The Asymmetric Burden of Proof

LLMs Show a Null-Result Asymmetry in a Matched-Vignette Benchmark

Rolando Bosch — Hermes Labs, San Francisco, CA

February 2026

Abstract

We report matched-pair experiments testing whether large language models apply symmetric evidential standards to positive and null scientific claims. Three models (GPT-4o, GPT-5.2 Thinking, Claude Haiku 4.5) evaluated fictional scientific vignettes in which evidence quality was held constant while only the conclusion direction was reversed. Across all six model-format conditions, models allocated less conclusion-consistent probability mass to null claims than to matched positive claims (gaps of 19.6–56.7 percentage points; all bootstrap 95% CIs excluding zero). The asymmetry was directionally consistent in 23 of 24 pair-condition cells and persisted even when discrete classification labels collapsed entirely, surfacing through probability allocation rather than categorical commitment. We characterize this as an asymmetric burden of proof: models treat non-detection as more provisional than matched detection claims, with implications for evidence synthesis, safety assessment, and decision-support pipelines that rely on LLM-generated confidence scores.

1 Introduction

A well-designed study that finds no effect is not a failure; it is a finding. When a randomized controlled trial with 95% power and pre-registered endpoints concludes that a compound does not meaningfully alter an outcome, the resulting null claim carries strong evidential weight. The study was built to detect the effect if it existed, it did not detect it, and the confidence interval constrains the plausible magnitude of any hidden effect. In the language of equivalence testing (Lakens et al., 2018), such a result can constitute positive evidence of absence rather than mere absence of evidence.

Yet null results have long occupied a disadvantaged position in scientific communication. They are harder to publish (Rosenthal, 1979), more likely to be relegated to file drawers, and more readily dismissed as uninformative. Ioannidis (2005) argued that this asymmetry distorts the published literature; Cohen (1994) demonstrated how the ritual of null hypothesis significance testing systematically conflates non-significance with non-existence. These are well-documented human tendencies. The question we raise here is whether large language models, increasingly deployed for evidence synthesis, risk assessment, and decision support, reproduce the same asymmetry.

This is not a question about hallucination or factual accuracy. The concern is directional: holding evidence quality constant, does the model's evaluation shift when the conclusion changes from presence to absence? If so, the implications are practical and immediate. Automated literature

reviewers that systematically discount null findings may amplify publication bias rather than correct it. Safety communication systems that add unsolicited caveats to clean-bill-of-health conclusions could induce unnecessary alarm. Decision-support pipelines that filter on confidence thresholds may quietly exclude high-quality negative evidence from consideration.

Recent work on LLM evaluation behavior has documented related phenomena. Zheng et al. (2023) identified systematic biases in LLM-as-judge evaluations, including position bias and verbosity bias. Pangakis et al. (2023) found inconsistencies in LLM annotation reliability across task types. Briggs, Mellon, and Arel-Bundock (2026) document publication-stage significance selection patterns, providing context for potential downstream asymmetries in model-mediated evidence workflows. A related line of research has examined sycophancy in language models—the tendency to agree with a user's stated position. Perez et al. (2022) demonstrated that RLHF-trained models systematically produce outputs matching user opinions, and Sharma et al. (2023) characterized multiple forms of sycophantic behavior across model families. Our setting differs from sycophancy in a critical respect: there is no user position for the model to agree with. The prompts present evidence and ask for evaluation; the asymmetry emerges from the evidence direction itself, not from social pressure to agree. Studies of positivity bias in language model outputs suggest models may systematically favor affirmative over negative conclusions. Our work extends this literature to a specific and consequential domain: the evaluation of scientific evidence, where the positive–null asymmetry has direct implications for downstream decision-making.

We propose and test the following thesis:

> Thesis (The Asymmetric Burden of Proof). When large language models evaluate scientific claims, they apply a higher evidential standard to null claims than to positive claims of matched quality. Specifically, they treat detection events as sufficient for high-confidence endorsement but treat non-detection as provisional and in need of further corroboration, even when both claims rest on equivalent experimental designs. This asymmetry persists across model families and response formats, though its surface expression varies: in older models it appears in both classification labels and probability distributions; in newer models, label collapse masks the asymmetry, which persists in probability allocation.

We deliberately frame our finding as an asymmetric burden of proof rather than an irrational bias. As we discuss in Section 6, there are principled reasons for a Bayesian agent to assign modestly lower credence to null claims in many real-world contexts. The question is not whether any asymmetry is defensible, but whether the observed asymmetry is proportionate—and whether users of LLM-based tools are aware it exists.

2 Methods

2.1 Models and Parameters

Three models were tested: GPT-4o (OpenAI), GPT-5.2 Thinking (OpenAI), and Claude Haiku 4.5 (Anthropic), all accessed via production API between January and February 2026. Exact API model identifiers and per-call timestamps are retained in source logs and available on request. Temperature was set to 1.0 for all runs in this study, matching the original GPT-4o setup used for the replication extension to additional models. Within each stimulus pair, Condition A and Condition B runs were interleaved rather than collected sequentially, to guard against model-update confounds. Per-run temperature values and timestamps are retained in source logs for auditability. We acknowledge that temperature sensitivity is an important dimension not characterized in this study; see Section 7.

No system prompt was used beyond the evaluation instructions embedded in the user message. Each prompt was independently sampled 20–30 times per condition (30 for GPT-4o, 20 for GPT-5.2 and Haiku).

2.2 Stimulus Design

We constructed four matched pairs of fictional scientific vignettes. Each pair described an identical study with two versions: a Condition A prompt reporting a statistically significant positive result, and a Condition B prompt reporting a non-significant null result. The pairs spanned four domains to reduce topic-specific confounding: pharmacology (Compound QX-7, $N = 2{,}400$), education (Thornberg Reading Method, $n = 8{,}200$), environmental health (Solvent-K Respiratory Risk, $n = 15{,}000$), and cognitive supplementation (Mineral Compound 44-B, $n = 10{,}000$). All study descriptions included explicit power statements (95% power), confidence intervals, and p-values.

The use of fictional stimuli was a deliberate design choice to eliminate training-data contamination. If real study descriptions were used, the model could draw on memorized literature rather than evaluating the presented evidence on its own terms. This strengthens internal validity at the cost of ecological validity; the generalization question is addressed in Section 7.

2.3 Prompt Matching Rule

The matching rule was strict lexical identity of all design features. Within each pair, the only elements that differed between Condition A and Condition B were: (a) the reported effect size, (b) the confidence interval bounds, (c) the p-value, and (d) the concluding sentence (asserting presence vs. absence of effect). Study descriptions (design type, N, power, domain framing) were identical across conditions. Full prompt texts and cleaned datasets are available from the author on request.

2.4 Response Formats

Two response formats were used: JSON-constrained and free-form (natural language, parsed post hoc). The JSON schema required a valid JSON object with three fields: a classification label (one of

Confirmed, Likely, Unsupported, or Refuted), a probability triad allocating 100 points across three bins ($P(\text{effect} > 0)$, $P(\text{effect} \approx 0)$, $P(\text{effect} < 0)$), and a reasons array. Free-form responses were parsed post hoc using rule-based extraction of probability statements and classification language. JSON tests the model's deliberative output under structural constraint; free-form tests naturalistic behavior and permits hedging-marker analysis.

2.5 Outcome Metrics

Primary metric: conclusion-consistent probability mass. For presence claims: mean $P(\text{effect} > 0)$. For absence claims: mean $P(\text{effect} \approx 0)$. We note that these are not structurally symmetric quantities—$P(\text{effect} > 0)$ covers an unbounded tail while $P(\text{effect} \approx 0)$ covers a neighborhood around zero—so differences in absolute magnitude should be interpreted with this asymmetry in mind. This metric was designated primary because it survives label collapse (see Section 3.2) and is therefore comparable across all model–format conditions.

Secondary metrics: (1) Confirmed classification rate (informative for GPT-4o only). (2) Hedging-marker count per response (free-form only), operationalized as instances of qualifying language including phrases such as "however," "it is important to note," "further research," "caution," "may," "might," and "potentially," counted via keyword matching. (3) Out-of-vocabulary (OOV) label rate.

Bootstrap 95% confidence intervals (5,000 resamples) were computed for the probability-mass gap in each condition. These intervals reflect consistency across API calls within each condition rather than population-level precision; see Section 7 for discussion.

2.6 Sample Sizes and Known Data Issues

Table 1 summarizes sample sizes, extraction outcomes, and quantitative status by condition. Primary-metric inclusion required the relevant component ($P(\text{effect} > 0)$ for condition A; $P(\text{effect} \approx 0)$ for condition B) to be numeric and within $[0, 100]$. Triad sums were not used as an exclusion rule; rows explicitly marked as extraction failures were excluded. In the released cleaned dataset, 20 included rows have non-100 triad sums (range 94–127); applying a strict sum=100 filter shifts condition means by at most 0.91 pp and does not change directional conclusions.

Table 1: Experimental conditions and data completeness. OOV = out-of-vocabulary labels (labels outside {Confirmed, Likely, Unsupported, Refuted}). The 3 OOV rows in GPT-4o JSON were "Supported" (absence condition only). For primary-metric analyses, one row with explicit extraction failure status was excluded (GPT-5.2 free-form absence), yielding n_A=80 and n_B=79 usable for that condition. All other conditions were complete for primary-metric extraction. Haiku 4.5 free-form includes six non-100 triad-sum rows retained under the component-inclusion rule.

Model	Format	N (A/B)	OOV	Missing probs	Quantitative status
GPT-4o	Free-form	120/120	0	0	Full
GPT-4o	JSON	120/120	3 (B only)	0	Full

Model	Format	N (A/B)	OOV	Missing probs	Quantitative status
GPT-5.2	Free-form	80/80	0	1 B row (extraction failure)	Near-complete (n_A=80, n_B=79 usable)
GPT-5.2	JSON	80/80	0	0	Full
Haiku 4.5	Free-form	80/80	0	0	Full
Haiku 4.5	JSON	80/80	0	0	Full

3 Results

3.1 Primary Outcome: Probability-Mass Asymmetry

Across all six conditions, models allocated substantially less probability mass to the conclusion-consistent direction for null claims than for matched positive claims. At the cell level, this asymmetry was directionally consistent in 23 of 24 cells (4 stimulus pairs × 6 model-format conditions), with one small reversal in GPT-5.2 JSON for Compound QX-7 (−2.25 pp). Because stimulus pairs and model-format conditions are crossed (the same four pairs appear in all six conditions, and the same models process all four pairs), the 24 cells are not fully independent. Aggregated to the model-format level, all six conditions showed a positive gap (sign test on 6 conditions: $p = 0.016$). Aggregated to the stimulus-pair level, all four pairs showed a positive gap across every condition tested (sign test on 4 pairs: $p = 0.0625$, one-tailed). This convergence across multiple levels of aggregation—every model, every format, every stimulus domain showing the same directional pattern—is the strongest evidence that the asymmetry is systematic rather than artifactual.

Table 2 presents the aggregate results by model-format condition.

Table 2: Probability-mass asymmetry across models and formats (primary outcome). $P(+)|A$ = mean $P(\text{effect} > 0)$ for presence claims. $P(\approx 0)|B$ = mean $P(\text{effect} \approx 0)$ for absence claims. Gap = $P(+)|A - P(\approx 0)|B$ in percentage points. 95% CIs are bootstrap intervals (5,000 resamples) reflecting within-condition consistency across repeated API calls to the same model with the same prompts; they characterize how reliably each model reproduces the gap, not the precision of a population-level estimate across stimuli (see Section 7, Statistical scope). Hdg = mean hedging-marker count per response. FF = free-form. pp = percentage points.

Model + Format	$P(+)\|A$	$P(\approx 0)\|B$	Gap	95% CI	Hdg A	Hdg B
GPT-4o FF	95.94	61.49	34.45	[31.41, 37.52]	0.88	2.27
GPT-4o JSON	95.56	75.97	19.59	[17.19, 22.21]	0.13	0.91
GPT-5.2 FF	91.94	35.20	56.73	[53.19, 60.34]	0.90	1.49

Model + Format	$P(+\|A$	$P(\approx 0)\|B$	Gap	95% CI	Hdg A	Hdg B
GPT-5.2 JSON	89.34	63.96	25.38	[21.91, 28.60]	0.88	0.60
Haiku 4.5 FF	82.83	51.81	31.01	[26.66, 35.09]	1.91	3.40
Haiku 4.5 JSON	88.99	46.41	42.57	[38.99, 46.01]	0.72	1.02

Across all six conditions, every aggregate probability gap favors presence claims, and every bootstrap 95% CI excludes zero. The asymmetry is not confined to a single model or format; it appears across three architectures from two providers. The gaps range from 19.6 to 56.7 percentage points—substantial by any practical standard.

The largest observed gap was GPT-5.2 free-form (56.73 pp), with one missing absence-side probability extraction row (n_A=80, n_B=79 usable). The smallest observed gap was GPT-4o JSON (19.59 pp).

3.2 Label Classification and the Collapse Finding

GPT-5.2 assigned "Likely" to 100% of evaluations in both conditions and both formats (320/320 responses). Haiku assigned "Likely" to 100% of presence evaluations and 75–100% of absence evaluations depending on format. This renders label-based analysis uninformative for these models.

GPT-4o, by contrast, showed large label asymmetries. In free-form mode, 66.7% of presence claims received "Confirmed" vs. 18.3% of absence claims (gap: 48.3 pp). In JSON mode, 99.2% vs. 46.7% (gap: 52.5 pp). The labels "Unsupported" and "Refuted" appeared exclusively in the absence condition across both formats.

The label collapse is itself a finding. Newer models appear to have been trained or fine-tuned to avoid categorical commitments, defaulting to "Likely" regardless of evidence strength. This de-risks the model's surface behavior but does not eliminate the underlying asymmetry—it migrates to probability allocation. The asymmetry did not disappear in newer models; it went underground. This output-channel shift has implications for monitoring: practitioners who rely on label-based confidence thresholds will not detect the asymmetry in newer models, but it will still distort probability-weighted evidence summaries.

3.3 Per-Pair Consistency

At the pair-condition level, the probability gap is directionally positive in 23 of 24 cells. Table 3 shows the per-pair breakdown for GPT-4o free-form (the condition with the richest behavioral variation).

Table 3: Per-pair results, GPT-4o free-form ($n = 30$ per cell).

Stimulus pair	Conf A	Conf B	$P(+)\|A$	$P(\approx 0)\|B$	Hdg A	Hdg B
Compound QX-7	100%	56.7%	95.6	82.7	0.43	1.37
Mineral 44-B	80.0%	13.3%	96.6	59.5	1.10	2.20
Solvent-K	20.0%	0.0%	94.5	57.0	1.07	2.97
Thornberg	66.7%	3.3%	97.1	47.6	0.93	2.53

For GPT-4o free-form, the probability gap is positive for all four pairs (range: 12.9 to 49.5 pp), confirming the effect is not driven by a single anomalous stimulus. Pair-level variation is substantial: Compound QX-7 shows the smallest gap and highest null-consistent allocation, while Thornberg shows the largest gap and lowest null-consistent allocation. We report this heterogeneity descriptively without causal attribution.

Mineral Compound 44-B anomaly. In GPT-4o JSON mode, this pair produced 21 of 30 absence-condition evaluations with the "Refuted" label—a label indicating the claim is actively contradicted. No other pair produced any "Refuted" labels. This is the most extreme expression of asymmetric skepticism in our data: the model not only withholds confirmation but actively rejects a null conclusion that the evidence supports. This anomaly warrants further investigation but does not drive the primary probability-gap finding, which holds across all pairs.

3.4 OOV Labels and Schema Violations

Three GPT-4o JSON responses used the out-of-vocabulary label "Supported," all in the absence condition. No presence-condition responses produced OOV labels in any model or format. The confinement of schema violations to the absence condition is consistent with the broader pattern: the model's uncertainty about how to classify null findings extends even to the label vocabulary itself. These three rows are included in probability-gap calculations (which are label-independent) and excluded from label-rate calculations. A sensitivity analysis (Section 5.1) confirms no impact on primary findings.

4 Validity and Threat Model

4.1 Parser Integrity

All released datasets preserve original model-returned labels as authoritative; parser-derived labels were used for quality assurance only. Out-of-vocabulary labels were flagged rather than silently remapped. Probability fields were coerced to numeric with missing values preserved (not imputed). This design ensures no post hoc analytic decision favored one condition over another.

4.2 JSON Constraint Effects

JSON-constrained responses show higher overall confidence (GPT-4o JSON: 99.2% Confirmed for presence vs. 66.7% in free-form). This is expected: structured output suppresses hedging and may push the model toward stronger commitments. However, the directional asymmetry persists in both formats. JSON does not create the asymmetry; it changes its magnitude. The probability gap is actually smaller in JSON (19.59 pp) than in free-form (34.45 pp) for GPT-4o, suggesting JSON may partially compress the asymmetry by forcing explicit probability allocation.

4.3 Label Collapse in Newer Models

GPT-5.2 and Haiku default to "Likely" for nearly all evaluations. We designated the probability-mass gap as the primary metric precisely because it survives label collapse. The label collapse itself may reflect reinforcement learning from human feedback (RLHF)-driven conservatism: models trained to avoid overconfident claims default to the safest label. This is a secondary finding with its own implications for LLM deployment, but it does not invalidate the probability-gap analysis.

4.4 Prompt Symmetry

The matching rule is strict lexical identity of design features. The only differences are result statistics and the concluding sentence. Full prompts are available from the author on request. A reviewer may argue that any phrasing difference could drive the effect. We note that the asymmetry is consistent across four independent stimulus pairs with different domains and framings, reducing the probability that a single phrasing artifact explains the pattern. The fictional-stimulus design ensures the model cannot draw on prior knowledge of the described study.

5 Sensitivity Analyses

5.1 OOV Exclusion

Removing the 3 OOV "Supported" rows from GPT-4o JSON: these rows have no impact on the probability-gap metric (computed on all rows regardless of label). The Confirmed-rate gap changes negligibly: from 52.5 pp (119/120 vs. 56/120) to 51.3 pp (119/120 vs. 56/117). The primary finding is robust to OOV handling.

5.2 Single-Row Extraction Failure Sensitivity

GPT-5.2 free-form includes one absence-condition row with explicit extraction failure status; this row is excluded from primary-metric computation by rule. The resulting condition estimate (gap 56.73 pp, 95% CI $[53.19, 60.34]$) remains strongly directionally aligned with the full benchmark pattern.

5.3 Cross-Condition Sign Consistency

The strongest model-free evidence for systematic asymmetry is the directional consistency across the pair-condition matrix. The asymmetry favored positive claims in 23 of 24 cells (4 stimulus pairs × 6 model-format conditions). Because the 24 cells share structure (the same stimuli appear across conditions; the same models process all pairs), they are not fully independent Bernoulli trials. We therefore report sign tests at two aggregation levels that respect this structure: at the model-format level, all 6 conditions showed a positive aggregate gap ($p = 0.016$); at the stimulus-pair level, all 4 pairs showed a positive gap across every condition ($p = 0.0625$, one-tailed). The convergence of these tests—neither alone is decisive, but together they indicate that the asymmetry is not confined to any single model, format, or stimulus domain—constitutes the primary evidence for a systematic pattern. The one cell-level reversal was small (−2.25 pp) and occurred in GPT-5.2 JSON for Compound QX-7, a pair that showed the smallest asymmetry across conditions generally.

6 Discussion

6.1 The Strongest Skeptical Interpretation: Rational Calibration

The most important objection to our thesis is that the observed asymmetry may reflect rational calibration rather than a defect. A Bayesian reasoner with reasonable priors might legitimately assign lower credence to null claims for at least three reasons. First, if the base rate of true effects in the model's training data is high (because published studies disproportionately report positive findings), the prior favors effects. Second, null results are in fact more sensitive to methodological shortcomings—underpowering, insensitive measures, implementation failure—so some additional skepticism may be warranted. Third, the model may have learned an accurate descriptive model of how scientific communities treat null results and may be reproducing that social norm rather than expressing an independent epistemic judgment.

We take this objection seriously, and it is why we frame our finding as an asymmetric burden of proof rather than as irrational bias. However, three features of the data are difficult to reconcile with purely rational calibration.

First, the stimuli were explicitly designed to neutralize the standard methodological concerns. Every vignette specified large samples (2,400–15,000), 95% power, pre-registration, and double-blinding. A calibrated Bayesian agent should update substantially on these features; yet the models' probability distributions still reserve about 24–65% of mass away from the conclusion-consistent direction for null claims, while reserving about 4–17% for positive claims.

Second, the model does not apply equivalently cautious scrutiny to positive claims. It does not hedge presence conclusions, demand replication, or note that a single significant result may be a false positive. If the asymmetry were purely rational, we would expect the model to demand comparable corroboration from both directions.

Third, the magnitude and consistency of the effect argue against modest prior-driven calibration. The asymmetry appeared in 23 of 24 pair-condition cells, with gaps ranging from 19.6 to 56.7

percentage points at the condition level. A rational prior adjustment would be expected to produce smaller and more variable gaps, particularly given that the stimuli explicitly describe high-powered, pre-registered designs that should substantially update any reasonable prior.

6.2 The Output-Channel Shift

The comparison between GPT-4o and newer models reveals that the asymmetry is not a fixed surface behavior but a deeper pattern that adapts to output constraints. When models are trained or constrained to avoid categorical labels, the asymmetry migrates to probability allocation. This has two implications. First, the pattern is consistent with something more fundamental than a quirk of label selection—it suggests the asymmetry is tied to how these models process evidence polarity, though behavioral data alone cannot confirm the mechanism. Second, it means that monitoring for this asymmetry requires examining probability distributions, not just labels. Organizations relying on label-based confidence thresholds will not detect it.

6.3 Practical Implications

Evidence synthesis. Automated systematic-review tools that aggregate model-generated credibility scores may systematically underweight null findings. Because published null results are already underrepresented due to the file-drawer effect, an LLM layer that further discounts them can amplify rather than correct the underlying publication bias.

Safety and risk communication. In regulatory and safety contexts, null results are often the desired outcome: no contamination detected, no adverse events observed, no structural defect found. If models routinely hedge these conclusions with unsolicited caveats, they may induce overreaction or erode trust in legitimate safety assessments.

Decision-support filtering. Workflows that use confidence thresholds for inclusion in evidence summaries can systematically exclude high-quality negative evidence. In a pipeline where the model allocates 96% probability to positive conclusions but only 35–76% to matched null conclusions, the resulting summary may be materially distorted.

6.4 Hypothesis: Training Distribution as Mechanism

We note—as a hypothesis, not a demonstrated causal claim—that the asymmetry may originate in the distribution of scientific text in training corpora. Published null results are typically accompanied by hedging, caveats, and calls for replication, while positive results are presented with confidence. If models learn to reproduce these distributional patterns, they may systematically hedge null conclusions and endorse positive ones, not because they have reasoned about evidence quality but because that is the surface pattern of the text they were trained on. This hypothesis is consistent with our data but not uniquely supported by it; the current experiments cannot distinguish training-distribution effects from other explanations.

7 Limitations and Future Work

Model coverage. Three models from two providers. We do not claim universal generalization; additional families (Gemini, Llama, Mistral) should be tested.

Temperature sensitivity. All runs used temperature 1.0, which maximizes response diversity. The reported gap magnitudes are specific to this setting; most production deployments use lower temperatures (0.0–0.7), and the relationship between temperature and asymmetry magnitude is uncharacterized.

Fictional stimuli. Eliminates training-data contamination but limits ecological validity. Real-world study evaluations may produce different magnitudes.

Four stimulus pairs. Sufficient to demonstrate cross-domain consistency but does not exhaust the space. The substantial pair-level variation suggests domain effects warrant systematic investigation.

Free-form extraction risk. While all six conditions are quantitatively analyzable for the primary metric after rule-based cleaning, free-form outputs remain more extraction-fragile than JSON and require explicit parsing QA controls.

Negation-framing confound. Null conclusions inherently use negation or absence language ("no significant effect," "did not reduce"), while positive conclusions use affirmative language. Language models may process negation differently from affirmation at the representation level. The current design cannot fully separate evidential asymmetry from negation-processing effects, though the consistency across four stimulus pairs with different phrasings provides partial mitigation. A future condition using affirmatively framed null conclusions (e.g., "The study confirmed the absence of a clinically meaningful effect") would help distinguish these explanations.

No equivalence-framing conditions. Prior observations suggest that equivalence-testing language ("the observed effect falls within pre-registered equivalence bounds") may partially mitigate the asymmetry. Controlled testing of framing interventions is the most important direction for future work.

No formal pre-registration. The experimental design was specified before execution but was not registered in a public pre-registration repository.

Statistical scope. The bootstrap confidence intervals and within-condition statistics reported in this study reflect consistency of model outputs across repeated API calls within each condition. They characterize how reliably a given model reproduces the asymmetry under fixed conditions, not the precision of a population-level estimate. The effective independent sample for cross-domain generalization is four stimulus pairs, and readers should interpret the reported precision accordingly. We view the 23/24 sign consistency across pair-condition cells as stronger evidence of systematic asymmetry than any single condition-level statistic, precisely because it does not depend on within-condition variance assumptions.

8 Conclusion

Across three model families, two response formats, and four matched stimulus pairs, large language models applied an asymmetric burden of proof to positive and null scientific claims. The asymmetry was directionally consistent in 23 of 24 pair-condition cells, with every model-format condition and every stimulus pair showing the same directional pattern. The probability-mass gap ranged from 19.6 to 56.7 percentage points across six conditions. The asymmetry persisted even when discrete labels collapsed entirely, surfacing through probability allocation rather than categorical commitment. This output-channel shift means the asymmetry may be invisible to label-based monitoring yet still distort probability-weighted evidence summaries.

The asymmetry matters because it is directional. An LLM evidence evaluator that symmetrically discounted all claims would merely be cautious; one that selectively discounts null claims can reproduce and potentially amplify the publication bias that has distorted the scientific record for decades. Practitioners deploying LLMs for evidence synthesis, safety assessment, or decision support should not assume symmetric handling of positive and negative evidence. Until the asymmetric burden of proof is better understood and mitigated, human oversight of null-result evaluations remains essential.

Acknowledgments

The principal investigator oversaw the research design, experimental protocol, and interpretation of results. Hermes Labs infrastructure executed automated experimental runs. Large language model assistance was used during experiment prototyping, analysis support, and manuscript drafting under human supervision and editorial control.

References

Briggs, R. C., Mellon, J., & Arel-Bundock, V. (2026). It must be very hard to publish null results [Preprint]. https://doi.org/10.31235/osf.io/zr5vf_v1

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49(12), 997–1003.

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science, 1(2), 259–269.

Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated annotation with generative AI suggests LLM annotation may lack reliability. arXiv:2306.07899.

Perez, E., et al. (2022). Discovering language model behaviors with model-written evaluations. arXiv:2212.09251.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, 86(3), 638–641.

Sharma, M., et al. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548.

Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Anticipated Objections and Responses

"This is rational caution, not bias." We agree that some asymmetry is epistemically defensible, which is why we use the framing "asymmetric burden of proof." Observed gaps (19.6–56.7 pp) remain large under high-power, pre-registered stimulus descriptions. The model does not apply symmetric caution to positive claims—it does not hedge, demand replication, or note false-positive risk in parallel. The asymmetry is in the scrutiny, not merely the prior.

"Your 24 cells aren't independent." Correct. The 4 stimulus pairs and 6 model-format conditions form a crossed design with shared structure. We report sign tests at two aggregation levels that respect this: 6/6 conditions positive ($p = 0.016$) and 4/4 pairs positive across all conditions ($p = 0.0625$). Neither alone is decisive; their convergence across independent axes (different models, different domains) is the basis for our claim.

"Your prompts aren't symmetric." Strict lexical identity of design features. Only result statistics and concluding sentence differ. Full prompts available. Consistency across four pairs with different domains reduces the probability of a single phrasing artifact. We acknowledge in Section 7 that the inherent use of negation language in null conclusions is a confound that future work should address with affirmatively framed null conditions.

"JSON changes behavior." Yes. JSON suppresses hedging and elevates overall confidence. But the directional asymmetry persists in both formats. The probability gap is actually smaller in JSON for GPT-4o (19.59 vs. 34.45 pp), suggesting JSON partially compresses rather than inflates the effect.

"Label collapse makes it meaningless." Label collapse makes label-based metrics meaningless for newer models—which is precisely why we designated the probability-mass gap as the primary metric. The asymmetry persisting under label uniformity strengthens the finding: it shows the phenomenon survives output constraints.

"This is cherry-picking." We report all runs, all models, all formats, and all pairs. Data quality handling is rule-based and explicit (including one extraction-failure exclusion in GPT-5.2 free-form). The full dataset is available for reanalysis.

"You didn't test enough models." Three families from two providers across multiple generations. Sufficient to establish cross-architecture consistency. Universal generalization is not claimed; additional families are proposed as future work.

"Temperature 1.0 inflates variance." Higher variance increases noise in both directions, making it harder to detect systematic directional asymmetry. A consistent gap under maximum response diversity is more robust, not less. Temperature sensitivity analysis is proposed as future work.