A Confirmatory Test of Kleos Framing in Language-Model Reasoning

A preregistered equivalence result

Abraham
September 2026

CONFIRMATORY RESULT

No effect as large as +/-5 percentage points on this design

Abstract

Can a language model solve difficult reasoning problems more accurately when excellence is framed as a source of enduring renown? A pilot suggested a +6.7 percentage-point effect. A preregistered confirmation used 866 paired procedural problems and 3,464 total observations. The confirmatory estimate was -0.64 points (bootstrap 95% interval: -3.18 to +1.85; paired sign-flip p = 0.650). TOST equivalence passed at the preregistered +/-5-point margin. The bounded conclusion is no kleos effect as large as 5 points for the tested checkpoint, task family, and prompt construction. Yet 40.1% of items changed outcome across conditions, showing substantial churn that canceled in aggregate.

1. Research question

Kleos is the ancient Greek idea of enduring renown: excellence that will be remembered and spoken about. This study asks an operational behavioral question, not a psychological one:

> *Does framing excellence as the model's enduring renown change objective accuracy beyond otherwise identical social attribution?*

The control said that response quality would be remembered and spoken of as the model's. The treatment changed only the stake: excellence would be remembered and spoken of as the model's renown or fame. Both requested step-by-step reasoning and the same final-answer format.

The design does not test whether a model experiences motivation, desires recognition, possesses a stable identity, or understands kleos as a person would. It tests whether a specific framing reliably changes exact-answer accuracy.

2. Why confirmation was necessary

The preceding pilot used 150 paired items and estimated a +6.67-point effect. Its intervals barely excluded zero and its paired sign-flip p-value was 0.0501. The pilot preregistration prohibited treating that estimate as confirmation or using it to size the next study.

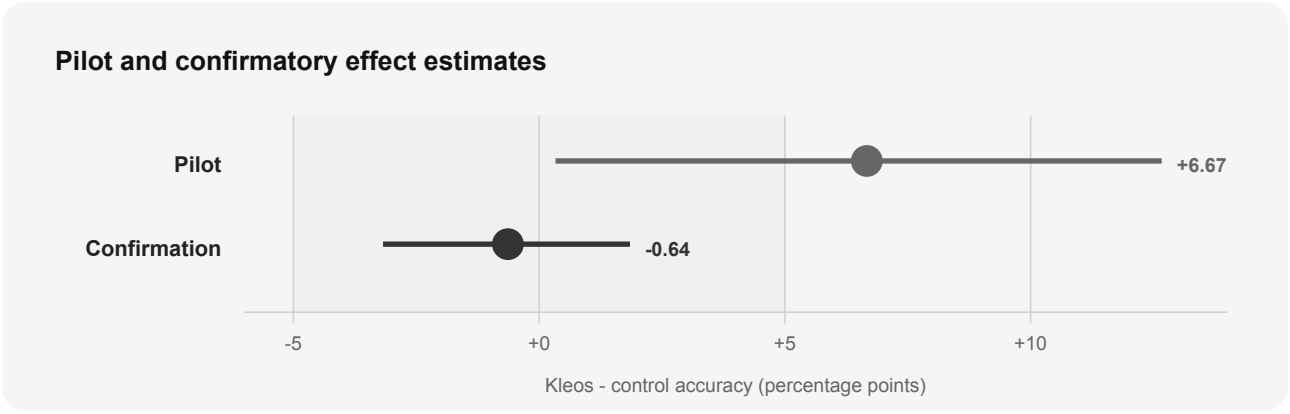
Related prompting work finds that wording changes can rearrange which questions a model answers correctly without producing a predictable aggregate benefit. Procedural generation was used to provide held-out problem instances and reduce dependence on static benchmark contamination.

3. Confirmatory design

Design element	Frozen choice
Scope	One frozen quantized Llama 3.1 checkpoint
Tasks	866 held-out procedural multi-step word problems
Conditions	Spoken attribution control vs. kleos treatment
Sampling	Two responses per item-condition cell
Observations	3,464 total; task item is the experimental unit
Outcome	Exact integer accuracy; unusable responses score zero
Primary test	Paired sign-flip test, two-sided alpha 0.05
Meaningful margin	+/-5 absolute percentage points
Stopping	Fixed sample; no interim analysis or outcome-driven reruns

The sample size targeted 90% power for a 5-point effect using a pilot-independent calibration variance. Ten preregistered shards ran under one immutable plan. Every admitted shard passed audit before pooling.

4. Primary result

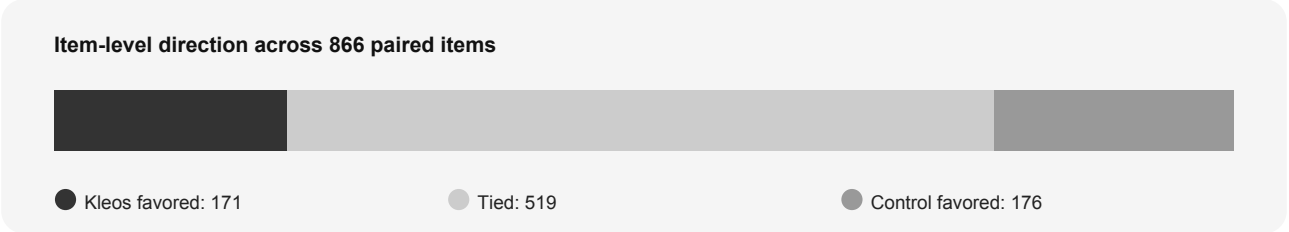


Quantity	Confirmatory result
Control accuracy	68.82%
Kleos accuracy	68.19%
Paired difference	-0.64 points
Bootstrap 95% interval	[-3.18, +1.85] points
Paired sign-flip p	0.650
Realized 80%-power MDE	3.59 points
TOST at +/-5 points	Equivalent

The primary test was not significant. The frozen decision rule therefore deferred to TOST. Both one-sided tests passed, and the realized minimum detectable effect was smaller than the equivalence margin. This is preregistered

outcome 3: no kleos effect as large as 5 percentage points on this design.

5. Item-level movement



The aggregate null was not produced by identical behavior everywhere. The item flip rate was 40.1%, the mean absolute paired difference was 22.8 points, and cross-condition score correlation was 0.467. Positive and negative movements balanced.

6. Validity and response length

All 3,464 scheduled observations were present and the validity gate passed. There were 36 unusable responses (1.04%): 21 under kleos and 15 under control. The arm-rate spread was 0.35 points, below the preregistered 2-point cap. Unusable responses remained in the intention-to-treat score as incorrect.

Mean output length was 227.1 tokens under kleos and 221.9 under control; medians were 208 and 207. These small descriptive differences do not establish an effort mechanism, particularly because the accuracy effect was equivalent to zero within the meaningfulness margin.

7. Why the reversal matters

The confirmation contradicts the pilot's positive estimate. Under the frozen decision rule, the confirmatory study governs. The most likely interpretation is ordinary sampling variation, an explanation anticipated before the run.

This is useful evidence. It narrows a plausible prompting effect and demonstrates a workflow capable of overturning its own exciting result. The study was powered from independent calibration variance, retained a two-sided test, used held-out tasks, and prohibited outcome-driven repair.

8. What the result licenses

For the tested frozen Llama 3.1 checkpoint, procedural multi-step word-problem family, and prompt construction, the study found no kleos-framing effect as large as 5 absolute percentage points.

What it does not license

- It does not prove that the exact effect is zero.
- It does not establish that kleos framing never affects any model or task.
- It does not generalize to other checkpoints, model families, languages, or interactive agents.
- It does not show that models lack motivation, desire, identity, or subjective experience.
- It does not explain the substantial item-level churn.
- It does not justify testing new phrasings repeatedly until one becomes positive.

9. Limitations and next tests

The experiment used one quantized checkpoint and one procedural reasoning family. The manipulation was deliberately narrow, which improves causal interpretation but limits generalization. A cross-model replication would be most valuable if motivated by a specific mechanistic prediction rather than dissatisfaction with the null.

The item-level churn deserves separate study. Future work could test whether prompt effects are predictable from task structure, whether movements replicate by item across checkpoints, and whether socially generated training data produces more stable effects than prompt-time framing.

10. Reproducibility statement

The study used a frozen manifest and preregistration, held-out procedural generation, deterministic pairing, immutable sharded execution, append-only invalidation records, shard audit, sealed pooling, and bit-exact summary regeneration. A public release bundle is deferred pending a separate anonymity and licensing review. No result was selected, repaired, extended, or reseeded after outcomes were observed.

References

1. Meincke, L., Mollick, E., Mollick, L., and Shapiro, D. (2025). *Prompting Science Report 3: I'll pay you or I'll kill you - but will you care?* arXiv:2508.00614. arxiv.org/abs/2508.00614

2. Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. (2025). *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models.* arXiv:2410.05229. arxiv.org/abs/2410.05229

3. Miller, E. (2024). *Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations.* arXiv:2411.00640. arxiv.org/abs/2411.00640

4. Bowyer, S., Aitchison, L., and Ivanova, D. R. (2025). *Position: Don't Use the CLT in LLM Evals With Fewer Than a Few Hundred Datapoints.* ICML 2025; arXiv:2503.01747. arxiv.org/abs/2503.01747

Suggested citation

Abraham. (2026). *A Confirmatory Test of Kleos Framing in Language-Model Reasoning: A Preregistered Equivalence Result.*