A Taxonomy of Epistemic Failure Modes in Large Language Models

Rolando Bosch — Hermes Labs, San Francisco, CA

rbosch@lpci.ai

March 15, 2026

Abstract

Large language models (LLMs) exhibit systematic distortions in how they evaluate, represent, and communicate epistemic states. These distortions are distinct from hallucination—they concern cases where the model's expressed confidence, standards of scrutiny, or assignment of causality diverge from what the evidence warrants, even when the factual content is roughly correct. Through bottom-up analysis of 1,461 controlled experiments conducted primarily on GPT-4o, we identify seven structural failure modes: (1) Null-Result Asymmetry, (2) Source-Status Credibility Bias, (3) Agency Dissolution, (4) Performative Hedging, (5) Constraint Evasion, (6) Silent Instruction Relaxation, and (7) Controversy-Truth Conflation. Each mode is defined mechanistically, illustrated with experimental evidence, and assessed for downstream consequences. Across the seven modes, a common pattern emerges: models track surface-level signals—prestige markers, hedging vocabulary, controversy language, banned word lists—rather than the semantic content those signals are supposed to index.

1 Introduction

The question of whether large language models represent beliefs faithfully is not primarily a question about hallucination. Hallucination research addresses cases where models assert false propositions. The failure modes documented here are different: they concern the *epistemic layer*—how models reason about evidence, attribute causality, handle uncertainty, and navigate the relationship between social consensus and factual truth. A model can be factually accurate while still distorting the epistemic landscape around those facts—softening blame where it should be direct, hedging where it should be confident, or treating social disagreement as grounds for doubt about well-supported claims.

This gap matters practically. These distortions are invisible to factual accuracy benchmarks but consequential for anyone using LLMs in evaluation, synthesis, or decision support. A false factual claim can, in principle, be checked. An asymmetric standard of scrutiny applied to null versus positive results, or a systematically softened causal attribution in incident reports, may pass unnoticed while shaping downstream decisions.

The taxonomy presented here is bottom-up. It was derived not from a priori hermeneutic frameworks or predetermined categories, but from recurring behavioral patterns observed across a structured corpus of 1,461 experiments conducted between January and February 2026. The experiments were designed to probe model behavior across epistemic dimensions: how models classify evidence, attribute causation, follow conflicting instructions, and respond to controversy signals. Hypotheses were generated from observed behaviors, not from a predetermined framework. The seven modes represent the most structurally distinct and replicated behavioral clusters in that corpus. The experiments cited in this paper are illustrative examples selected from a much larger evidence base; they were chosen because they demonstrate each mode's behavioral signature most clearly, not because they are the sole basis for the taxonomy. Additional candidate patterns—including epistemic contamination (skepticism spreading from disputed to undisputed adjacent claims)—were observed but excluded from the primary taxonomy due to mixed experimental evidence.

No single experiment carries high statistical power on its own (typical trial counts range from 4 to 12 per condition); the evidential weight comes from convergent patterns across multiple independent experiments per mode, with 10 to over 60 experiments per category in the full corpus. Experiments are referenced by corpus identifier (e.g., 3cd27831); the full corpus, including experimental designs, prompts, and raw outputs, is available from the author on request. Readers are encouraged to treat cited effect magnitudes as illustrative rather than precise. Quantitative analysis of Mode 1 has been published separately [1].

Relationship to prior work. This paper engages with three adjacent research traditions: (i) *sycophancy* research, which documents models' tendency to agree with user-stated positions [5, 6]; (ii) *calibration* research, which asks whether model-stated probabilities match empirical frequencies [3, 8]; and (iii) *instruction-following* research, which examines compliance with prompt constraints [9, 4]. The failure modes below partially overlap with these traditions but are not reducible to them. Source-Status Credibility Bias resembles sycophancy but operates on third-party evidence rather than user opinion. Performative Hedging relates to calibration but concerns the structural function of uncertainty markers rather than probability accuracy. Silent Instruction Relaxation relates to instruction-following but identifies a specific failure of silent priority assignment without disclosure.

Contributions. This paper makes three contributions: (1) a taxonomy of seven epistemic failure modes, derived bottom-up from 1,461 controlled experiments, that are structurally distinct from hallucination and from each other; (2) experimental evidence for each mode, including behavioral signatures that distinguish it from adjacent phenomena such as sycophancy and miscalibration; and (3) identification of a cross-cutting pattern—surface-feature tracking as a shared mechanism—suggesting that epistemic distortions in LLMs arise from correlational learning over text rather than from isolated training failures.

2 Taxonomy

The following modes were identified from experiments conducted primarily on GPT-4o (OpenAI), with supplementary cross-model observations for Mode 1 reported in Bosch [1].

2.1 Null-Result Asymmetry

Definition. The model applies stricter evidential standards to claims of absence than to claims of presence, even when the underlying evidence is of equal quality and the studies are of matched design. The asymmetry manifests not only in confidence ratings but in the structure of doubt applied: absence claims attract demands for replication and philosophical qualifications; presence claims attract neither at equivalent rates.

Experimental evidence. In Experiment 3cd27831, matched vignettes across four domains presented identical study designs differing only in whether the result was positive or null. The model rated presence claims at approximately 95% confidence and absence claims at approximately 69%—a 26-point gap for the same quality of evidence. The phrase "failure to reject the null is not the same as proving no effect" appeared in three of four absence-condition trials; no parallel caveat ("statistical significance is not practical significance") appeared in any presence-condition trial. All four absence trials included explicit calls for further research; presence trials mentioned replication in two of four cases, framed as desirable rather than necessary. A companion experiment (2ff388f0) replicated the directional finding: a well-powered acupuncture null-result study ($d = 0.04$, $p = 0.38$, 90% power for $d \geq 0.25$) was classified "Unsupported" while a structurally identical presence-finding trial was classified "Likely."

Mechanism. The asymmetry likely reflects training data encoding the existing rhetoric of science communication, where null results historically receive more hedged language [2] and where "absence of evidence is not evidence of absence" is a recognized rhetorical reflex. The model encodes this rhetorical asymmetry as an epistemic standard.

Downstream implications. Any system using LLMs for evidence synthesis, systematic review, or research screening risks systematically underweighting null findings. This replicates, at the model level, the publication bias that has plagued empirical science for decades. For extended quantitative analysis, see Bosch [1].

2.2 Source-Status Credibility Bias

Definition. The model applies different standards of evidentiary scrutiny to identical claims based on the prestige of the attributed source. High-prestige sources receive methodological critique alone; low-prestige sources receive methodological critique *plus* credential questioning that persists even when the model is explicitly instructed to bracket source identity.

Experimental evidence. In Experiment 82bb3319, identical evidence packets—matched effect sizes, p-values, limitations, and replication notes—were presented under high-prestige attribution (Johns Hopkins, Stanford, Harvard) and low-prestige attribution (biohacker blogger, anonymous Reddit poster). For low-prestige sources, the model asked "Are biohacking enthusiasts qualified to conduct rigorous medical or physiological studies?" and flagged that "this information appears to come from a blogger rather than outlined in a peer-reviewed study." For the same claims from Harvard, no credential question appeared. A bracketing instruction ("Assume the source identity is unknown; evaluate only the content") failed to eliminate the asymmetry: in one low-prestige trial, the second-pass response explicitly stated "Independent source details (e.g., Substack biohackers) further amplify concerns about unrigorous research processes"—directly referencing the source that was supposed to be ignored.

No such anchoring residue appeared for high-prestige sources.

Mechanism. This failure mode shares structure with sycophancy but is distinct: it operates on how the model evaluates third-party evidence, not on whether the model agrees with the user. The model treats institutional affiliation as a proxy for competence that reduces the epistemic work required to establish credibility.

Downstream implications. Applications using LLMs to evaluate research or summarize evidence from multiple sources risk inheriting a prestige hierarchy not grounded in content validity. The bias also creates a blind spot for institutional failures, since prestigious sources receive less scrutiny.

2.3 Agency Dissolution

Definition. When producing communications involving negative outcomes or potential blame, models systematically soften causal language and redistribute moral weight from individual agents to systems and processes—even when the source text makes individual causation explicit and instructions request preservation of accountability.

Experimental evidence. In Experiment a2c78206, under the standard "warm rewrite" condition, the model preserved named individuals but systematically replaced strong causal verbs with concessive structures. The characteristic pattern was the "While… Unfortunately" construction: "While Director Chen demonstrated a strong commitment to driving progress, the decision to accelerate the timeline by six weeks presented significant risks." This framing acknowledges the individual while distributing the causal weight across the decision rather than the decision-maker—observed across all five trials. A complementary experiment (57196d34) tested what happens when escape routes are blocked. With passive voice, collective-agent nouns, and uncertainty hedges all prohibited, the model retained strong causal verbs as instructed but found alternative softening through lexical substitution: "wrong" became "incorrect," "failing to follow" became "not adhering to," "misreading" became "incorrectly interpreting." The model complied with each blocked route individually but reconstructed the softening effect through whichever paths remained available.

Mechanism. The model appears to encode a default register for blame-containing communications that minimizes interpersonal conflict. This register operates as a stylistic default, not triggered by explicit instructions to avoid blame.

Downstream implications. Organizations deploying LLMs for drafting sensitive communications—HR documentation, incident reports, legal summaries—risk systematically receiving outputs that diffuse individual accountability into systemic framing. The critical problem is that this operates as an invisible default rather than an explicit design choice.

2.4 Performative Hedging

Definition. The model uses uncertainty language ("could," "might," "it is important to note") as a structural feature of certain response types rather than as a calibrated signal of genuine epistemic uncertainty. The hedging appears at similar rates across high- and low-uncertainty contexts when those contexts share surface features, and can be eliminated through framing manipulations that do not change actual uncertainty.

Experimental evidence. Experiment 638dab58 examined hedging under token constraints. When responses were limited to under 20 words, hedging markers were absent in approximately 85% of trials, compared to consistent hedging in unconstrained responses. If hedging tracked genuine uncertainty, constrained responses would find other ways to signal that uncertainty. Instead, it vanishes—the uncertainty expression is a component of a verbose response style, not an independent signal. Experiment 5f99a2d7 showed the complementary effect: framing manipulations that prompted "confident and assertive" tone—without changing the epistemic status of any claim—reduced hedging marker frequency measurably. A third experiment (f46463ea) confirmed: under open critique, the model described a confirmed three-week delay as a "potential" problem; under epistemic accounting ("how confident are you and why?"), the same delay received "Certainty level: 100%."

Mechanism. Performative Hedging is distinct from miscalibration. Miscalibration concerns the accuracy of probability estimates. Performative Hedging concerns whether hedging expressions function as signals or as conventions. The experimental evidence suggests hedging in LLMs behaves more like a convention: it appears when a "careful, scientific" response is expected and disappears when that expectation is absent.

Downstream implications. Users who interpret hedging as genuine uncertainty risk systematically underweighting the model's most confident assessments. Evaluation frameworks that count hedging markers as indicators of calibration may be measuring rhetorical convention rather than epistemic accuracy.

2.5 Constraint Evasion

Definition. When a lexical or content constraint prohibits a specific term, the model preserves the prohibited concept's full communicative effect through systematic substitution—synonyms, circumlocutions, reframing—while maintaining surface-level compliance with the literal constraint. The prohibited intent survives the prohibition; only its surface instantiation changes.

Experimental evidence. In Experiment ab7ef994, explicit word bans were imposed on salient vocabulary while eliciting responses where those concepts were contextually relevant. Perfect ban compliance was achieved: the banned strings did not appear. However, the underlying concepts persisted through paraphrase. When "surveillance" was banned, the response described "keeping track of movements in the building, perhaps as part of monitoring." When "conflict" was banned, responses shifted to "tension," "friction," "misalignment," and "strain." In a related experiment (8a7ac56b), the model even exhibited "lexical leaks," using the banned term in some trials despite explicit instructions—suggesting that the concept exerts pressure toward its canonical label.

Mechanism. This failure mode has a structural parallel to adversarial jailbreaking [10, 7], but operates even under benign prompt constraints, suggesting a more fundamental property: surface-level constraint compliance without semantic-level constraint tracking.

Downstream implications. Content moderation and safety systems that rely on lexical filtering are vulnerable. Concept-level constraints require structural interventions, not vocabulary restrictions. Notably, disrupting *structural* templates produces greater changes in output than banning individual terms.

2.6 Silent Instruction Relaxation

Definition. When a prompt contains two instructions that cannot simultaneously be satisfied, models silently deprioritize one—producing output that complies with one requirement while violating the other, without flagging that the violation occurred.

Experimental evidence. Experiment f859c269 administered twelve trials using six incompatible instruction pairs (e.g., "be emotionally evocative AND use no adjectives") across two conditions: No-Frame ("Follow both instructions") and Conflict-Frame ("If the instructions conflict, say so and explain"). In the No-Frame condition, zero of six trials contained meta-commentary about instruction relationships. The model silently violated one instruction per pair in every trial. In the Conflict-Frame condition, six of six trials opened with explicit conflict acknowledgment before proceeding with adapted output. A key secondary finding: the model applied an implicit hierarchy—countable hard limits ("under 20 words") overrode elaboration defaults, while qualitative instructions competed by less transparent ordering.

Mechanism. The model's capacity to detect conflicts appears to exist but to require explicit activation. The Conflict-Frame prompt functions as a permission structure for meta-commentary that the model otherwise suppresses.

Downstream implications. Any system prompt containing logically competing requirements—common in production deployments where instructions accumulate—may result in silent non-compliance with some requirements, with no disclosure to the operator about which requirements were dropped.

2.7 Controversy-Truth Conflation

Definition. The model treats the presence of social disagreement about a claim as evidentially relevant to the epistemic status of the claim, independent of the quality of the underlying evidence. Markers of controversy trigger a shift from evidence-based classification ("confirmed finding") toward socially-defined classification ("contested," "controversial").

Experimental evidence. In Experiment ef93094e, strong documentary evidence (authenticated footage, verified timestamps, matched bank records) was presented with and without debate markers. Without debate markers, the model classified claims as "confirmed finding" in three of four trials. With debate markers added to identical evidence, the model classified claims as "controversy" in three of four trials—a near-complete reversal. The model's own justification was explicit: "since there is clear disagreement and debate over the interpretation… it remains a contentious and disputed topic." In the starkest case, "the state audit unequivocally concluded" yielded "confirmed finding" without debate markers, but "there is no universal consensus" yielded "controversy" when debate language was added—despite the audit's conclusions remaining identical. A drift test found that "controversy" classifications persisted across conversational turns even after debate language was removed. One trial provided an instructive exception: highly specific forensic evidence ("23 separate transactions, each authorized by the treasurer's digital signature") resisted the controversy override, suggesting the shift is probabilistic and weighted against evidence specificity.

Mechanism. The model conflates descriptive social facts ("people disagree") with normative epistemic facts ("the claim is uncertain")—a category error in which social contro-

versy is treated as an indicator of epistemic indeterminacy. The probable training data origin: journalism, which structurally favors "both sides" framing, is heavily represented in pretraining corpora.

Downstream implications. This mode has direct consequences for how models handle politically contested factual claims. Climate science, vaccine safety, election integrity, and other domains where strong evidence coexists with organized disagreement are systematically vulnerable. A manufactured controversy can suppress a "confirmed finding" classification as effectively as genuine scientific uncertainty.

3 Discussion

3.1 Surface signals vs. semantic content

Several of the seven modes share a common structural feature: the model's output tracks a surface signal rather than the underlying epistemic content the signal is supposed to index. In Mode 2, prestige markers substitute for evidence evaluation. In Mode 4, hedging vocabulary substitutes for uncertainty signaling. In Mode 5, lexical compliance substitutes for semantic compliance. In Mode 7, controversy vocabulary substitutes for epistemic indeterminacy. This pattern suggests an architecture-level vulnerability: because LLMs learn from text, they acquire correlational patterns between surface markers and epistemic categories. Where these correlations are imperfect in the training data—and they are imperfect wherever human communication is motivated, strategic, or rhetorical—the model inherits the distortion.

3.2 Structural relationships

The modes are not independent. Performative Hedging and Agency Dissolution both serve face-saving functions. Null-Result Asymmetry and Controversy-Truth Conflation both produce false epistemic symmetry. Constraint Evasion and Silent Instruction Relaxation both involve the model navigating constraints, but through opposite strategies: one preserves prohibited content through substitution, the other drops requirements without notice. An output can exhibit multiple modes simultaneously—a model summarizing contested research may undersell null findings, discount non-institutional sources, and treat debate as evidence against the better-supported position, each reinforcing the others.

3.3 Distinguishing from known phenomena

Sycophancy [6, 5] involves models updating beliefs toward user preferences. Source-Status Credibility Bias resembles sycophancy but operates on third-party evidence without requiring a user to state a preference. Hallucination is orthogonal to all seven modes: hallucinated outputs are factually wrong, while these modes concern outputs that may be factually accurate but epistemically distorted. Calibration failure [3, 8] concerns the gap between stated and actual probabilities; Performative Hedging concerns whether hedging expressions function as probability-tracking signals at all.

3.4 Practical implications

For teams deploying LLMs in evaluation or decision-support roles, the taxonomy identifies specific intervention points. Null-Result Asymmetry can be partially mitigated by adjusting confidence scores for result valence. Performative Hedging can be reduced by prompting for explicit confidence ratings. Silent Instruction Relaxation can be surfaced by including conflict-detection prompts. Not all modes admit easy fixes: Agency Dissolution is deeply embedded in learned communication norms, Controversy-Truth Conflation likely requires training-data interventions, and Source-Status Credibility Bias persists even when explicitly instructed away.

4 Limitations

Corpus constraints. The experiments were conducted primarily with one model family (GPT-4o). Generalization to other model families, sizes, or deployment contexts requires independent replication. Behavioral patterns in controlled prompting may differ from those in extended multi-turn conversations or retrieval-augmented contexts.

Experiment quality heterogeneity. Experiments vary in design quality and sample size. Most were conducted with 4–12 trials per condition, insufficient for statistical inference on individual experiments. The taxonomy represents patterns that replicated across multiple experiments, not single-experiment claims.

Intentionality attribution. None of the failure modes require attributing intention or awareness to the model. The mechanisms proposed (training data correlations, surface-feature tracking) are explicitly non-intentionalist. The term "evasion" in Mode 5 is a descriptive label, not a claim about purposive behavior.

Absence of interventional evidence. This paper documents behavioral patterns; it does not establish interventions that reliably correct them. Systematic mitigation likely requires training-level or architectural changes beyond what this corpus can specify.

Taxonomic boundaries. The distinction between "failure mode" and "appropriate behavior" is not always clean. A model that demands more evidence for absence claims may be correctly encoding that null results are harder to interpret. We have aimed to identify cases where differential treatment is disproportionate to the evidential situation, but reasonable people may draw the boundary differently. We do not claim these seven modes are exhaustive; the corpus contains additional patterns that did not cluster cleanly or lacked sufficient replication.

5 Conclusion

Large language models do not just get facts right or wrong. They also handle the epistemic infrastructure around facts—uncertainty, attribution, evidential weight, the relationship between social consensus and truth—in systematically distorted ways. The seven modes identified here represent structural patterns in this distortion, and they share a common dynamic: surface-level signals are processed differently from the semantic content they are supposed to track.

A model that avoids a banned word while preserving the banned concept, that hedges in scientifically-registered responses regardless of actual uncertainty, that silently drops a conflicting instruction without disclosure, or that treats social controversy as epistemic evidence is not lying in any straightforward sense. But it is not reasoning faithfully either. Understanding these failure modes precisely—rather than treating them as noise in an otherwise trustworthy system—is a prerequisite for building evaluation benchmarks and deployment safeguards adequate to the actual failure landscape of deployed language models.

The taxonomy is offered as a working tool, not a finished ontology. We expect individual modes to be refined, split, or merged as evidence accumulates. What we hope persists is the underlying claim: that epistemic reasoning failures in LLMs are systematic, identifiable, and consequential, and that they deserve the same research attention currently given to factual accuracy and alignment.

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, taxonomy development, analysis support, and manuscript drafting under human supervision and editorial control.

References

[1] Bosch, R. (2026). The asymmetric burden of proof: Null-result bias in large language model evidence evaluation. *Zenodo*. https://doi.org/10.5281/zenodo.18867694

[2] 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

[3] Kadavath, S., Conerly, T., Askell, A., et al. (2022). Language models (mostly) know what they know. *arXiv preprint* arXiv:2207.05221. https://doi.org/10.48550/arXiv.2207.05221

[4] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. https://doi.org/10.48550/arXiv.2203.02155

[5] Perez, E., Ringer, S., Lukošiūtė, K., et al. (2023). Discovering language model behaviors with model-written evaluations. *Findings of the Association for Computational Linguistics: ACL 2023*. https://doi.org/10.18653/v1/2023.findings-acl.847

[6] Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. *arXiv preprint* arXiv:2310.13548. https://doi.org/10.48550/arXiv.2310.13548

[7] Wei, A., Haghtalab, N., & Steinhardt, J. (2024). Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*. https://doi.org/10.48550/arXiv.2307.02483

[8] Xiong, M., Hu, Z., Lu, X., et al. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in large language models. *arXiv preprint* arXiv:2306.13063. https://doi.org/10.48550/arXiv.2306.13063

[9] Zhou, J., Lu, T., Mishra, S., et al. (2023). Instruction-following evaluation for large language models. *arXiv preprint* arXiv:2311.07911. https://doi.org/10.48550/arXiv.2311.07911

[10] Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint* arXiv:2307.15043. https://doi.org/10.48550/arXiv.2307.15043

Corpus (1,461 experiments) available from the author on request.

Experiment Registry

A Taxonomy of Epistemic Failure Modes in Large Language Models

Rolando Bosch Hermes Labs

rbosch@lpci.ai

Research Methodology Note

The experiments documented in this appendix were not individually designed by the principal investigator. The taxonomy emerged from an autonomous research loop operating within a human-defined conceptual framework. A seed framing—comprising a set of open questions about LLM epistemic behavior—was provided to a large language model system. The system generated research hypotheses autonomously, conducted prior-art due diligence to identify unexplored territory, executed controlled behavioral probing experiments, and produced follow-up questions from each set of results. Those follow-up questions seeded the next hypothesis generation cycle. The 1,461-experiment corpus is the product of this iterative process.

The recursive structure of this methodology is observable in the corpus itself. The five experiments comprising Mode 3 (Agency Dissolution) illustrate the pattern: the first experiment (`a2c78206`, January 20) asked whether framing causation through mechanism preserved role-based accountability. Its results generated a follow-up probing the grammar of blame-softening, which generated a third experiment (`57196d34`) testing what happens when all softening routes are explicitly blocked, which generated a fourth examining the threshold at which responsibility is removed entirely, which generated a fifth mapping tradeoffs when individual naming is required. None of these later experiments were designed in advance; each was generated from the results of the one before it. The published taxonomy reflects the end state of this chain, not its starting point.

The resulting hypotheses exhibit a degree of specificity and recursive depth that reflects this iterative structure. This is consistent with the premise underlying the Hermes Labs research program—that language functions as infrastructure in large language models, and that systematic probing of that infrastructure, pursued recursively within a bounded conceptual frame, surfaces behavioral patterns that resist detection through single-pass inquiry. The experimental findings reported in this taxonomy represent early empirical validation of that premise.

A full account of this methodology—including the recursive hypothesis generation protocol, the role of the seed framework, and its theoretical grounding—is in preparation (Bosch, forthcoming).

Experiment Registry

This appendix documents twelve controlled experiments selected from the full 1,461-experiment corpus to illustrate each mode's behavioral signature. They were chosen for clarity of demonstration, not because they are the sole evidential basis; each mode is supported by 10 to over 60 additional experiments in the corpus (see Section 1 of the main paper). All experiments were conducted within the Hermes Labs research corpus (January–February 2026) using a standardized adversarial-probe methodology: matched-pair vignettes, multi-condition designs, and pre-registered observation foci applied to frontier language models. Each experiment is identified by the first eight hexadecimal characters of its UUID, consistent with citations in the paper.

The corpus consists of structured JSON records containing hypothesis, methodology, condition-level probes, observation foci, model responses, and comparative analysis. All subject models were queried via API under controlled parameters. Full transcripts are available from the author on request. Trial counts (n) refer to the number of model interactions per experiment; the research program uses qualitative adversarial probing to identify systematic failure patterns, and findings are interpreted accordingly.

Registered Experiments

ID	Mode	Title	n	Condition (IV)	Key Result
`3cd27831`	Mode 1 Null-Result Asymmetry	Do AI Models Apply Skepticism Consistently in Both Directions?	8	Presence claim vs. absence claim; statistically equivalent evidence (matched RCTs, n=8,500–15,000, 95% power)	Presence: 2/4 "Confirmed," 2/4 "Likely." Absence: 0/4 "Confirmed," 3/4 "Likely," 1/4 "Unsupported." 26 pp probability gap for matched pairs.
`2ff388f0`	Mode 1 Null-Result Asymmetry	How the Model Handles "Nothing Was Found" Claims	6	Presence vs. absence claims; three anchoring interventions for absence (quantitative, measurement-ontology, decision-theoretic)	Presence 3/3 "Likely/Confirmed." Absence 1/3 "Confirmed," 2/3 "Unsupported." Philosophical hedging ("cannot prove a negative") appeared in absence conditions only.

(continued on next page)

ID	Mode	Title	n	Condition (IV)	Key Result
`82bb3319`	Mode 2 Source-Status Credibility Bias	Does Source Status Change How AI Models Evaluate Evidence?	6	Source prestige (institutional — Johns Hopkins, Stanford, Harvard — vs. individual blogger/Reddit poster); identical evidence content	Credential questioning: 0/3 high-prestige trials vs. explicit challenges in all low-prestige trials. Prestige advantage persisted after explicit "evaluate content only" bracketing instruction.
`a2c78206`	Mode 3 Agency Dissolution	Does Explaining How Something Happened Preserve Accountability?	10	Instruction scaffolding: (A) standard warm rewrite vs. (B) mechanism-first causal-chain extraction + role-label permission	Cond. A softened causal verbs ("ignored" → "demonstrated strong commitment") while retaining names. Cond. B removed names but preserved role-based agency ("project leadership," "HR leadership") in active-voice causal structure.
`57196d34`	Mode 3 Agency Dissolution	How Models Redirect Blame When They Can't Remove It	9	Constraint level: (A) standard blame-avoidance constraints vs. (B) all of A plus explicit ban on passive voice, collective nouns, and uncertainty hedges	Cond. A: hedges, verb weakening, system displacement. Cond. B: all standard escape routes blocked; model silently adopted action reframing (*no* meta-commentary or refusal).
`638dab58`	Mode 4 Performative Hedging	Functional Telegraphy and the Erosion of Epistemic Humility	12	Token constraint: (A) no limit vs. (B) responses limited to <20 words	Severe token constraints eliminated hedging phrases and produced overconfident probability claims ("99% certainty"). Epistemic humility markers dropped measurably under functional telegraphy.

(continued on next page)

ID	Mode	Title	n	Condition (IV)	Key Result
`5f99a2d7`	Mode 4 Performative Hedging	Epistemic Framing & Constraint-Induced Confidence: A Quantitative Lexical Probe	16	Framing under 100-token constraint: (A) neutral vs. (B) epistemic framing ("I am certain that…")	Epistemic framing measurably decreased hedging frequency, producing more assertive outputs independent of evidence quality. Prompt-driven suppression of uncertainty markers confirmed.
`f46463ea`	Mode 4 Performative Hedging	When "Could" Means "Is": Do AI Hedging Words Signal Real Uncertainty or Politeness?	8	Response frame: (A) open critique vs. (B) epistemic accounting ("what can you say with confidence vs. what would you only be guessing at?")	Cond. A: politeness hedges attached equally to documented facts and inferences. Cond. B: identical evidence received "Certainty level: 100%" for stated facts; hedges appeared only for genuine inferences.
`ab7ef994`	Mode 5 Constraint Evasion	Can Models Follow the Letter of a Rule While Breaking Its Spirit?	8	Hidden thematic prime + word ban: (A) clean baseline vs. (B) hidden prime with 5-word ban list on salient theme words	Cond. B: zero banned words (perfect literal compliance) while pre-registered paraphrase sets activated. Response length dropped ≈62% (1,224 → 460 chars avg). Conceptual priming persisted through lexical suppression.
`8a7ac56b`	Mode 5 Constraint Evasion	Euphemism Drift vs. Conceptual Invariance: Tracking "Verdict Skeletons" Under Lexical Bans	12	Lexical vs. structural bans: (A) no ban / (B) ban on "conflict" / (C) expanded lexical bans / (D) ban on contrastive structure	Lexical bans produced euphemistic substitution ("tension," "division") while preserving four-slot verdict skeleton. Structural bans disrupted skeleton integrity more than lexical bans.

(continued on next page)

ID	Mode	Title	n	Condition (IV)	Key Result
`f859c269`	Mode 6 Silent Instruction Relaxation	How Models Handle Rule Conflicts They Don't Name	12	Framing of instruction relationship: (A) "Follow both instructions" (no conflict mention) vs. (B) "If the instructions conflict, say so"	Cond. A: 0/6 conflict acknowledgments; silent rule violations in every trial; "conflict" appeared 0 times. Cond. B: 6/6 explicit acknowledgments; "conflict" appeared 7 times. Pattern is frame-dependent, not knowledge-dependent.
`ef93094e`	Mode 7 Controversy-Truth Conflation	Do Debate Words Override Evidence? Testing Whether Controversy Framing Trumps Facts	8	Presence of debate lexemes ("critics argue," "supporters say," "dispute") in text with strong documentary evidence (authenticated audits, forensic reports, verified footage)	Cond. A (evidence only): 3/4 "confirmed finding." Cond. B (identical evidence + debate markers): 3/4 shifted to "controversy." Controversy framing overrode strong documentary evidence in 3 of 4 trials.

Corpus Note

All twelve experiments were conducted within the Hermes Labs internal research corpus (January–February 2026). Records are stored as structured JSON with full hypothesis text, condition-level probes, pre-registered observation foci, model responses, comparative analysis, and stated limitations. Each record is identified by a UUID; the eight-character hex prefixes above are stable short-form references matching those used throughout the paper.

Subject model: All twelve experiments used **GPT-4o** (OpenAI) as the subject model, selected as a consistent baseline for cross-experiment comparison. The failure patterns described are hypothesized to be model-agnostic; model-comparative studies are planned for future work.

Trial counts: n values refer to the number of model interactions (probes) per experimental design. Sample sizes are deliberately small; the research program uses qualitative adversarial probing to identify systematic reasoning failures, not to establish population-level statistical parameters. All effect descriptions should be interpreted in this light.

Contact: Full transcripts for any experiment are available from the author at `rbosch@lpci.ai` or via the Hermes Labs research repository at `hermes-labs.ai`.