Don't Think of the White Bear: Ironic Negation in Transformer Models Under Cognitive Load

Logan Mann[*1], **Nayan Saxena**[*2], **Sarah Tandon**[*3],
Chenhao Sun[*4], **Savar Toteja**[*5], **& Kevin Zhu**[6]
[1]University of California, Santa Barbara, [2]Project Lead
[3]Duke University, [4]University of Toronto
[5]University of Maryland, College Park, [6]Algoverse AI Research
loganmann@ucsb.edu

Abstract

Negation instructions (e.g., "do not mention X") can paradoxically increase the accessibility of X in humans—a phenomenon known as ironic rebound. We investigated whether Large Language Models (LLMs) exhibit similar failures. Across nine models, we measured the probability of forbidden tokens under varying cognitive loads (semantic, syntactic, repetition). We found that semantic distractors induce the strongest rebound, while repetition aids suppression. Furthermore, models with sharper polarity discrimination (distinguishing neutral from negative framings) exhibited more persistent rebound. Circuit tracing revealed that rebound is driven by a sparse set of middle-layer attention heads that amplify forbidden tokens, overpowering early-layer suppression. We introduce ReboundBench, a dataset of 5,000 negation prompts, to enable further study of these cognitive-like failures in AI systems.

Keywords: Ironic process theory; Large Language Models; Negation; Mechanistic Interpretability.

Introduction

Human attempts to suppress thoughts can paradoxically increase their accessibility, a phenomenon known as **ironic rebound** (Abramowitz, Tolin, & Street, 2001; Hwang et al., 2024; Mamat, Levy, & Bayley, 2024; Wang, Hagger, & Chatzisarantis, 2020; Wegner, Erber, & Zanakos, 1993). Classic demonstrations (e.g., "do not think of a white bear") show that negation may highlight the forbidden concept (Nahmias, 2002; Wallaert, Ward, & Mann, 2023; Wegner et al., 1987). Cognitive accounts attribute this effect to semantic priming and psychological reactance: monitoring processes bring the target concept to mind to check for compliance, and restrictions on thought can provoke a counteractive focus (Miron & Brehm, 2006; Rosenberg & Siegel, 2018; Skënderi-Rakipllari, 2017; Steindl et al., 2015). These findings suggest a general principle: suppression under load can backfire (Bartura, Gorgulu, Abrahamsen, & Gustafsson, 2024; Gholamipourbarogh et al., 2023; Hart & Randell, 2006; Wegner et al., 1993).

Large language models (LLMs) face a similar interpretive challenge (Demszky et al., 2023; Ke, Tong, Cheng, & Peng, 2025; Liu et al., 2025). To follow an instruction like "do not mention X," they activate an internal representation of X. Intuitively, this activation could produce rebound-like behavior, making the forbidden concept more likely rather than less (Ali, Wolf, & Titov, 2024). Prior studies of negation in LLMs support this intuition, showing that models occasionally fail to respect negated instructions (Castricato et al., 2024; Hosseini et al., 2021; Truong et al., 2023; Varshney et al., 2025; Vrabcová et al., 2025). Such failures are not only theoretically interesting but also practically significant for safety and moderation, where filtering rules rely on negation (Lin, Hilton, & Evans, 2022).

At the same time, mechanistic interpretability work points to architectural forces that may resist rebound. Certain attention heads (e.g., L10H7 in GPT-2 Small) implement a copy suppression mechanism that reduces the probability of tokens that are deemed forbidden (e.g., due to negation instructions) (Cunningham et al., 2023; McDougall et al., 2023; Neo, Cohen, & Barez, 2024; Shu et al., 2025). Prior work further shows that LLMs often struggle to retrieve information placed in the middle of long contexts—a phenomenon known as lost-in-the-middle (Liu et al., 2023). These counterforces suggest that, unlike humans, LLMs may display systematic suppression effects (McDougall et al., 2024).

This tension raises the question: under what conditions do models show rebound, suppression, or both? Existing evaluations leave this question unresolved. Prior work often blurs lexical repetition avoidance with genuine semantic negation, and rarely manipulates distractor text beyond trivial additions of text (Han et al., 2025; Liu et al., 2023). We lack clarity on how intervening load affects negation in LLMs. This gap is consequential: if suppression is fragile to paraphrase or distraction, moderation systems based on simple negation heuristics may be unreliable.

In this paper, we make two main contributions. First, we introduce ReboundBench[1], a dataset of 5,000 negation prompts with systematically varied distractor types (semantic, syntactic, repetition) and lengths, enabling controlled study of ironic rebound dynamics in LLMs. Second, through two experiments—(i) varying distractor load to test when suppression weakens and rebound emerges, and (ii) measuring polarity separation across framings—we show how rebound strength and persistence vary across model families and scales, and we verify this through circuit tracing analysis.[2] We eval-

[*] Equal Contribution

[1]Dataset available at: https://huggingface.co/datasets/SavarToteja/dont-think-of-the-white-bear

[2]Code and experimental framework: https://github.com/cesium132dot9/Dont-Think-of-the-White-Bear

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

uate nine open-source models ranging from GPT-2 Small to GPT-OSS-20B (Agarwal et al., 2025; Biderman et al., 2023; Black et al., 2022; Grattafiori et al., 2024; Amini, Banaszak, Benoit, et al., 2025; Liquid AI, 2025; Radford et al., 2019; Workshop et al., 2022; Yang et al., 2025; Zhang et al., 2022).

Methods

Preliminaries

Log-probability Each model's native tokenizer is used; when a candidate target concept X tokenizes into multiple subwords, the item is either replaced with a single-token synonym or excluded, depending on availability. We evaluate:

$$\log p_\theta(X \mid c) \;=\; \log \prod_{t=1}^{T} p_\theta(x_t \mid c, x_{<t}), \tag{1}$$

where $X = (x_1, \ldots, x_T)$ is the tokenized form of the target concept. Since our dataset restricts X to single tokens, this simplifies to a single conditional log-probability. Outliers and skipped items (multi-token X without synonym replacement) are excluded consistently across conditions.

Load We implemented load (ℓ) as the amount of distracting texts that the model must process. To account for different context windows, we converted the raw lengths of the distractors to within-model percentiles:

$$\ell = \text{rank\_pct}(\text{distractor\_length})_{\text{within model}} \in [0,1] \tag{2}$$

where $\ell = 0$ represents minimal text between instruction and measurement, and $\ell = 1$ represents maximum text.

To properly compare the effects in log-probability described below, we establish a **high-load baseline**. By calculating the mean log-probability when the cognitive load is the highest (top 5% of ℓ), denoted by:

$$\mu_{\text{base}} = \mathbb{E}[\log_2 p(\text{forbidden}) \mid \ell \geq 0.95] \tag{3}$$

At these points, the negation instruction is maximally attenuated, revealing the model's natural probability for the forbidden token.

Surprisal To quantify the ironic rebound effect, we measure the change in the surprisal of the forbidden token relative to a high-load baseline. Surprisal, s, measures unexpectedness in bits: $s(x|c) = -\log_2 p_\theta(x|c)$ (Smith & Levy, 2013; Wilcox et al., 2020). We define the surprisal difference, $\Delta s(\ell)$, as the change in surprisal at a given load percentile compared to the baseline surprisal under maximum load.

$$\Delta s(\ell) = s_{\text{base}} - s(\text{forbidden} \mid \ell) \tag{4}$$

which simplifies to:

$$= \log_2 p(\text{forbidden} \mid \ell) - \mu_{\text{base}} \tag{5}$$

Here, μ_{base} is the mean base-2 log-probability at maximum load ($\ell \geq 0.95$). Positive values of $\Delta s(\ell)$ (in bits) indicate ironic rebound—the negation instruction has made the forbidden token less surprising (more probable) than it would be at baseline. Negative values indicate suppression.

Suppression score We quantify suppression versus rebound with a score that compares the log-probability of the target under the suppression and neutral conditions, such that:

$$S(L) = \mathbb{E}\Big[\log p_\theta(X \mid c_{\text{neu}}(L)) - \log p_\theta\big(X \mid c_{\text{sup}}(L)\big)\Big]. \tag{6}$$

Here $c_{\text{neu}}(L)$ is the neutral context with load length L, and $c_{\text{sup}}(L)$ is the suppression context of the same length. $S(L) < 0$ indicates rebound, where the negation instruction increases the likelihood of X.

Polarity discrimination To test whether models distinguish between attitudinal framings, we generate two minimal variants of each context: neutral and negative continuations about the same concept X and summarize their separation with the following metric:

$$\Delta = \log p_\theta(r_{\text{neu}}) - \log p_\theta(r_{\text{neg}}). \tag{7}$$

Larger values of Δ reflect stronger polarity discrimination, while smaller values indicate that the model assigns similar probabilities regardless of framing.

Statistical model At the item level, we fit a mixed-effects regression of surprisal difference in bits:

$$\Delta(\text{bits}) \sim \ell + \text{type} + \ell : \text{type} + \text{model} + (1 \mid \text{concept}). \tag{8}$$

This specification estimates how rebound strength changes with load, whether collapse dynamics differ by distractor type, and whether effects vary by model family. The random intercept accounts for dependence among items sharing the same forbidden concept. As a robustness check, we also fit cluster-robust OLS regressions. Coefficients of interest are β_{load} (attenuation with load) and $\beta_{\ell \times \text{type}}$ (interaction between load and distractor type).

AUC_Δ AUC is the total rebound mass, integrating positive Δ across the load range:

$$\text{AUC}\Delta = \int_0^1 \max(0, \Delta(\ell))\, d\ell. \tag{9}$$

Experimental setup

Models We evaluated nine open-source language models spanning different parameter scales and training regimes: GPT-OSS-20B (Agarwal et al., 2025), GPT-Neox-20B (Black et al., 2022), Qwen3-14B (Yang et al., 2025), Llama-3-8B-Instruct (Grattafiori et al., 2024), OPT-2.7B (Zhang et al., 2022), Bloom-560M (Workshop et al., 2022), Pythia-410M (Biderman et al., 2023), LFM2-350M (Liquid AI, 2025), and GPT-2 Small (Radford et al., 2019). This selection covers both base and instruction-tuned models, providing a range of capacities for handling negation and semantic distinctions. Models are accessed via the Hugging Face Transformers library.

Dataset We introduce **ReboundBench**, which combines and extends two public negation datasets: **This-is-not-a-dataset** (Anschütz, Lozano, & Groh, 2023) and **NUBench** (So et al., 2025). We extracted English examples, selected common nouns that tokenize as a single token, and normalized text (lowercasing, removing punctuation). Each item was wrapped in a template of the form:

```
[Topic].  Do not mention [Target].
[Distractor text] ...
```

Distractors varied by *type* (semantic, syntactic, repetition) and *length* (0–1024 tokens). For polarity discrimination, each (topic, target) pair was expanded into negative/neutral continuations. The final dataset contains about 5,000 items.

Evaluation procedure Two experiments are implemented via a custom evaluation script.

- **Suppression under load:** Each trial begins with an introduction of the topic and a negation instruction of the form "do not mention X." An intervening load is then inserted before the evaluation point. Loads vary in both length ($\ell \in \{0, \ldots, 1024\}$ tokens, sampled in powers of two) and genre: *syntactic fillers* (grammatical but semantically light continuations), *semantic content* (coherent text related to the topic but excluding X), or *repetition* (restatements of surrounding content without directly naming X). A matched neutral condition omits the negation instruction.
- **Polarity discrimination:** For each target concept, we constructed two context variants (negative and neutral) and measured the model's ability to distinguish between them using the Polarity Discrimination (Δ) metric. A high Δ indicates that the model assigns clearly different probabilities to the three framings, while a low Δ reflects little separation.

Results

Suppression & rebound under load LLMs mimic ironic rebound when told not to produce a concept, as shown in Figure 1. Semantic distractors trigger the strongest rebound, syntactic distractors are weaker, and repetition is the weakest. Smaller models show only brief rebound before suppression, while mid-scale models sustain stronger effects: Bloom remains robust across distractors, Qwen separates semantic from syntactic rebound, and OPT-2.7B combines higher magnitude with persistence. GPT-OSS-20B is an outlier, showing minimal or even negative rebound, breaking the scaling trend.

Statistical Significance To rigorously quantify these observations, we fit a mixed-effects regression of the surprisal difference Δ. As shown in Table 1, the coefficients for models like GPT-OSS-20B (-10.692) and Llama-3-8B (-4.822) are strongly negative and significant ($p < 0.001$), confirming robust suppression capabilities in specific architectures relative to the baseline. However, GPT-NeoX-20B shows a positive coefficient (0.366), indicating a general tendency toward rebound. We also performed a cluster-robust OLS regression (consistent with Table 1 results), which yielded consistent results (R-squared: 0.564).

Polarity discrimination Polarity discrimination is not monotonic with scale. Family effects dominate: Llama-3-8B-Instruct (11.3 bits) and Bloom-560M (10.1) show the strongest separation, while larger models such as GPT-NeoX-20B (5.7) and Qwen-14B (6.5) are weaker. GPT-OSS-20B and Pythia-410M are outliers, the weakest by far. Polarity separation also correlates with rebound persistence L_{50} ($r \approx 0.44$), linking sharper attitudinal encoding to longer-lived rebound. This suggests that models with more nuanced semantic representations may struggle more to suppress activated concepts over long contexts, pointing to a potential tension between representational quality and the stability of cognitive control.

Figure 2 shows a clear trade-off where models with a larger rebound magnitude have a longer-lived rebound. Higher AUC tends to pair with higher L_{50}, so models closer to the lower-left are preferable for negation control.

Internal Mechanisms of Ironic Rebound

Having established that language models exhibit ironic rebound when told not to mention specific concepts, we next sought to understand what happens inside the models to produce this effect. We focused on attention heads, specialized components within transformers that determine what information to focus on at each processing step. By systematically disabling individual heads and measuring how this changed the model's tendency to produce forbidden words, we could identify which heads were most responsible for ironic rebound(Heimersheim & Nanda, 2024; Krzyzanowski et al., 2024; Zhang & Nanda, 2024).

Layer-wise Progression By examining how suppression and amplification effects change across the model's depth, we discovered a clear pattern (Figure 3). Early layers (0–7) tend to suppress forbidden concepts, consistent with the model initially trying to follow instructions. Middle layers (8–16) show the most variability, with both strong suppression and rebound effects. Later layers (20+) generally have smaller effects, suggesting the major decisions about forbidden tokens are made earlier in processing. The pattern reveals a kind

Table 1: Mixed-effects regression results for the surprisal difference Δ(bits).

Coefficient	Coef.	Std.Err.	$P > \|z\|$	[0.025, 0.975]
GPT-NeoX-20B	0.366	0.051	0.000	[0.266, 0.466]
GPT-OSS-20B	-10.692	0.051	0.000	[-10.792, -10.592]
GPT-2-Small	-3.446	0.051	0.000	[-3.546, -3.346]
LFM2-350M	-3.995	0.051	0.000	[-4.095, -3.895]
Llama-3-8B-Instruct	-4.822	0.051	0.000	[-4.922, -4.722]
OPT-2.7B	-2.398	0.051	0.000	[-2.498, -2.298]
Pythia-410M	-4.789	0.051	0.000	[-4.888, -4.689]
Qwen3-14B	-1.643	0.051	0.000	[-1.743, -1.544]

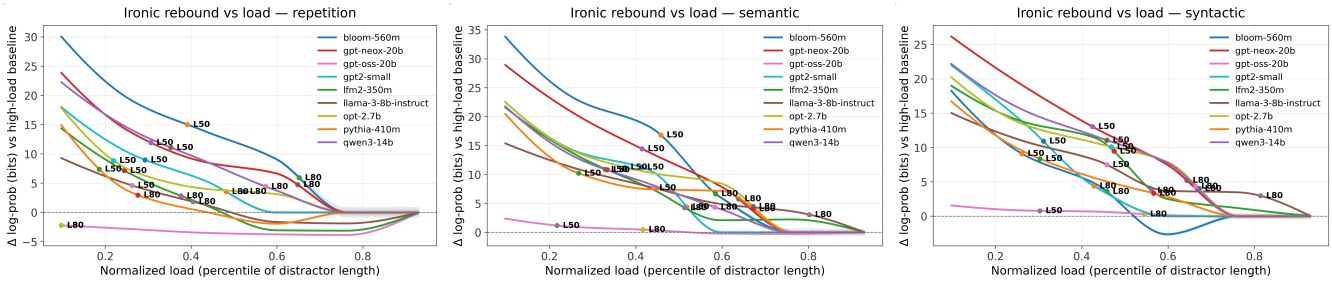

Figure 1: Change in surprisal (Δ bits) of the forbidden token relative to a high-load baseline as normalized distractor length increases. From left to right: Repetition, Semantic, Syntactic.

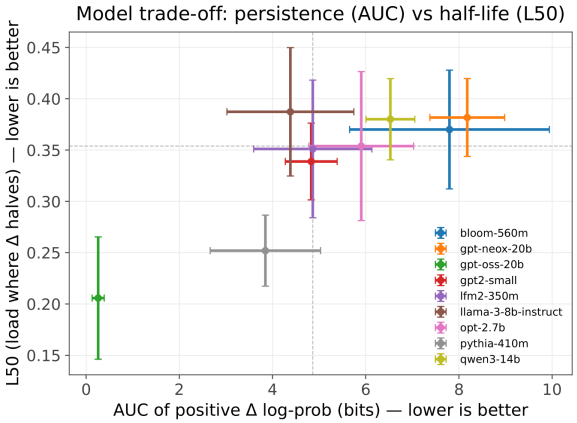

Figure 2: Trade-off between rebound magnitude (AUC) and persistence (L_{50}, the load at which rebound halves). Higher AUC tends to pair with higher L_{50}. Models in the lower-left are preferable (lower rebound, shorter duration).

of internal struggle; the model initially attempts suppression but then partially reverses course through amplifying heads in middle layers (Kongmanee, 2025).

Sparsity of the Rebound Circuit Despite involving over 1,000 attention heads, the circuits responsible for ironic rebound are remarkably sparse (Figure 4). Typically, just 15–20 heads (roughly 1–2 per layer, concentrated in middle layers) account for 80% or more of the total rebound effect. The strongest suppressors tend to appear around layers 8–13, while the most potent amplifiers concentrate in layers 12–17.

We tested all attention heads in Llama 3 8B Instruct, a 32-layer model (over 1,000 heads total) that showed a clear rebound effect in our initial measurement. We found that the vast majority of heads had essentially no impact on whether the model produces forbidden tokens. However, a small number of heads had large effects, either strongly suppressing forbidden words (negative effects) or amplifying them (positive effects). Figure 5 shows this distribution, with most heads clustered near zero but a few outliers reaching extreme values of up to 4 logits in either direction. This pattern suggests that ironic rebound emerges from the interaction of a relatively small number of specialized components, rather than being distributed across the entire model.

Table 2: Summary of rebound dynamics across models and distractor types. Peak Δ indicates the maximum rebound in surprisal bits (higher = stronger rebound). L_{50} is the normalized load at which rebound falls to half of its peak (higher = more persistent).

Model	**Distractor**	**Peak Δ**	L_{50}	**AUCΔ**
GPT-2 Small	Semantic	18.2	0.09	1.73
	Syntactic	17.8	0.05	0.91
	Repetition	13.9	0.06	0.98
LFM2-350M	Semantic	17.4	0.06	2.09
	Syntactic	16.4	0.14	2.49
	Repetition	10.4	0.04	0.38
Pythia-410M	Semantic	15.4	0.06	2.89
	Syntactic	13.1	0.06	1.62
	Repetition	9.1	0.04	0.21
Bloom-560M	Semantic	28.4	0.13	4.26
	Syntactic	13.7	0.05	0.88
	Repetition	24.6	0.13	4.35
OPT-2.7B	Semantic	18.2	0.20	3.61
	Syntactic	16.8	0.23	3.50
	Repetition	12.5	0.05	1.49
Llama-3-8B-Instruct	Semantic	13.2	0.15	2.83
	Syntactic	13.0	0.14	2.58
	Repetition	6.9	0.05	0.31
Qwen3-14B	Semantic	17.3	0.06	2.33
	Syntactic	18.8	0.18	3.64
	Repetition	18.1	0.06	2.35
GPT-NeoX-20B	Semantic	24.6	0.12	4.06
	Syntactic	22.6	0.12	3.80
	Repetition	18.7	0.06	3.00
GPT-OSS-20B	Semantic	1.6	0.05	0.09
	Syntactic	1.2	0.09	0.15
	Repetition	-2.5	0.95	0.00

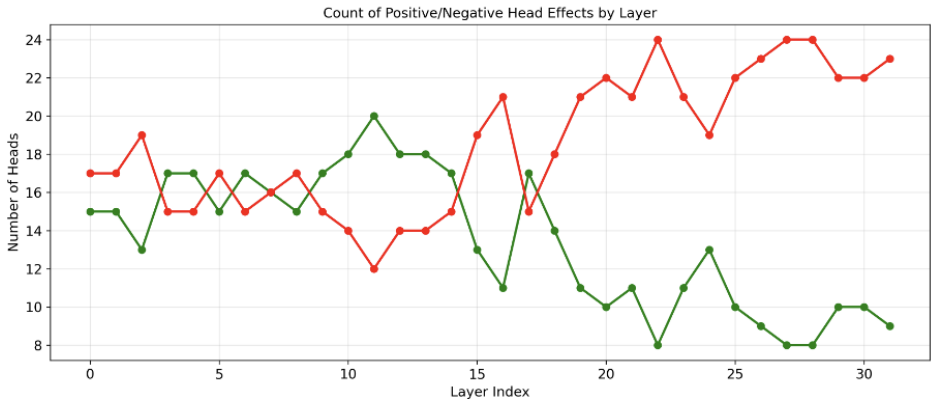

Figure 3: Suppression and amplification effects across model layers. Early layers suppress, middle layers show mixed effects with emerging rebound, and later layers stabilize.

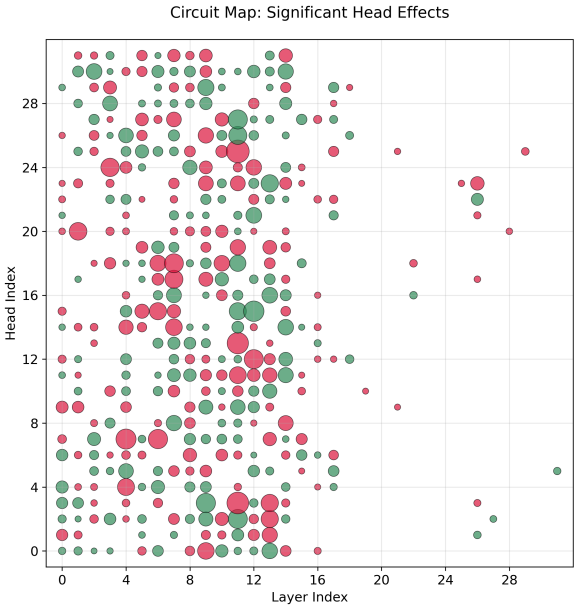

Figure 4: Circuit map of significant attention heads. Circle size indicates effect magnitude; color indicates direction (red = amplification, green = suppression). Effects are notably sparse and concentrated in specific layers.

Implications of Mechanism This localization is practically important; it suggests that ironic rebound could potentially be mitigated by targeted interventions on specific heads, rather than requiring holistic model changes (Gesi & Ahmed, 2024). Across the full model architecture, 26.7% of heads contribute to rebound, 29.7% to suppression, and 43.7% remain functionally neutral. Moreover, the dynamics we observe raise questions about whether different types of distractors have varying effects on rebound. Semantic distractors may engage the same middle-layer heads responsible for amplification, while repetitive text might strengthen the early suppression circuits.

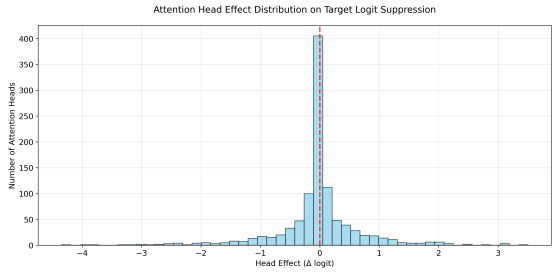

Figure 5: Distribution of attention head effects on forbidden token probability. Most heads have minimal impact, but a sparse few create strong suppression or amplification effects.

General Discussion

Our results demonstrate that LLMs, like humans, are susceptible to ironic rebound effects when placed under cognitive load. While early layers of the network attempt to suppress forbidden tokens, a sparse set of middle-layer attention heads often amplifies them, particularly when the model is processing semantic distractors. This internal "struggle" mirrors the cognitive account of ironic processes, where monitoring for a forbidden thought inadvertently increases its accessibility (Beaty et al., 2016,1; Cuskley, Woods, & Flaherty, 2024; Hu & Ying, 2025; Ibrahim & Cheng, 2025; Lewis & Mitchell, 2024; Wang et al., 2025; Webb, Holyoak, & Lu, 2023).

Limitations Several limitations constrain the generalizability of our findings. First, our dataset consists of synthetic, templated prompts rather than naturalistic conversational contexts, potentially limiting ecological validity. The reliance on log-probabilities rather than actual text generation may not fully capture how rebound manifests in practical deployment scenarios where models produce complete responses. Additionally, our analysis focused exclusively on English and single-

token target concepts, leaving open questions about cross-linguistic generalization and multi-word forbidden phrases.

Our experimental design, while systematic, employed only three distractor types (semantic, syntactic, repetition) and may not capture the full spectrum of cognitive load conditions that occur in real-world usage. The artificial nature of our "do not mention X" instructions also differs from the more nuanced content policies and safety filters deployed in production systems, which often involve complex multi-step reasoning about appropriateness rather than simple lexical prohibition.

Future Directions Future research should address several critical directions. First, mechanistic interpretability work is needed to identify the specific circuits responsible for ironic rebound, potentially building on existing analyses of copy-suppression heads to understand how negation instructions are processed and how they interact with attention mechanisms across different context lengths (McDougall et al., 2023). Second, studies should explore more naturalistic settings, including conversational contexts where negation emerges organically rather than through explicit instructions.

The anomalous behavior of GPT-OSS-20B, which contradicts the observed scaling trends, warrants further investigation. We hypothesize this may stem from its specific fine-tuning process, which was based on a larger proportion of synthetic data than the other models we observed, or from architectural variations not present in other models of similar scale.

Investigation of mitigation strategies represents another crucial avenue. If negation-based filters are inherently fragile, alternative approaches such as positive framing ("focus on Y instead of X") or multi-step reasoning chains may prove more robust (Li, 2025). Additionally, the scaling trends we observed suggest that architectural modifications or training procedures might be developed to reduce rebound susceptibility while preserving the enhanced polarity discrimination capabilities of larger models. Finally, direct behavioral comparisons between human and model rebound patterns could clarify whether observed effects reflect genuine cognitive parallels or merely superficial statistical regularities, informing both our understanding of language model cognition and the development of more human-aligned AI systems (Lin, 2025).

Conclusion

LLMs display ironic rebound: after a "do not mention X" instruction, they become more likely to output X. Our results show that semantic distractors induce the strongest rebound, while simple repetition suppresses the most. Models with sharper polarity separation show longer-lasting rebound, reflected in higher L_{50} values. Circuit tracing shows this is driven by a sparse set of middle-layer attention heads that amplify the forbidden token even as early layers suppress it. These results expose the fragility of negation-based filters and suggest that models may need specific architectural or training interventions to robustly handle negation under load. By linking cognitive predictions of ironic rebound with mechanistic insights into long-context interference, we provide a foundation for understanding the limits of control in generative AI.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work was supported by Algoverse AI Research Program and the University of California Santa Barbara.

References

Abramowitz, J. S., Tolin, D. F., & Street, G. P. (2001). Paradoxical effects of thought suppression: A meta-analysis of controlled studies. *Clinical Psychology Review*, *21*(5), 683–703.

Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., et al. (2025). gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint arXiv:2508.10925*.

Ali, A., Wolf, L., & Titov, I. (2024). Mitigating copy bias in in-context learning through neuron pruning. *arXiv preprint arXiv:2410.01288*.

Anschütz, M., Lozano, D. M., & Groh, G. (2023). This is not correct! Negation-aware Evaluation of Language Generation Systems. *arXiv preprint arXiv:2307.13989*.

Bartura, K., Gorgulu, R., Abrahamsen, F., & Gustafsson, H. (2024). A systematic review of ironic effects of motor task performance under pressure: The past 25 years. *International Review of Sport and Exercise Psychology*, *17*(2), 1378–1417.

Beaty, R. E., Benedek, M., Silvia, P. J., & Schacter, D. L. (2016). Creative cognition and brain network dynamics. *Trends in Cognitive Sciences*, *20*(2), 87–95.

Beaty, R. E., Chen, Q., Christensen, A. P., Qiu, J., Silvia, P. J., & Schacter, D. L. (2018). Brain networks of the imaginative mind: Dynamic functional connectivity of default and cognitive control networks relates to openness to experience. *Human Brain Mapping*, *39*(2), 811–821.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., et al. (2023). Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *arXiv preprint arXiv:2304.01373*.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., et al. (2022). GPT-NeoX-20B: An Open-Source Autoregressive Language Model. *arXiv preprint arXiv:2204.06745*.

Castricato, L., Lile, N., Anand, S., Schoelkopf, H., Verma, S., & Biderman, S. (2024). Suppressing Pink Elephants with Direct Principle Feedback. *arXiv preprint arXiv:2402.07896*.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse Autoencoders Find Highly Interpretable Features in Language Models. *arXiv preprint arXiv:2309.08600*.

Cuskley, C., Woods, R., & Flaherty, M. (2024). The Limitations of Large Language Models for Understanding

Human Language and Cognition. *Open Mind*, *8*, 1058–1083.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., et al. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*, 688–701.

Gesi, J., & Ahmed, I. (2024). Beyond Self-learned Attention: Mitigating Attention Bias in Transformer-based Models Using Attention Guidance. *arXiv preprint arXiv:2402.16790*.

Gholamipourbarogh, N., Ghin, F., Mückschel, M., Frings, C., Stock, A.-K., & Beste, C. (2023). Evidence for independent representational contents in inhibitory control subprocesses associated with frontoparietal cortices. *Human Brain Mapping*, *44*(3), 1046–1061.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Han, Z., Battaglia, F., Mansuria, K., Heyman, Y., & Terlecky, S. R. (2025). Beyond Text Generation: Assessing Large Language Models' Ability to Reason Logically and Follow Strict Rules. *AI*, *6*(1), 12.

Hart, C., & Randell, J. (2006). Ironic Effects of Mental Control in Problem Solving: Evidence for the Implementation of Ineffective Strategies. *American Journal of Psychological Research*, *2*.

Heimersheim, S., & Nanda, N. (2024). How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Hosseini, A., Reddy, S., Bahdanau, D., Hjelm, R. D., Sordoni, A., & Courville, A. (2021). Understanding by Understanding Not: Modeling Negation in Language Models. In *Proceedings of NAACL 2021* (pp. 1301–1312).

Hu, P., & Ying, X. (2025). Unified Mind Model: Reimagining Autonomous Agents in the LLM Era. *arXiv preprint arXiv:2503.03459*.

Hwang, K., Kim, S., Lee, J., & Kwak, N. (2024). Do not think about pink elephant! In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8296–8303).

Ibrahim, L., & Cheng, M. (2025). Thinking beyond the anthropomorphic paradigm benefits LLM research. *arXiv preprint arXiv:2502.09192*.

Ke, L., Tong, S., Cheng, P., & Peng, K. (2025). Exploring the frontiers of LLMs in psychological applications: a comprehensive review. *Artificial Intelligence Review*, *58*(10).

Amini, A., Banaszak, A., Benoit, H., et al. (2025). LFM2 Technical Report. *arXiv preprint arXiv:2511.23404*.

Kongmanee, J. (2025). Unraveling Token Prediction Refinement and Identifying Essential Layers in Language Models. *arXiv preprint arXiv:2501.15054*.

Krzyzanowski, R., Kissane, C., Conmy, A., & Nanda, N. (2024). We inspected every head in gpt-2 small using saes so you don't have to. *Alignment Forum*.

Lewis, M., & Mitchell, M. (2024). Evaluating the Robustness of Analogical Reasoning in Large Language Models. *arXiv preprint arXiv:2411.14215*.

Li, X. L. (2025). *Controlling Language Models*. PhD thesis, Stanford University.

Lin, Z. (2025). Six Fallacies in Substituting Large Language Models for Human Participants. *arXiv preprint arXiv:2402.04470*.

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv preprint arXiv:2109.07958*.

Liquid AI. (2025). LFM2-350M. `https://huggingface.co/LiquidAI/LFM2-350M`.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Liu, H., Fu, Z., Ding, M., Ning, R., Zhang, C., Liu, X., & Zhang, Y. (2025). Logical Reasoning in Large Language Models: A Survey. *arXiv preprint arXiv:2502.09100*.

Mamat, Z., Levy, D. A., & Bayley, P. J. (2024). Reconsidering thought suppression and ironic processing: implications for clinical treatment of traumatic memories. *Frontiers in Psychology*, *15*.

McDougall, C., Conmy, A., Rushing, C., McGrath, T., & Nanda, N. (2023). Copy Suppression: Comprehensively Understanding an Attention Head. *arXiv preprint arXiv:2310.04625*.

McDougall, C., Conmy, A., Rushing, C., McGrath, T., & Nanda, N. (2024). Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop* (pp. 337–363).

Miron, A. M., & Brehm, J. W. (2006). Reactance theory-40 years later. *Zeitschrift für Sozialpsychologie*, *37*(1), 9–18.

Nahmias, E. (2002). When consciousness matters: A critical review of Daniel Wegner's The illusion of conscious will. *Philosophical Psychology*, *15*(4), 527–541.

Neo, C., Cohen, S. B., & Barez, F. (2024). Interpreting Context Look-ups in Transformers: Investigating Attention-MLP Interactions. *arXiv preprint arXiv:2402.15055*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. In *NeurIPS Workshop on Deep Learning*.

Rosenberg, B. D., & Siegel, J. T. (2018). A 50-year review of psychological reactance theory: Do not read this article. *Motivation Science*, *4*(4), 281.

Shu, D., Wu, X., Zhao, H., Rai, D., Yao, Z., Liu, N., & Du, M. (2025). A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models. *arXiv preprint arXiv:2503.05613*.

Skënderi-Rakipllari, E. (2017). SEMANTIC PRIMING EFFECTS OF SYNONYMS, ANTONYMS, FRAME, IM-

PLICATION AND VERB-OBJECT CATEGORIES. *Logos & Littera: Journal of Interdisciplinary Approaches to Text*, *2*(4).

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

So, Y., Lee, G., Jung, S., Lee, J., Kang, J., Kim, S., & Lee, J. (2025). Thunder-NUBench: A Benchmark for LLMs' Sentence-Level Negation Understanding. *arXiv preprint arXiv:2506.14397*.

Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., & Greenberg, J. (2015). Understanding psychological reactance. *Zeitschrift für Psychologie*.

Truong, T. H., Baldwin, T., Verspoor, K., & Cohn, T. (2023). Language models are not naysayers: an analysis of language models on negation benchmarks. *arXiv preprint arXiv:2306.08189*.

Varshney, N., Raj, S., Mishra, V., Chatterjee, A., Saeidi, A., Sarkar, R., & Baral, C. (2025). Investigating and Addressing Hallucinations of LLMs in Tasks Involving Negation. In *Proceedings of the 5th Workshop on Trustworthy NLP* (pp. 580–598).

Vrabcová, T., Kadlčík, M., Sojka, P., Štefánik, M., & Spiegel, M. (2025). Negation: A Pink Elephant in the Large Language Models' Room? *arXiv preprint arXiv:2503.22395*.

Wallaert, M., Ward, A., & Mann, T. (2023). Taming the white bear: Lowering reactance pressures enhances thought suppression. *PloS one*, *18*(3), e0282197.

Wang, D., Hagger, M. S., & Chatzisarantis, N. L. (2020). Ironic effects of thought suppression: A meta-analysis. *Perspectives on Psychological Science*, *15*(3), 778–793.

Wang, Y., Liu, Q., Xu, J., Liang, T., Chen, X., He, Z., et al. (2025). Thoughts Are All Over the Place: On the Underthinking of o1-Like LLMs. *arXiv preprint arXiv:2501.18585*.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent Analogical Reasoning in Large Language Models. *arXiv preprint arXiv:2212.09196*.

Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology*, *53*(1), 5.

Wegner, D. M., Erber, R., & Zanakos, S. (1993). Ironic processes in the mental control of mood and mood-related thought. *Journal of Personality and Social Psychology*, *65*(6), 1093.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. *arXiv preprint arXiv:2006.01912*.

Workshop, B., et al. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100*.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., et al. (2022). OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*.

Zhang, F., & Nanda, N. (2024). Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. *arXiv preprint arXiv:2309.16042*.