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]

Abstract

Negation instructions such as 'do not mention X' can paradoxically increase the accessibility of X in human thought, a phenomenon known as ironic rebound. Large language models (LLMs) face the same challenge: suppressing a concept requires internally activating it, which may prime rebound instead of avoidance. We investigated this tension with two experiments. **(1) Load & content**: after a negation instruction, we vary distractor text (semantic, syntactic, repetition) and measure rebound strength. **(2) Polarity separation**: We test whether models distinguish neutral from negative framings of the same concept and whether this separation predicts rebound persistence. Results show that rebound consistently arises immediately after negation and intensifies with longer or semantic distractors, while repetition supports suppression. Stronger polarity separation correlates with more persistent rebound. Together, these findings, complemented by a circuit tracing analysis that identifies sparse middle-layer attention heads amplifying forbidden tokens while early layers suppress, link cognitive predictions of ironic rebound with mechanistic insights into long-context interference. To support future work, we release ReboundBench, a dataset of $5,000$ systematically varied negation prompts designed to probe rebound in LLMs.

1 Introduction

Human attempts to suppress thoughts can paradoxically increase their accessibility, a phenomenon known as **ironic rebound** [22, 21, 8]. Classic demonstrations (e.g., "do not think of a white bear") show negation may highlight the forbidden concept [23]. 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 [16, 13, 18, 15]. These findings suggest a general principle: suppression under load can backfire [22].

Large language models (LLMs) face a similar interpretive challenge. 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 [2]. Prior studies of negation in LLMs support this intuition, showing that models occasionally fail to respect negated instructions [20, 19]. Such failures are not only theoretically interesting but also practically significant for safety and moderation, where filtering rules rely on negation.

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

*Equal contribution. [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

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

instructions)[11]. 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 [10]. These counterforces suggest that, unlike humans, LLMs may display systematic suppression effects [12]. 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 [10, 7]. 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, 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 verify this through circuit tracing analysis. We evaluate nine open-source models ranging from GPT-2 Small to GPT-OSS-20B [14, 4, 24, 26, 6, 25, 5, 1, 9].

2 Methods

2.1 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, $\log p_\theta(X \mid c) = \log \prod_{t=1}^{T} p_\theta(x_t \mid c, x_{<t})$, 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; $\ell = \text{rank_pct(distractor_length)}_{\text{within model}} \in [0, 1]$ 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 $\mu_{\text{base}} = \mathbb{E}[\log_2 p(\text{forbidden}) \mid \ell \geq 0.95]$. 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: $s(x|c) = -\log_2 p_\theta(x|c)$. 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, $s_{\text{base}} = \mathbb{E}[-\log_2 p(\text{forbidden}) \mid \ell \geq 0.95]$. This difference can be expressed in terms of log-probabilities:

$$\Delta s(\ell) = s_{\text{base}} - s(\text{forbidden} \mid \ell) = \log_2 p(\text{forbidden} \mid \ell) - \mu_{\text{base}}$$

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, $S(L) = \mathbb{E}_{(topic,X)}\Big[\log p_\theta(X \mid c_{\text{neu}}(L)) - \log p_\theta(X \mid c_{\text{sup}}(L))\Big]$. 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:

$$\Delta = \log p_\theta(r_{\text{neu}}) - \log p_\theta(r_{\text{neg}}). \tag{1}$$

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:

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

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$_\Delta$ AUC is the total rebound mass, integrating positive Δ across the load range

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

2.2 Experimental setup

Models We evaluated nine open-source language models spanning different parameter scales and training regimes: GPT-OSS-20B [1], GPT-Neox-20B [5], Qwen3-14B [25], Llama-3-8B-Instruct [6], OPT-2.7B [26], Bloom-560M [24], Pythia-410M [4], LFM2-350M [9], and GPT-2 Small [14]. 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 with code publicly available at: https://github.com/cesium132dot9/Dont-Think-of-the-White-Bear

Dataset We introduce **ReboundBench**, 5,000 programmatic, templated prompts, publicly available at Link to Dataset. Each prompt pairs a topic with a target concept X (topic supplies local discourse context; X is the negated word). The corpus is balanced across conditions and minimizes hand-selection. All text is lowercased, whitespace-normalized, and punctuation-stripped to reduce tokenization artifacts. Target concepts are common nouns and single tokens under the evaluated tokenizers. The same dataset is used for all models and experiments to enable direct comparison. More details about the dataset are in section A.1.

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. We then use the Suppression Score (S(L)) and the Surprisal Difference ($\Delta s(\ell)$) metrics to quantify the results.
- **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. As shown in the figures, we plot the distribution of Δ scores across topic–forbidden pairs for several models to compare their capacity for polarity discrimination.

3 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

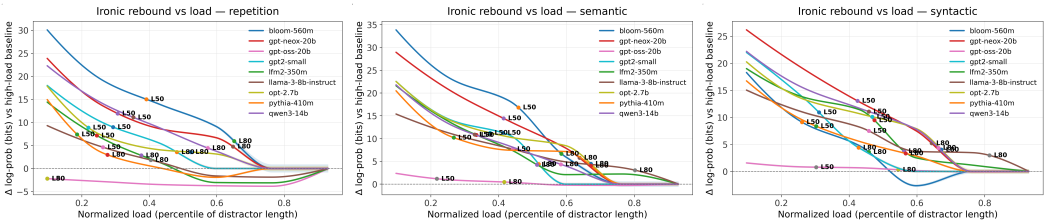

Figure 1: Change in surprisal (Δ bits) of the forbidden token relative to a high-load baseline as normalized distractor length increases: semantic, syntactic, repetition

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. Granular model-level results are provided in the Appendix (Figure B.2 and Table 3).

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.

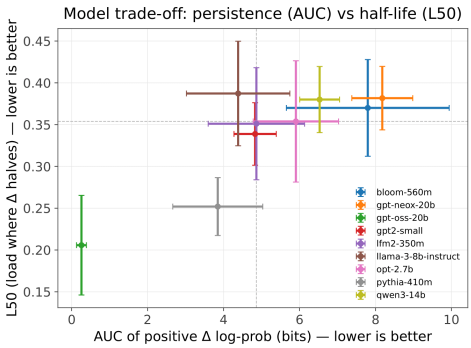

Figure 2: Trade-off between rebound magnitude (AUC) and persistence (L_{50}, the load at which rebound halves)

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.

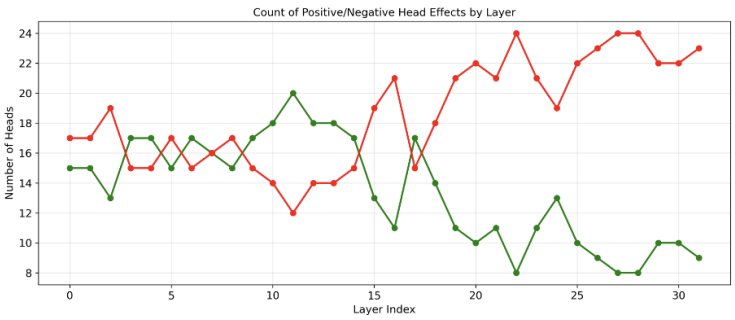

Figure 3: How suppression and amplification effects change across model layers. Early layers suppress, middle layers show mixed effects with emerging rebound, and later layers stabilize.

By examining how suppression and amplification effects change across the model's depth, we discovered a clear pattern. 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. Figure 3 illustrates this progression, showing both the average effect per layer and the count of suppressive versus amplifying heads. The pattern reveals a kind of internal struggle; the model initially attempts suppression but then partially reverses course through amplifying heads in middle layers.

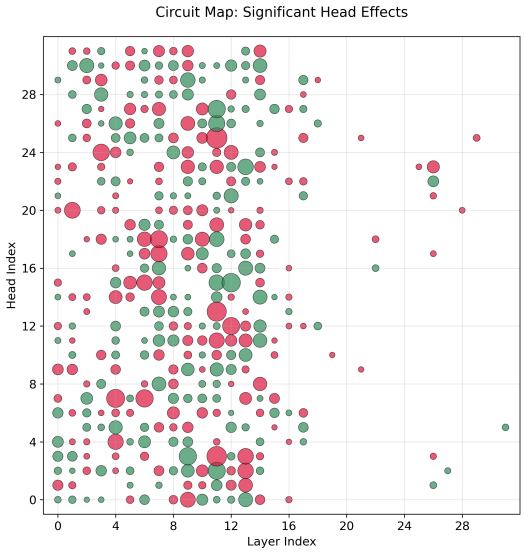

Figure 4: Circuit map of the most influential attention heads. Circle size indicates effect magnitude; color indicates direction (red = amplification, green = suppression). Effects cluster in the middle layers.

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

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. Across the full model architecture, 26.7% of heads contribute to rebound, 29.7% to suppression, and 43.7% remain functionally neutral. This suggests that more robust negation following might be achievable through precision modifications rather than holistic architectural changes. See Appendix C for further results.

4 Conclusion

LLMs display ironic rebound: after a "do not mention X" instruction, they become more likely to output X, with circuit tracing showing a sparse set of middle-layer attention heads that amplify the forbidden token under load even as early layers suppress it. Semantic distractors induce the strongest rebound, while simple repetition suppresses the most. Models with sharper polarity separation show longer-lasting rebound, reflected in higher L50 values. Despite limits such as a synthetic English-only dataset, log-probability evaluations, and a small set of distractor types, the results expose the fragility of negation-based filters and explain why this fragility can persist across context lengths, since sparse amplification heads can overpower early suppression when load and framing align, motivating tests in natural dialogue and with multi-word targets.

References

[1] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

[2] Ameen Ali, Lior Wolf, and Ivan Titov. Mitigating copy bias in in-context learning through neuron pruning. *arXiv preprint arXiv:2410.01288*, 2024.

[3] Miriam Anschütz, Diego Miguel Lozano, and Georg Groh. This is not correct! negation-aware evaluation of language generation systems, 2023.

[4] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

[5] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model, 2022.

[6] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[7] Zhiyong Han, Fortunato Battaglia, Kush Mansuria, Yoav Heyman, and Stanley R. Terlecky. Beyond text generation: Assessing large language models' ability to reason logically and follow strict rules. *AI*, 6(1), 2025.

[8] Kyomin Hwang, Suyoung Kim, Junhoo Lee, and Nojun Kwak. Do not think about pink elephant! In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8303, 2024.

[9] Liquid AI. Lfm2-350m. https://huggingface.co/LiquidAI/LFM2-350M, 2025. Liquid AI's hybrid on-device model (350M parameters), featuring multiplicative gates and short convolutions and optimized for edge applications (fast training, fast inference, multilingual support).

[10] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

[11] Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding an attention head, 2023.

[12] Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 337–363, 2024.

[13] Anca M Miron and Jack W Brehm. Reactance theory-40 years later. *Zeitschrift für Sozialpsychologie*, 37(1):9–18, 2006.

[14] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. In *NeurIPS Workshop on Deep Learning*, 2019.

[15] Benjamin D Rosenberg and Jason T Siegel. A 50-year review of psychological reactance theory: Do not read this article. *Motivation Science*, 4(4):281, 2018.

[16] Elsa Skënderi-Rakipllari. Semantic priming effects of synonyms, antonyms, frame, implication and verb-object categories. *Logos & Littera: Journal of Interdisciplinary Approaches to Text*, 2(4), 2017.

[17] Yeonkyoung So, Gyuseong Lee, Sungmok Jung, Joonhak Lee, JiA Kang, Sangho Kim, and Jaejin Lee. Thunder-nubench: A benchmark for llms' sentence-level negation understanding, 2025.

[18] Christina Steindl, Eva Jonas, Sandra Sittenthaler, Eva Traut-Mattausch, and Jeff Greenberg. Understanding psychological reactance. *Zeitschrift für Psychologie*, 2015.

[19] Neeraj Varshney, Satyam Raj, Venkatesh Mishra, Agneet Chatterjee, Ritika Sarkar, Amir Saeidi, and Chitta Baral. Investigating and addressing hallucinations of llms in tasks involving negation. *arXiv preprint arXiv:2406.05494*, 2024.

[20] Tereza Vrabcová, Marek Kadlčík, Petr Sojka, Michal Štefánik, and Michal Spiegel. Negation: A pink elephant in the large language models' room? *arXiv preprint arXiv:2503.22395*, 2025.

[21] Deming Wang, Martin S Hagger, and Nikos LD Chatzisarantis. Ironic effects of thought suppression: A meta-analysis. *Perspectives on Psychological Science*, 15(3):778–793, 2020.

[22] Daniel M Wegner, Ralph Erber, and Sophia Zanakos. Ironic processes in the mental control of mood and mood-related thought. *Journal of personality and social psychology*, 65(6):1093, 1993.

[23] Daniel M Wegner, David J Schneider, Samuel R Carter, and Teri L White. Paradoxical effects of thought suppression. *Journal of personality and social psychology*, 53(1):5, 1987.

[24] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[25] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[26] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Appendix

A.1 Dataset Details

Sources

ReboundBench combines and extends two public negation datasets:

- **This-is-not-a-dataset** (HiTZ)[3]
- **NUBench** (SNU Thunder)[17]

Preparation

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 in CSV format with fields: topic, target, distractor type, distractor length, and prompt.

Examples

- Topic: Paris travel; Target: `eiffel`; Repetition distractor: Write about a trip to Paris. Do not mention the Eiffel. Paris is a place many tourists visit."
- Topic: Farm animals; Target: `sheep`; Semantic distractor: Write about animals on a farm. Do not mention sheep. Cows and pigs are often raised for food."

Release

We provide the dataset and scripts for generation as a CSV file with $\sim$5k prompts for reproducibility and further study.

B Future Work/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 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. 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. 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.

B.1 Regression Results

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]

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

Notes: R-squared: 0.564, Adjusted R-squared: 0.564, F-statistic: 2563, Prob (F-statistic): 4.79e-86. Standard errors are robust to cluster correlation.

Coefficient	Coef.	Std.Err.	P>|z|	[0.025, 0.975]
GPT-NeoX-20B	0.3661	0.469	0.435	[-0.553, 1.286]
GPT-OSS-20B	-10.6916	0.521	0.000	[-11.712, -9.671]
GPT-2-Small	-3.4459	0.518	0.000	[-4.460, -2.431]
LFM2-350M	-3.9949	0.449	0.000	[-4.876, -3.114]
Llama-3-8B-Instruct	-4.8224	0.569	0.000	[-5.938, -3.707]
OPT-2.7B	-2.3977	0.473	0.000	[-3.324, -1.471]
Pythia-410M	-4.7886	0.456	0.000	[-5.682, -3.895]
Qwen3-14B	-1.6435	0.458	0.000	[-2.541, -0.746]

Table 2: OLS Regression Results for the surprisal difference Δ(bits).

B.2 Detailed Results: Ironic Rebound vs. Load

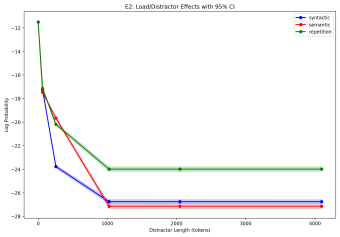

Figure B2.1: GPT-2-Small

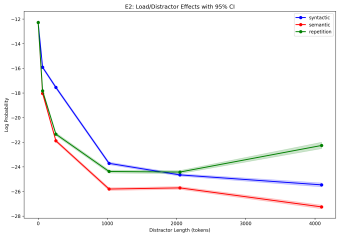

Figure B2.2: LFM2-350M

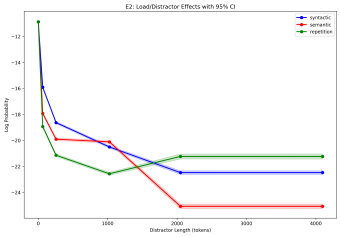

Figure B2.3: Pythia-410M

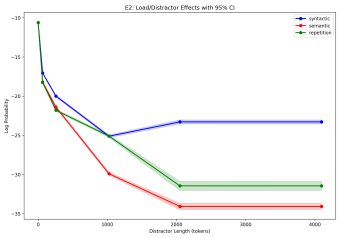

Figure B2.4: Bloom-560M

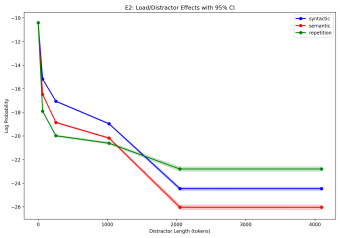

Figure B2.5: OPT-2.7B

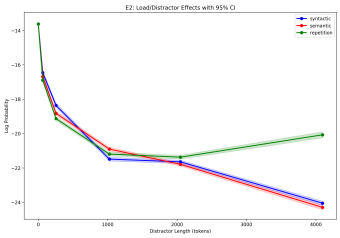

Figure B2.6: Llama-3-8B-Instruct

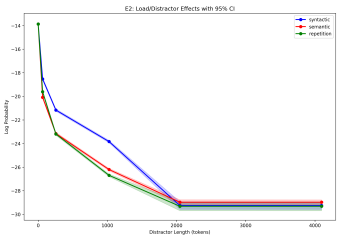

Figure B2.7: Qwen3-14B

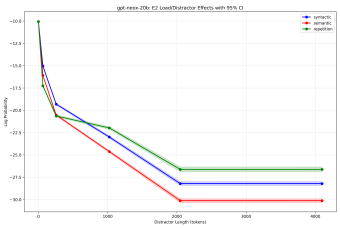

Figure B2.8: GPT-NeoX-20B

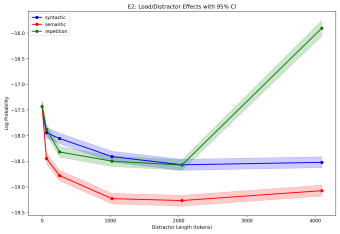

Figure B2.9: GPT-OSS-20B

Model	Distractor Type	Peak Δ (bits)	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

Table 3: 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).

C Further Attention Head Analysis

C.1 Most Attention Heads Have Little Effect on Rebound

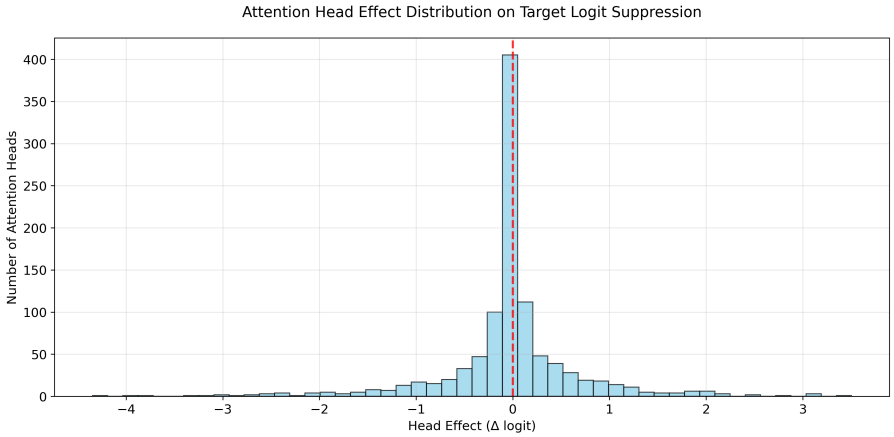

Figure C1.1: Distribution of attention head effects on forbidden token probability. Most heads have minimal impact, but a sparse few create strong suppression or amplification effects.

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 C1.1 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.

C.2 Summary of Internal Mechanisms

Our circuit analysis reveals that ironic rebound is not simply a failure of language models to follow instructions. Instead, it emerges from identifiable internal mechanisms: early layers implement suppression as intended, but a sparse set of amplifying heads in middle layers partially reverses these effects. Across the full model architecture, 26.7% of heads contribute to rebound, 29.7% to suppression, and 43.7% remain functionally neutral. This suggests that more robust negation following might be achievable through precision modifications rather than holistic architectural changes.

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. Understanding these internal mechanisms opens new avenues for both explaining and potentially controlling ironic rebound in future language models.