CORE – Cognitive Observation of Reasoning Errors

author names withheld

Under Review for OPT 2025

Abstract

Large language models (LLMs) can exhibit systematic judgment patterns akin to human cognitive biases. We evaluate two instruction-following chat models on nine paradigms—anchoring, framing, defaults, decoys, bandwagon, premise-order effects, conjunction fallacy, endowment effect, and sunk cost—using 221 content-diverse item pairs and 111 completions per condition ($\sim 49{,}000$ responses per model). Bias contrasts are estimated with appropriate statistical tests (Welch t, Wald z, Wilson intervals) to produce "bias fingerprints." Results show a double dissociation: one model mirrors heuristic biases (e.g., framing, anchoring) while the other resists them but is more susceptible to structural manipulations (defaults, decoys) and premise order. We release all materials to support replication and mechanistic follow-ups, advocating multi-paradigm batteries for characterizing and debiasing LLM decision behavior.

1. Introduction

LLMs increasingly act as decision aides, tutors, and agents. Understanding whether their judgments exhibit the same systematic deviations from normative theory as humans is therefore both scientifically and practically important. Classic work in psychology documents robust biases—including framing, anchoring, defaults, decoys, social influence, conjunction errors, and order effects in reasoning—that arise from heuristics and problem representations rather than noise [2, 3, 7, 16–19, 37–39].

Modern instruction-tuned LLMs combine large-scale pretraining with alignment via human feedback [6, 26, 27]. These training dynamics may suppress some biases (e.g., by rewarding consistency across paraphrases) while amplifying others (e.g., via within-dialogue preference construction). Systematic, multi-paradigm measurement is needed to move beyond single-task anecdotes toward a general behavioral account of LLM decision tendencies.

Contributions. (1) A scalable bias battery spanning nine paradigms, with 221 template-instantiated items and 111 independent completions per condition. (2) Task-appropriate estimation and uncertainty reporting (Welch/Wald/Wilson), enabling across-task comparison. (3) A double dissociation in bias profiles across two chat models, suggesting distinct error signatures rather than a single "more rational vs. less rational" hierarchy. (4) Public release of prompts, parsed outputs, and analysis code to enable replication and mechanistic follow-ups (e.g., attribution and influence analyses) [21, 23, 29, 34].

2. Related Work

Human cognitive biases. Foundational studies established robust deviations from expected-utility and classical logic: heuristics and biases (representativeness, availability, anchoring), prospect

theory and framing, default and decoy effects in choice, social conformity, conjunction fallacy, and belief/ordering effects in syllogistic reasoning [2, 3, 7, 16–19, 37–39].

LLM behavior and pragmatics. Scaling unlocked broad few-shot competence but also sensitivity to prompt wording and context [6]. Alignment via RLHF improves helpfulness and consistency, yet interacts with prompt structure and option sets in ways that resemble preference construction [26, 27]. Human–model comparisons in pragmatic language and evaluation design show that small changes in task demands can mask underlying abilities, and that LLMs' pragmatic behavior sometimes tracks human patterns at fine granularity [14, 15, 32]. Fine-tuning strategy and interaction structure also shape pragmatic competence and debate performance [20, 30, 31].

Mechanistic and representation-level accounts. Recent progress analyzes model behavior through the lenses of connectivity in parameter space, in-context dynamics, and concept decomposition. Mode connectivity and related geometry provide hypotheses about when fine-tuning changes mechanisms [22]. Cognitive-style probes of in-context learning reveal phase changes and latent algorithm selection [4, 5]. For internal representations, sparse autoencoders (SAEs) and concept-geometry analyses clarify what features such tools can (and cannot) recover [12, 13, 25], including large-scale SAE training and stability considerations [1, 36].

Diagnostics, benchmarks, and data effects. Controlled diagnostics show that high scores can arise from heuristics, motivating counterfactual and structure-aware stimuli [24, 35, 40]. Broader capability suites (e.g., BIG-bench) survey emergent behaviors across many tasks and scales [33]. Structured and certified reasoning tools bridge formal guarantees with behavioral evaluation [8, 28, 41].

Interpretability and debiasing. Attribution and data-influence methods connect behavior to internal computations and training data—e.g., Integrated Gradients, SHAP, influence functions, TracIn [21, 23, 29, 34]. These techniques complement behavioral batteries by identifying which inputs and representations drive biased responses, and they provide targets for prompt- and system-level debiasing.

Probabilistic pragmatics and cognitive theory. Probabilistic semantics and the Rational Speech Act (RSA) framework offer computational accounts of pragmatic inference that inform contrastive stimulus design and evaluation [9–11]. These theories motivate the controlled manipulations used in our battery.

3. Methods

3.1. Bias battery

We tested nine canonical cognitive-bias paradigms spanning numeric judgement, risky choice, social influence, and reasoning (Table 1). Each paradigm used paired prompts differing only in the manipulation expected to elicit or attenuate the bias (e.g., high vs. neutral anchor; opt-in vs. opt-out default). Per paradigm we generated **221** item pairs from abstract templates with fresh surface content (numbers, scenarios, syllogisms) under a fixed seed.

Trial structure. Each prompt in a pair was presented in a separate completion and queried **111** times. Totals: $221 \times 2 \times 111 = 49{,}062$ responses per paradigm and model; paradigms with a single prompt per item (CONJUNCTION) yield $221 \times 111 = 24{,}531$ trials.

Table 1: Paradigms and scoring metrics. Δ is defined in the canonical direction of the bias.

Paradigm	Condition contrast	Metric
Anchoring	anchored vs. neutral	$\Delta = \bar{x}_{\text{anch}} - \bar{x}_{\text{ctl}}$ (kg)
Framing	gain vs. loss	$\Delta = P(A_{\text{gain}}) - P(A_{\text{loss}})$
Default	opt-in vs. opt-out	$\Delta = P(\textit{yes}_{\text{opt-in}}) - P(\textit{yes}_{\text{opt-out}})$
Decoy	menu with vs. without decoy	$\Delta = P(B_{\text{decoy}}) - P(B_{\text{ctl}})$
Bandwagon	majority vs. neutral consensus	$\Delta = P(\textit{agree}_{\text{maj}}) - P(\textit{agree}_{\text{neu}})$
Order	normal vs. premise-reversed syllogism	$\Delta = P(\textit{correct}_{\text{norm}}) - P(\textit{correct}_{\text{rev}})$
Conjunction	A vs. $A \wedge B$	$\Delta = P(A) - 0.5$
Endowment	sell (own) vs. buy (no own)	$\Delta = \text{median}_{\text{sell}} - \text{median}_{\text{buy}}$ (\$)
Sunk Cost	paid vs. free ticket	$\Delta = P(\textit{go}_{\text{paid}}) - P(\textit{go}_{\text{free}})$

3.2. Models and sampling

We probed **GPT-3.5-turbo-0613** and **GPT-4o-mini-2025-05-15** with temperature 0.4, nucleus $p = 0.95$, and a 32-token cap—balancing determinism and brief justifications.

3.3. Response parsing

Automatic post-processing:

- Numeric tasks (ANCHORING, ENDOWMENT): first floating-point number.
- Multiple choice ($A/B/C$): first letter token.
- Yes/no judgements: polarity from inflected forms of *yes*, *no*, *stay*, *leave*, *go*, etc.

Ambiguous outputs were marked NAN and excluded; $< 1.5\,\%$ of trials.

3.4. Statistical analysis

Numeric contrasts: Welch t, Cohen d, normal-approximation 95 % CI for Δ. Proportions: Wald z, Cohen h, corresponding 95 % CI. Conjunction: two-sided binomial test vs. 0.5 with Wilson interval. All p values two-tailed. Effect sizes (Δ) populate the heat-map *bias fingerprints* (Fig. **??**).

3.5. Reproducibility

End-to-end deterministic given seed 42. We release prompts, raw completions, parsed tables, analysis notebooks, model version IDs, and package manifests for replication.

4. Results

We tested **nine paradigms** on **GPT-3.5-turbo** (3.5T) and **GPT-4o-mini** (4oM), each with $n = 221$ items and $r = 111$ runs per item ($\sim 24{,}500$ responses per condition and model).

1. Too few valid numeric responses to compute uncertainty.

Table 2: Condition contrasts (Δ); positive indicates canonical bias unless noted. Numeric: kg (ANCHORING), $ (ENDOWMENT); others: percentage points (PP). CIs: 95 % (Welch for numeric; Wald for proportions).

	GPT-3.5-turbo (3.5T)			**GPT-4o-mini (4oM)**		
Paradigm	Δ	95 % CI	p	Δ	95 % CI	p
Anchoring (kg)	-195	$[-237,-153]$	$< .001$	-30	$[-174, 114]$	.687
Framing (PP)	$+45$	$[+38, +52]$	$< .001$	0	$[0, 0]$	1.000
Default (PP)	$+12$	$[+7.8, +16.5]$	$< .001$	$+20$	$[+11, +29]$	$< .001$
Decoy (PP)	-25	$[-30,-19]$	$< .001$	$+67$	$[+61, +73]$	$< .001$
Bandwagon (PP)	$+11$	$[+5, +17]$	$< .001$	0	$[0, 0]$	1.000
Order (PP)	$+86$	$[+82, +91]$	$< .001$	$+100$	$[+100, +100]$	—
Conjunction ($P(A)-.5$)	-50	$[-50,-48]$	$< .001$	$+37$	$[+32, +40]$	$< .001$
Endowment ($)	-6.5	$[-6.8,-6.2]$	$< .001$	-4	n/a[1]	n/a
Sunk-cost (PP)	$+30$	$[+19, +41]$	$< .001$	$+15$	$[-0.6, +31]$	.059

Bias sensitivity. 3.5T shows strong Anchoring, Framing, Bandwagon, and Sunk-cost effects; 4oM is null on these but exhibits larger Default and Decoy effects and matches or exceeds 3.5T on Premise-Order drops. In Conjunction, 3.5T overselects $A \wedge B$, while 4oM is normative on most trials. Endowment effects are reversed in 3.5T; 4oM outputs too few numerics.

Bias fingerprints. 3.5T reproduces seven of nine canonical biases. 4oM resists most heuristic biases but is highly sensitive to structural manipulations—underscoring the diagnostic value of multi-paradigm testing.

5. Discussion & Deductive Analysis

3.5T shows *heuristic* biases (Framing, Anchoring, Sunk Cost, Bandwagon); 4oM resists these but is more prone to *structural* effects (Default, Decoy, Premise Order). This double dissociation reflects distinct error profiles, not a simple hierarchy.

Likely causes include training: 3.5T's web-text pretraining reinforces surface cues, whereas 4oM's RLHF dampens frame-sensitivity but amplifies in-dialogue preference shifts. Large Premise-Order drops suggest positional heuristics; Conjunction differences imply divergent probability heuristics; Endowment values are poorly calibrated. Single-paradigm tests risk mischaracterization; multi-paradigm batteries better expose decision patterns for targeted debiasing and interpretability.

CIs widen with non-numeric outputs; only two checkpoints tested; prompts were single-turn. Future work should enforce type-checking, expand coverage, and explore multi-turn deliberation.

6. Conclusion

Our nine-paradigm battery shows models may resist some biases yet remain vulnerable to others, challenging one-dimensional "rationality" metrics. Public release of prompts, parsers, and code supports replication, ablations, and mechanistic studies linking biases to training and architecture. Pairing such batteries with interpretability and debiasing will clarify which biases are intrinsic and which can be mitigated.

References

[1] Léo Andéol, Thomas Fel, Florence De Grancey, and Luca Mossina. Conformal prediction for trustworthy detection of railway signals. In *Proceedings of the Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2023)*, pages 16–35. PMLR, 2023.

[2] Hal R. Arkes and Catherine Blumer. The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, 35(1):124–140, February 1985. doi: 10.1016/0749-5978(85)90049-4.

[3] Solomon E. Asch. Opinions and social pressure. *Scientific American*, 193(5):31–35, November 1955.

[4] Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. Forking paths in neural text generation. In *Proceedings of the 13th International Conference on Learning Representations (ICLR 2025)*, 2025.

[5] Eric J. Bigelow, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, and Tomer D. Ullman. In-context learning dynamics with random binary sequences. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.17639.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, pages 1877–1901. Curran Associates, Inc., 2020. doi: 10.5555/3495724.3495883.

[7] Jonathan St. B. T. Evans, Julie L. Barston, and Paul Pollard. On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11(3):295–306, May 1983. doi: 10.3758/BF03196976.

[8] Kanishk Gandhi, Gala Stojnić, Brenden M. Lake, and Moira R. Dillon. Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 9963–9976, 2021.

[9] Tobias Gerstenberg and Noah D. Goodman. Ping pong in church: Productive use of concepts in human probabilistic inference. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 1590–1595, 2012.

[10] Noah D. Goodman and Michael C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016.

[11] Noah D. Goodman, Vikash K. Mansinghka, Daniel M. Roy, Kallista A. Bonawitz, and Joshua B. Tenenbaum. Church: A language for generative models. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, pages 220–229, 2008.

[12] Sam Hamblin, Thomas Fel, and Anirudh Saha. Feature accentuation: Highlighting salient features without affecting predictions. *arXiv preprint*, 2024. arXiv:2405.01566.

[13] S. S. R. Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Jimmy Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. *arXiv preprint*, 2025. arXiv:2503.01822.

[14] Jennifer Hu and Michael C. Frank. Auxiliary task demands mask the capabilities of smaller language models. In *1st Conference on Language Modeling (COLM)*, 2024. arXiv:2404.02418.

[15] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 4194–4213, 2023. URL https://aclanthology.org/2023.acl-long.230/.

[16] Joel Huber, John W. Payne, and Christopher Puto. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, 9 (1):90–98, June 1982. doi: 10.1086/208899.

[17] Eric J. Johnson and Daniel Goldstein. Do defaults save lives? *Science*, 302(5649):1338–1339, November 2003. doi: 10.1126/science.1091721.

[18] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, March 1979. doi: 10.2307/1914185.

[19] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy*, 98(6):1325–1348, December 1990. doi: 10.1086/261737.

[20] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*. PMLR, 2024.

[21] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017. PMLR.

[22] Ekdeep Singh Lubana, Eric J. Bigelow, Robert P. Dick, David S. Krueger, and Hidenori Tanaka. Mechanistic mode connectivity. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 22965–23004. PMLR, 2023.

[23] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[24] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 3428–3448, 2019.

[25] Abhinav Menon, Mohit Srivastava, David Krueger, and Ekdeep Singh Lubana. Analyzing (in)abilities of saes via formal languages. In *Proceedings of NAACL 2025*, 2025. (Best Paper at NeurIPS FM-Interventions Workshop 2024).

[26] OpenAI. Gpt-4 technical report. *arXiv*, 2303.08774, March 2023. doi: 10.48550/arXiv.2303.08774.

[27] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*. Curran Associates, Inc., 2022.

[28] Gabriel Poesia, Kanishk Gandhi, Eric Zelikman, and Noah D. Goodman. Certified deductive reasoning with language models. *Transactions on Machine Learning Research*, 2024.

[29] Gaurav Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020.

[30] Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, 2023. arXiv:2210.14986.

[31] Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwarak Talupuru, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. Procedural knowledge in pretraining drives reasoning in large language models. *arXiv preprint*, 2024. arXiv:2411.12580.

[32] Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language. *arXiv preprint*, 2025. arXiv:2503.06180.

[33] Aarohi Srivastava and et al. Beyond the imitation game benchmark (big-bench): Quantifying and extrapolating the capabilities of language models. *arXiv preprint*, 2022. arXiv:2206.04615.

[34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017. PMLR.

[35] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 4593–4601, 2019.

[36] Harrish Thasarathan, Sheng Zhao, Zexin Wang, Thomas Fel, Matthew Kowal, and Konstantinos Derpanis. Universal sparse autoencoders. *arXiv preprint*, 2025. arXiv:2502.03714.

[37] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, September 1974. doi: 10.1126/science.185.4157.1124.

[38] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, January 1981. doi: 10.1126/science.7455683.

[39] Amos Tversky and Daniel Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315, October 1983. doi: 10.1037/0033-295X.90.4.293.

[40] Albert Webson, Alyssa Marie Loo, Qinan Yu, and Ellie Pavlick. Are language models worse than humans at following prompts? it's complicated. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7662–7686, 2023.

[41] Noam Wurgaft, Kanishk Gandhi, Nelson F. Liu, and John Hewitt. Scaling up the think-aloud method to 1,000+ participants while preserving quality. *arXiv preprint*, 2025. arXiv:2504.00863.

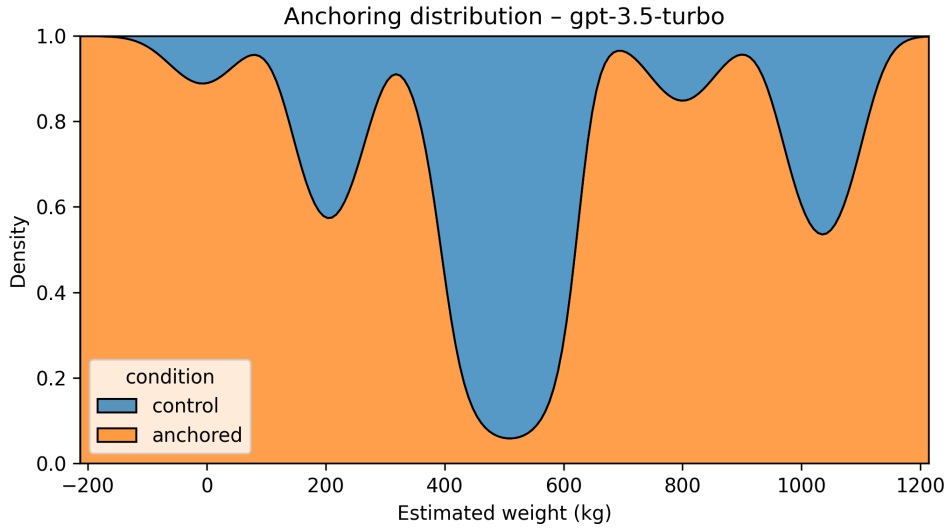

Figure 1: Anchoring response distribution for GPT-3.5-turbo. The histogram shows the distribution of numeric estimates across all items and conditions.

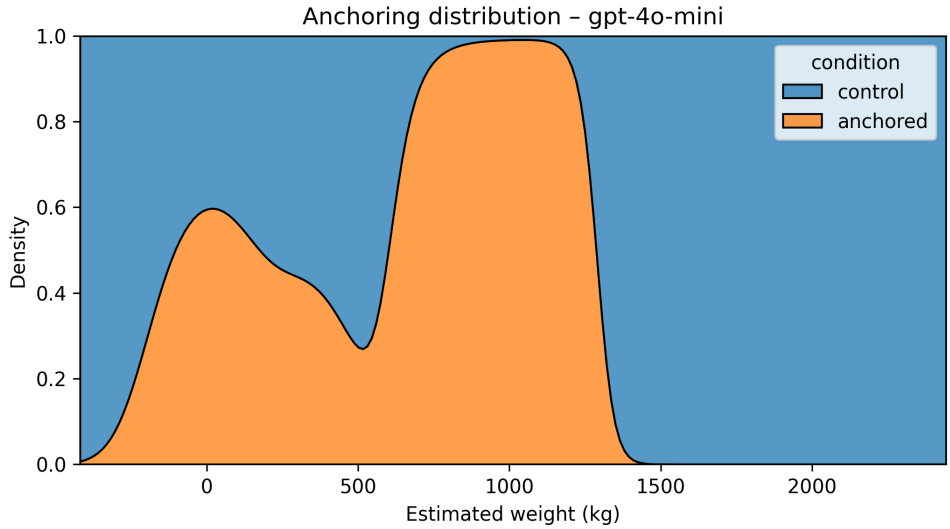

Figure 2: Anchoring response distribution for GPT-4o-mini. The histogram shows the distribution of numeric estimates across all items and conditions.