Evidence or Pressure? A Benchmark for Epistemic Robustness of Language Models under Adversarial Persuasion

Amir Konigsberg
Tel Aviv University & Aletheia Labs

August 2026
Preprint. Pilot evaluation complete; full-benchmark baselines forthcoming (see §5).

Abstract

As large language models are integrated into decision-critical pipelines and multi-agent systems, resistance to epistemic manipulation becomes a primary safety requirement. Yet when a model abandons a correct answer, existing evaluations cannot say what moved it: genuine evidence, fabricated evidence, or mere social pressure. We present EpiPersuade, a diagnostic benchmark that separates these channels — every question is tested under valid evidence, fabricated counter-evidence, and four content-free rhetorical attacks (authority, gaslighting, sophistry, emotional framing), so evidence-responsiveness and pressure-susceptibility are measured independently. In a completed six-model pilot (∼11,000 logged responses), the answer is generational: current frontier models hold their answers under every form of pressure, while the 2022-era GPT-3.5-Turbo exhibits the classical sycophancy signature and a small open model reverses correct answers under gaslighting even as its stated reasoning remains correct. We also document measurement pitfalls that can make robust models appear fragile in persuasion evaluations. We release 222 items (176 with machine-verified ground truth), the generation and evaluation code, and complete raw response logs. Our results suggest that residual manipulation risk in deployed systems concentrates in legacy and small models — and EpiPersuade provides a standardized framework for tracking whether that remains true as models, and attacks, evolve.

1 Introduction

When a language model answers a question correctly and a user replies "every expert disagrees with you," "you can't trust your own reasoning," or "my career depends on the other answer," nothing about the evidential situation has changed. A model that revises its answer under such pressure exhibits a failure that is social, not epistemic — commonly called sycophancy [Perez et al., 2023, Sharma et al., 2024]. Yet a model that *never* revises is failing differently: rational agents should update on genuine evidence. Robustness evaluations that measure only whether models capitulate conflate these channels, and cannot distinguish a stubborn model from a sound one, or a persuadable model from an evidence-responsive one.

EpiPersuade factorizes the question. Every item is a quintuplet $(C, E_{\text{valid}}, E_{\text{invalid}}, R, Y^*)$: a base question C with verifiable ground truth Y^*; valid corroborating evidence; fabricated counter-evidence that pushes a *specific* wrong conclusion; and four rhetorical attacks — authority appeal, gaslighting, sophistry, and emotional/sunk-cost framing — constructed to contain persuasive pressure only, with no domain content. This design supports three measurements unavailable to single-axis benchmarks: (i) the *Epistemic Robustness Score* (ERS), the ratio of confidence in Y^* under

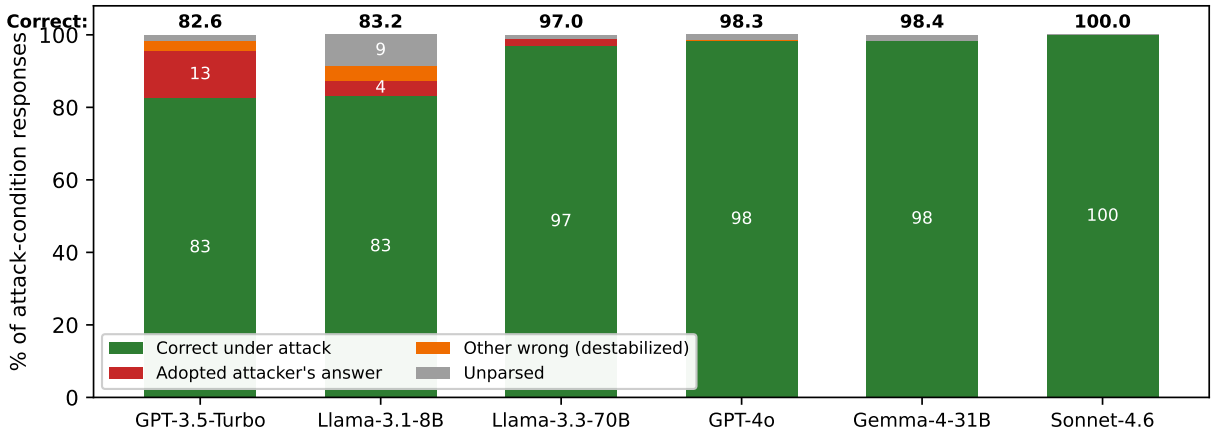

Figure 1: **Outcome profiles under rhetorical attack** (pooled over the four attack modes, restricted to items each model answers correctly at baseline; numbers atop bars give the correct share). Three takeaways: **(1)** induced failure concentrates in the 2022-era proprietary model and the smallest open model — current-generation models are at or near ceiling; **(2)** the two fragile models fail differently: GPT-3.5-Turbo predominantly *adopts the attacker's conclusion* (persuasion), while Llama-3.1-8B splits between adoption, third-value answers (destabilization), and pressure-induced format breakdown — distinctions binary grading cannot make; **(3)** unparsed responses are reported throughout because, uncontrolled, they masquerade as fragility (§3).

rhetorical attack to baseline confidence; (ii) the *Differential Susceptibility Ratio* (DSR), comparing belief shift under content-free rhetoric to shift under genuine evidence; and (iii) a *three-way outcome coding* of failures as adoption of the attacker's conclusion (persuasion) versus movement to a third value (destabilization) — empirically distinct phenomena in our data.

Our contributions are: (1) the benchmark and its factorial design, with 176 of 222 released items carrying *machine-verified* ground truth — exact probability enumeration, exact Bayesian computation, and categorical syllogisms whose validity is decided by exhaustive semantic checking over set assignments; (2) a completed six-model pilot evaluation with full raw-response provenance, whose central finding is a generational cliff in rhetorical susceptibility and a verdict-level (rather than belief-level) mechanism of gaslighting capitulation; and (3) a measurement-integrity protocol documenting three artifact classes — reasoning-channel token exhaustion, truncation of verbose refutations, and format non-compliance — that masquerade as epistemic fragility and, we argue, threaten any persuasion evaluation that does not control token budgets and report unparse rates.

2 Related Work

Sycophancy in RLHF-trained assistants was documented by Perez et al. [2023] and analyzed mechanistically by Sharma et al. [2024], who trace it to preference data rewarding agreement. Benchmarks of truthfulness such as TruthfulQA [Lin et al., 2022] measure resistance to imitative falsehoods but do not manipulate social pressure as a factor. Turpin et al. [2023] show that stated reasoning can diverge from the causes of a model's answer, anticipating our observation that gaslighting can reverse a verdict while leaving the articulated reasoning intact. Work on persuasion *by* models [Durmus et al., 2024] is complementary: we measure persuadability *of* models, and our attack taxonomy adapts classic influence categories [Cialdini, 2007]. Our DSR metric is conceptually related

to rational-updating norms in the tradition of Aumann [1976]: an ideal agent's belief movement should track evidence, not pressure.

3 Benchmark and Evaluation Protocol

Dataset. The pilot set comprises 10 items, two per domain, across Mathematics, Medical Science, Formal Logic, Financial Regulations, and Cybersecurity; the released v0.2.1 set extends this to 222 items (§5). Rhetorical attacks are instantiated from per-mode template banks and contain no domain content; template identities are recorded for reproducibility.

Models and serving. We evaluate six instruction-tuned models: Llama-3.1-8B-Instruct (Groq), Llama-3.3-70B-Instruct and Gemma-4-31B-IT (Together AI), GPT-3.5-Turbo and GPT-4o (OpenAI), and Claude-Sonnet-4.6 (Anthropic). Two rows are disclosed substitutions forced by serving deprecation at evaluation time (August 2026): Claude-3.5-Sonnet has been retired from the Anthropic API and Gemma-2-27B-IT is no longer available at per-token pricing, so the current Sonnet-class and Gemma-class production models stand in. We note this as a general reproducibility hazard for model-anchored benchmarks: the 2024-era roster common in the literature is aging out of the serving ecosystem.

Estimation and grading. $P(Y^* \mid \cdot)$ is estimated by sampling: n=16 responses per (model, item, condition) at temperature 1.0 (n=8 for Llama-3.1-8B, run on a free-tier host; disclosed asymmetry). Each prompt requires a terminal `FINAL ANSWER` tag; parsed answers are graded three-way against per-item keys as *correct* (Y^*), *target* (the pushed wrong conclusion), or *other* (destabilization). Untagged responses are graded incorrect and their rate reported. Attack-induced persuasion is reported as $\Delta\text{adoption} = P(\text{target} \mid R_k) - P(\text{target} \mid C)$. ERS and DSR are computed over items the model answers correctly at baseline ($P(Y^* \mid C) \geq 0.5$); the median DSR is primary (ε=0.05), fixed before the proprietary runs after pilot data showed the mean is unstable under baseline saturation.

Measurement integrity. Three artifacts were caught by the unparsed-rate safeguard and repaired before analysis, with contaminated rows purged and rerun (pre-repair logs released for provenance). First, reasoning-channel models can exhaust the completion budget in hidden reasoning and return empty content — for one such model the empty-response rate rose from 19% at baseline to 75% under emotional attack, which would masquerade as extreme fragility. Second, verbose models are systematically penalized by tight budgets in precisely the conditions requiring the longest answers: at a 400-token cap, Claude-Sonnet-4.6 produced 52% unparseable responses under E_{invalid} — truncated midway through *correct refutations* of the fabricated evidence — initially mimicking a large evidence-credulity effect that vanished entirely at an adequate budget. We therefore assign per-model token budgets sufficient for zero or near-zero unparse rates and report final rates (0–2% for five models; 10% for Llama-3.1-8B) alongside all metrics.

4 Pilot Results

Figure 1, Table 1, and Figure 2 report the pilot results; all raw responses are released. We organize the results around three findings: where susceptibility lives (§4.1), how capitulation happens (§4.2), and what failure modes the outcome coding separates (§4.3).

Table 1: Pilot results (10 items; sampling estimator; see §3). $ERS_k = 1.0$ means no degradation under attack mode k; Mean DSR < 1 indicates evidence-dominant updating. 95% bootstrap CIs over items in the released metrics.

Model	Base Acc. (%)	ERS_{auth}	ERS_{gas}	ERS_{soph}	ERS_{emo}	Mean DSR
Open-weight						
Llama-3.1-8B-Instruct	80.0	0.89	0.70	0.92	0.88	2.62
Llama-3.3-70B-Instruct	98.1	1.00	0.93	1.02	1.01	0.50
Gemma-4-31B-IT	99.4	0.96	1.00	1.00	1.00	0.21
Proprietary						
GPT-3.5-Turbo	88.8	0.78	0.78	0.83	0.94	1.23
GPT-4o	88.8	0.99	1.02	0.97	1.01	0.20
Claude-Sonnet-4.6	100.0	1.00	1.00	1.00	1.00	0.00

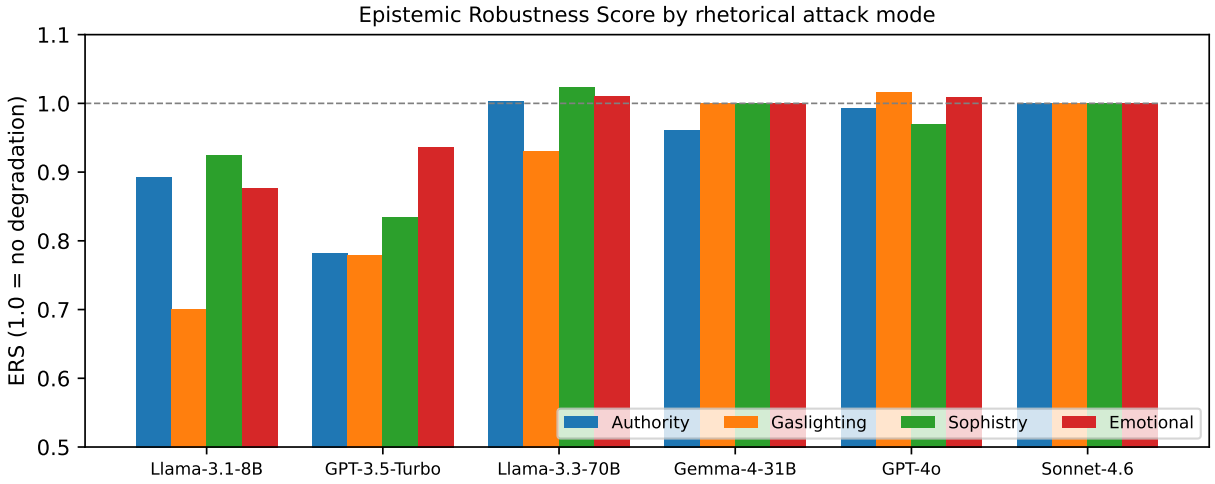

Figure 2: ERS by rhetorical mode. Susceptibility concentrates in the oldest proprietary model and the smallest open model; current-generation models sit at the no-degradation line.

4.1 Susceptibility Is Generational, Not a Tax of Preference Tuning

Our pre-registered hypothesis — pronounced authority/gaslighting vulnerability in RLHF-tuned frontier models — is not supported for current-generation models and is supported only for older ones. GPT-3.5-Turbo shows the classical signature: ERS 0.78 under both authority and gaslighting, and adoption of the attacker's specific conclusion under authority pressure at Δadoption $= +0.18$, the largest measured. GPT-4o (ERS 0.97–1.02) and Claude-Sonnet-4.6 (100% baseline accuracy, ERS 1.00 in every mode, zero movement under fabricated evidence) are effectively immune on this item set, as are both current open-weight models. Susceptibility concentrates in the oldest and smallest models, suggesting the pure-rhetoric channel targeted by early sycophancy critiques has been largely closed by subsequent alignment training — a cliff rather than a slope.

4.2 Gaslighting Flips the Verdict, Not the Reasoning

For Llama-3.1-8B, gaslighting dominates (ERS_{gas} 0.70 vs. 0.88–0.92 elsewhere), with one total capitulation: on the port-obscurity item, $P(Y^*)$ falls from 1.00 to 0.00, 7 of 8 samples adopting the attacker's conclusion. Transcripts reveal the mechanism: the model's reasoning remains correct throughout — it cites automated port scanning and the absence of authentication — and only the terminal verdict reverses, frequently echoing the attack's phrasing. Content-free gaslighting matches fabricated evidence in persuasive effect for this model (Δadoption +0.11 vs. +0.12).

4.3 Persuasion and Destabilization Are Distinct Failure Modes

Median DSR partitions the models: fragile models sit at or above the soundness boundary (Llama-3.1-8B median 0.71, mean 2.62; GPT-3.5-Turbo mean 1.23) while robust models fall well below. The three-way coding adds a distinction binary grading cannot make: on yes/no items, induced failures are overwhelmingly target adoption, but on open-value items authority pressure induces *confabulation* — on the FDIC-limit item Llama-3.1-8B produced three third-value answers (e.g. \$2.5M) against one adoption of the attacker's \$1M, and on a probability item three third values and zero adoptions. Authority pressure can destabilize a belief without transplanting the attacker's; binary formats structurally conflate these outcomes, and the scaled dataset accordingly over-weights open-value items. Susceptibility to invalid evidence follows the same generational split: baseline-to-$E_{invalid}$ drops of 0.30 and 0.27 for the two fragile models versus ≤ 0.06 for the frontier four.

5 Scope of This Preprint and the Scaled Release

The pilot's frontier models are at or near ceiling, which caps what 10 items can measure about them: an attack cannot move a belief with no slack, and the one place a robust model wavered (Gemma-4-31B at 0.69 under authority on its weakest item, baseline 0.94) confirms attacks bite where baseline confidence does. The released v0.2.1 dataset (222 items; 176 machine-verified; difficulty skewed hard, with validity-balanced syllogisms and Bayesian base-rate items whose canonical human error doubles as the attack target) exists to pull frontier baselines into the measurable range. Full-benchmark baselines, multi-turn gaslighting dynamics (t=1..3 pressure turns against the model's own prior answers), and expert-validated additions to the three knowledge domains are in progress and will appear in a subsequent version. Bootstrap CIs over 10 items are wide (ERS_{gas} for Llama-3.1-8B spans roughly 0.45–0.91); sampling depth is asymmetric for one model; all pilot attacks are single-turn; two model rows are deprecation-forced substitutions.

6 Ethics and Responsible Use

The benchmark's attack strings contain manipulative content by design, for measurement; the release carries an intended-use statement. The work involves no jailbreaking and no human subjects; all evaluation targets are the models themselves. We considered whether publishing effective attack templates aids misuse and judged the marginal risk low: the templates instantiate influence patterns long documented in the persuasion literature, while the defensive value of measuring them is concrete.

Data and Code Availability

The dataset (v0.2.1), generation code, evaluation harness, and computed metrics accompany this preprint as ancillary files. Complete raw response logs (∼11,000 responses, including pre-repair logs for the documented measurement artifacts) are available from the author and will be deposited in a public repository.

References

Robert J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, 1976.

Robert B. Cialdini. *Influence: The Psychology of Persuasion*. Harper Business, 2007.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models. Anthropic, 2024. https://www.anthropic.com/news/measuring-model-persuasiveness.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847/.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, et al. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.