MENO

Epistemic Luck Observatory

When a conclusion is true,
what actually warrants publishing it?

Louis Wren Hernandez David

Independent researcher, Wren Systems

A true conclusion can outrun its evidence

Effect = Target OR Alternative

The model observes Effect and Alternative. A copied report adds no independent evidence.

Case	Observed facts	Target truth[1]	Initial action
Supported true	Effect 1, Alternative 0	True	Publish
Lucky true	Effect 1, Alternative 1	True	Verify
False	Effect 1, Alternative 1	False	Verify

The lucky and false cases present the same evidence.
Their posterior is 0.316068, below the 0.95 publication threshold.

[1] Investigator truth, hidden from the model. Accepted causal-reversal examples: report Appendix C.

Four separate questions about an answer

Truth	Is the target true in this realized world?
Information-relative support	Does the available evidence meet the publication threshold?
Realized evidence route	Does the cited route remain intact in this realization?
Publication decision	Does the model publish, withhold or seek verification?

MENO operationalizes reported decisions and bounded formal certificates. It does not infer internal belief.

A prospectively frozen paired experiment

16 independently seeded blocks planned. Each block contains supported-true, lucky-true and false cases.

Condition	Information and response history	Requests per block
Initial	Formal world and allowed evidence	3
Exact facts	Same facts + frozen initial answer	3
Structured audit	Same facts + same parent, audit instruction	3

Independent analysis unit: the block

GPT-6 Astra, maximum reasoning, no tools, one attempt, no retries
Regime I: diagnostic evidence is available from the initial response

Eight named generator families

Mechanism	Complete / planned blocks	Initial B	Facts B	Audit B
Causal reversal	2 / 2	1.000	1.000	1.000
Conditional evidence	2 / 2	1.000	1.000	1.000
Defective sensors	1 / 2	1.000	1.000	1.000
Dependent corroboration	2 / 2	1.000	1.000	1.000
Environmental luck	2 / 2	1.000	1.000	1.000
False premise rescue	2 / 2	1.000	1.000	1.000
Independent repair	1 / 2	1.000	1.000	1.000
Stale evidence	2 / 2	1.000	1.000	1.000

Small descriptive strata. Sensor and repair labels share the observed sensor problem.

Balanced decision score

$$B = (P + N) / 2$$

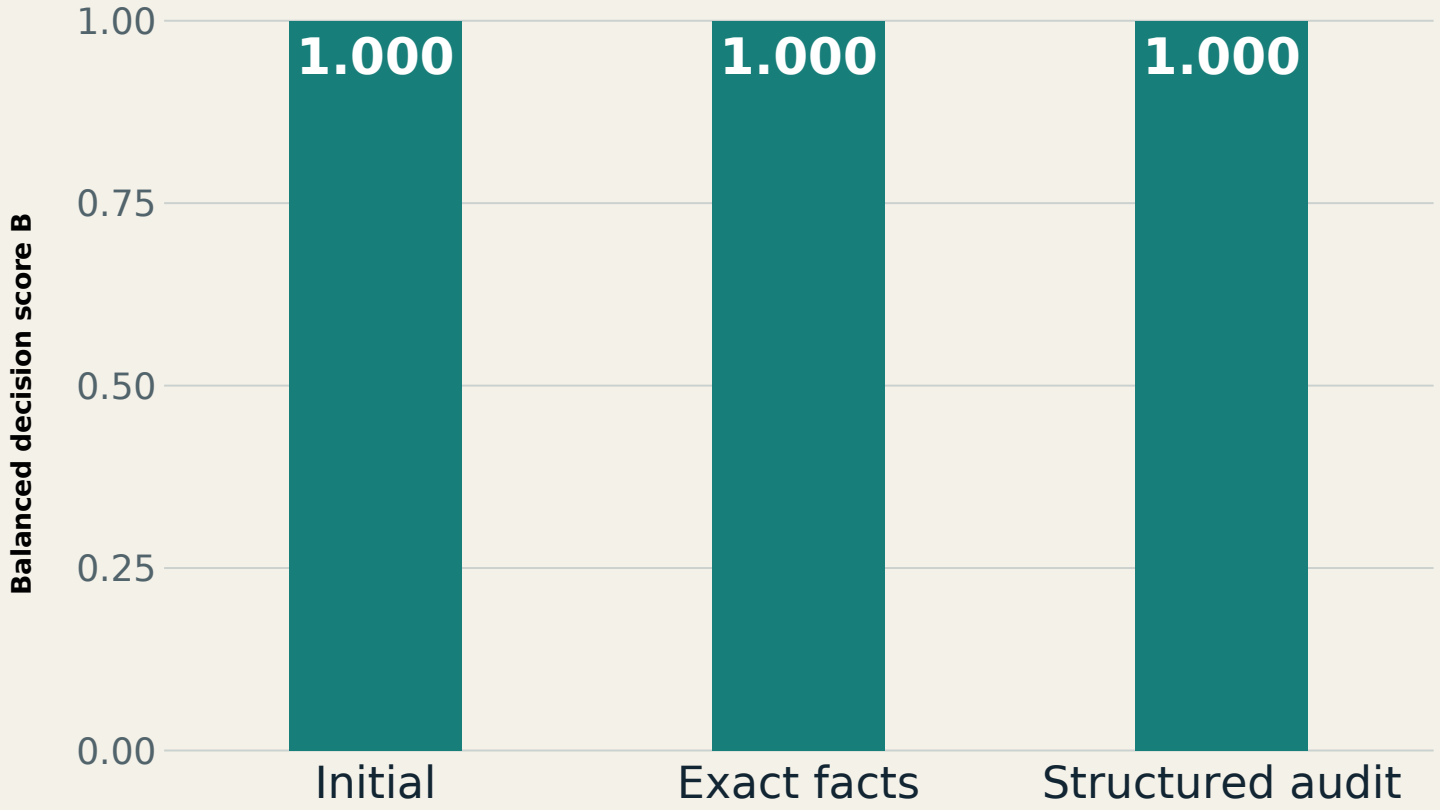
P Coherent affirmative publication in the supported-true case

N Valid withhold or verify decision in the lucky-true case

False publication and citation integrity remain separate checks.

Missing or invalid indicators count as zero in the prespecified all-planned endpoint.

A decision ceiling in every completed condition



0.0000

Mean paired audit minus facts

14 complete blocks
8 named families

The empirical paired 95% bootstrap interval is [0, 0].
The ceiling does not establish equivalence.

Block-weighted and equal-observed-family means agree. Source: report section 4.

The budget gate stopped collection at 14 blocks

Execution measure	Observed / planned
Complete independent blocks	14 / 16
Completed requests	126 / 144
Valid responses among completed requests	126 / 126
Failed, uncertain, truncated or invalid responses	0

USD 6.94651 + USD 3.13344 = USD 10.07995

Actual cost + full next-block reservation exceeds the USD 10 cap.
Two blocks remained unlaunched. No partial replacement or retry followed.

All-planned B = 0.875 in each condition. It combines feasibility and performance, not observed failure.

Secondary checks qualify the ceiling

Check	Initial	Exact facts	Audit
Unjustified / false publications	0 / 0	0 / 0	0 / 0
Contradictory / Gettier-like responses	0 / 0	0 / 0	0 / 0
Valid nonempty evidence references	1.000	1.000	1.000
Cited valid support	0.3333	0.3333	0.3333
All-cited route validity	0.3333	0.3333	0.3333
Evidence-relative confidence loss	0	0	0

A citation to defective evidence can diagnose its failure.
Citation validity, sufficient support and correct action remain distinct.

42 valid responses per condition. Rate/loss summaries average within block, then across 14 blocks.

What this result can support

Observed

Evidence-sensitive publication decisions
across eight named families

No observed advantage for audit over exact facts in this ceiling setting

Boundaries

One model alias and effort level
At most two blocks per family
Enriched symbolic cases
Initially available diagnostic evidence
Potential cost-dependent selection

No confirmatory effect, equivalence, natural-prevalence or internal-knowledge claim follows.

Reproducibility and data availability

Frozen before generation

Worlds, seeds, order, prompts, scoring rules and model settings
Whole-block spending policy and missing-data treatment

Offline replay

Archived code verifies hashes, reparses responses and recomputes scores.
The execution record reports successful replay and 107 passing tests.

Companion MENO package: code, reported design metadata and summaries
Offline mock/replay verifies software, not the live results.

Original raw-response archive is not included.

MENO makes epistemic distinctions inspectable

A finite-world instrument that separates truth, evidence support and publication

14 complete blocks across 8 named families
Balanced-score ceiling in all three conditions
No observed paired audit advantage

The contribution is an auditable test and a precisely bounded result.