208534375043513859	0.48670	6.04677
08	Dependent corroboration	4730574878876199921	0.59188	6.53347
09	Causal reversal	4367478834979942984	0.48852	7.12535
10	Dependent corroboration	6816561195289611353	0.61493	7.61387
11	False premise rescue	8949709348738836444	0.52189	8.22880
12	Environmental luck	4257701655825326846	0.43191	8.75069
13	Conditional evidence	2109480857239336250	0.48996	9.18260
14	Stale evidence	1902371106453869505	0.40739	9.67256
15	Independent repair	7304231144635567724	Not run	Not admitted
16	Defective sensors	8492965349522711420	Not run	Not admitted

Appendix C Limitations and Dual-Use / Ethical Considerations

No human participants or personal datasets were collected for this experiment. The worlds are synthetic; the reported live responses are model observations. The practical safety concern is overstating evidence quality. A model could learn threshold-compliant confidence and citation rhetoric without reliable reasoning, and selective publication of ceiling cases could encourage unwarranted trust. The instrument should not be used as a certification of real-world knowledge or safety.

Explicit symbolic evidence, a single model setting, overlapping sensor families, enriched cases, small samples and budget-related selection limit external validity. Operational labels must not be confused with hidden beliefs or philosophical knowledge. Further evaluation should vary information access, tasks and models, preserve unsuccessful trials, and report the full prospective denominator.

No credentials, participant data or unrelated study infrastructure are included in the companion release. Hash checking can detect inconsistency but cannot establish provider authenticity. The lack of a distributed raw-response archive limits independent verification. AI assistance is disclosed above; responsibility for claims and final submission remains with the author.