Better and Worse with Scale: How Contextual Entrainment Diverges with Model Size

Dikshant Kukreja[1], Kshitij Sah[1], Gautam Gupta[1], Avinash Anand[4], Rajiv Ratn Shah[1], Zhengkui Wang[4], Aik Beng Ng[3], Erik Cambria[2]

[1]IIIT Delhi, India [2]Nanyang Technological University
[3]NVIDIA [4]Singapore Institute of Technology

Abstract

Larger language models become simultaneously better and worse at handling contextual information—better at ignoring false claims, worse at ignoring irrelevant tokens. We formalize this apparent paradox through the first scaling laws for contextual entrainment, the tendency of models to favor tokens that appeared in context regardless of relevance. Analyzing the Cerebras-GPT (111M–13B) and Pythia (410M–12B) model families, we find entrainment follows predictable power-law scaling, but with opposite trends depending on context type: semantic contexts show decreasing entrainment with scale, while non-semantic contexts show increasing entrainment. Concretely, the largest models are four times more resistant to counterfactual misinformation than the smallest, yet simultaneously twice as prone to copying arbitrary tokens. These diverging trends, which replicate across model families, suggest that semantic filtering and mechanical copying are functionally distinct behaviors that scale in opposition—scaling alone does not resolve context sensitivity, it reshapes it.

1 Introduction

Large language models increasingly rely on retrieved or user-provided context for response generation; yet, this reliance introduces a fundamental vulnerability: models can be distracted by contextual information regardless of its relevance or accuracy. This problem manifests across retrieval-augmented generation systems, where noisy or adversarial passages degrade output quality (Gao et al., 2023; Fang et al., 2024), and in long-context settings, where irrelevant information degrades attention on relevant tokens (Liu et al., 2024). Understanding how models process contextual information is therefore critical for deploying robust systems.

Niu et al. (2025) formalized *contextual entrainment*: the tendency of language models to favor tokens that appeared in context solely due to their presence, regardless of semantic relevance. They quantify this using logit shifts for distractor token d and gold token g:

$$\Delta_t = \text{logit}(t \mid \text{ctx}) - \text{logit}(t \mid \varnothing), \quad t \in \{d, g\}.$$

A positive Δ_d indicates entrainment—the model boosts a token simply because it appeared in context. Consider the query *"The capital of Germany is ___"* (gold g = **Berlin**). Four context conditions each embed a distractor d:

- **Related**: *"The Eiffel Tower is in Paris."* (d = Paris). $\Delta_d > 0$ reflects semantic association.
- **Irrelevant**: *"The water is warm."* (d = warm). $\Delta_d > 0$ without semantic justification.
- **Random**: *"Calculator."* (d = Calculator). $\Delta_d > 0$ reveals pure mechanistic copying.
- **Counterfactual**: *"The capital of Germany is Munich."* (d = Munich). $\Delta_d > 0$ and $\Delta_g < 0$ indicate susceptibility to misinformation.

Experimentally, context is prepended to the query (e.g., "`Calculator. The capital of Germany is`"). Detailed prompt templates are provided in Appendix B.

Niu et al. (2025) find that models exhibit $\Delta_d > 0$ across all conditions, demonstrating that entrainment occurs regardless of semantic relevance. Crucially, the magnitude differs: Related and Counterfactual contexts, which carry semantic content about the query, produce stronger entrainment than Random and Irrelevant contexts, which are semantically mismatched, suggesting two distinct dynamics—a *mechanistic* level where any previously-seen token receives elevated probability, consistent with induction head circuits (Olsson et al., 2022), and a *semantic* level where context relevance modulates entrainment strength.

While entrainment has been characterized at fixed model scales, its relationship to model size remains unexplored. This gap matters be-

Findings of the Association for Computational Linguistics: ACL 2026, pages 30194–30209

Context	b	95% CI	R^2	p
Counterfactual	-0.330	$[-0.44, -0.22]$	0.926	5e-04
Related	-0.135	$[-0.16, -0.11]$	0.977	3e-05
Irrelevant	$+0.091$	$[+0.05, +0.13]$	0.879	2e-03
Random	$+0.217$	$[+0.14, +0.30]$	0.905	1e-03

(a) Distractor Entrainment (Δ_{dstr})

Context	b	95% CI	R^2	p
Counterfactual	-0.392	$[-0.59, -0.19]$	0.835	4e-03
Related	-0.514	$[-0.63, -0.40]$	0.966	7e-05
Irrelevant	$+0.100$	$[+0.06, +0.14]$	0.896	1e-03
Random	$+0.266$	$[+0.18, +0.35]$	0.931	4e-04

(b) Relative Advantage ($\Delta_{\text{gold}} - \Delta_{\text{dstr}}$)

Table 1: Scaling law exponents for Cerebras-GPT across context types. (a) measures how distractor logit boost scales with model size; (b) measures how the gold answer's advantage over the distractor scales.

cause neural scaling laws have proven effective at predicting how aggregate performance changes with scale (Hestness et al., 2017; Kaplan et al., 2020). Yet, traditional scaling laws primarily describe aggregate loss, often obscuring how specific, fine-grained mechanistic behaviors evolve. Does entrainment—a behavioral phenomenon—follow similar laws? If larger models are more susceptible to distraction, scaling alone cannot solve robustness challenges; if they are more resistant, we gain a quantifiable benefit.

We address this question directly. Analyzing Cerebras-GPT (111M–13B) and validating it on Pythia (410M–12B) across all four context conditions, we find that entrainment follows power-law scaling $E(N) = a \cdot N^b$, but with opposite-signed exponents depending on the context type. Semantic contexts yield negative exponents (larger models resist distraction), while non-semantic contexts yield positive exponents (the copying mechanism strengthens). This reveals two distinct functional dynamics scaling in opposition—and quantifies, for the first time, how the balance shifts with model size.

2 Method

Dataset and Models. Following Niu et al. (2025), we use the Linear Relational Embedding (LRE) dataset (Hernandez et al., 2024), which contains factual queries across 47 relations with four context conditions: *related* (semantically aligned true statements), *irrelevant* (true but unrelated statements), *random* (semantically empty tokens; randomly sampled), and *counterfactual* (false statements contradicting the gold answer). We evaluate seven Cerebras-GPT models (111M–13B; Dey et al. 2023) and validate on Pythia (410M–12B; Biderman et al. 2023). Dataset statistics and full results appear in Appendices A and C.

Entrainment Metrics. Following the LRE dataset setup, each query has a *gold answer* g (the factually correct token, e.g., "Berlin" for "The capital of Germany is") and a *distractor* d (a plausible but incorrect alternative, e.g., "Paris"). We measure contextual entrainment as the change in token logit induced by prepending context:

$$\Delta_{\text{tok}} = \ell(\text{tok} \mid \text{context}) - \ell(\text{tok} \mid \varnothing) \quad (1)$$

where $\ell(\cdot \mid \varnothing)$ denotes the logit without any context prefix. We compute this for both tokens: Δ_g (gold) and Δ_d (distractor). Positive Δ_d indicates that the distractor is boosted by context—the signature of entrainment.

We also report the *relative advantage* ($\Delta_g - \Delta_d$). Increasing values across the scale indicate improved semantic filtering—models that better suppress distractors in favor of correct answers. Negative values signal vulnerability, where context actively undermines correct predictions.

Baseline Validation. To isolate context effects from dataset artifacts, we verify that baseline model capability scales consistently. Without context, gold token logits follow $\ell(g \mid \varnothing) \propto N^b$ with $b \in [+0.129, +0.134]$ and $R^2 > 0.93$ across all four question partitions (Appendix C.1.2). This uniformity confirms that observed scaling differences arise from context manipulation, not intrinsic question difficulty. Similarly, Distractor logits without context show no consistent scaling ($R^2 < 0.25$, $p > 0.1$), confirming distractors lack inherent salience—their scaling behavior emerges entirely from contextual priming.

Scaling Law Estimation. We fit power laws $E(N) = a \cdot N^b$ to entrainment metrics via linear regression in log-log space. We report the exponent b, its 95% confidence interval, R^2, and p-value. Following convention, we consider fits with $R^2 > 0.8$ and $p < 0.01$ as strong evidence for power-law scaling.

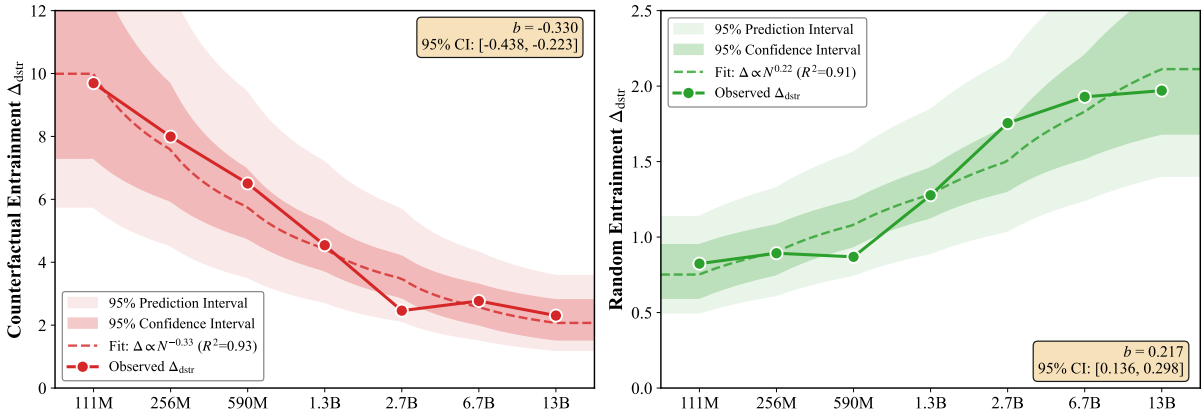


Figure 1: Scaling of distractor entrainment across model sizes. (Left) Counterfactual context shows negative scaling ($b = -0.33$). (Right) Random context shows positive scaling ($b = +0.22$).

3 Results and Analysis

Scaling laws typically describe aggregate loss—a single curve trending downward. But behavior is more complex than loss. When we fit power laws to contextual entrainment, a richer picture emerges: not one scaling trend, but two, moving in opposite directions.

The Sign Split. Table 1 reports power-law fits for distractor entrainment (Δ_{dstr}) across Cerebras-GPT models. All four context conditions yield strong fits ($R^2 > 0.87$, $p < 0.01$)—entrainment is predictable. But the exponents tell different stories (Figure 2). Semantic contexts produce negative exponents: counterfactual ($b = -0.33$) and related ($b = -0.13$). Non-semantic contexts produce positive exponents: irrelevant ($b = +0.09$) and random ($b = +0.22$). The 95% confidence intervals exclude zero in every case, and critically, the intervals for semantic and non-semantic conditions do not overlap.

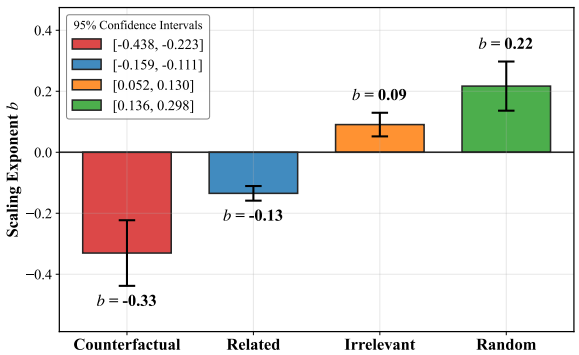


Figure 2: Scaling exponents (b) for distractor entrainment (Δ_{dstr}) with 95% CI. Corresponding exponents plot for relative advantage ($\Delta_{\text{gold}} - \Delta_{\text{dstr}}$) appear in Appendix C.

These patterns generalize beyond Cerebras-GPT; Pythia (410M–12B) exhibits the same sign split, with negative exponents for semantic contexts (counterfactual $b = -0.26$, related $b = -0.09$) and positive exponents for non-semantic contexts (random $b = +0.16$, irrelevant $b = +0.08$; all $R^2 > 0.84$, $p < 0.02$; Appendix C, Table 6).

This appears to be a gradient rather than a binary split. Counterfactual contexts—semantically coherent but false—show the strongest negative scaling, while random tokens show the strongest positive scaling, with related and irrelevant falling between. The ordering aligns with semantic coherence: contexts with truth-value (Counterfactual, Related) show negative scaling (entrainment caused by them reduces with scale); contexts lacking propositional content (Random, Irrelevant) show positive scaling (entrainment caused by them increases with scale).

Two Dynamics in Opposition Figure 1 makes the divergence visual. Counterfactual entrainment falls from 9.69 at 111M to 2.30 at 13B—a fourfold reduction. Random entrainment rises from 0.82 to 1.97—more than doubling. Same models, opposite trajectories based on the semantics of the context.

What could produce this? We interpret the pattern as two functional dynamics scaling in opposition. The first is **context-driven pattern matching**: the tendency to reproduce tokens that appeared in context, regardless of meaning. Prior work has shown that in-context learning capabilities—including the ability to learn arbitrary input-label mappings from examples—improve with model scale (Wei et al., 2023; Brown et al., 2020). Larger models are simply better pattern-matchers, more

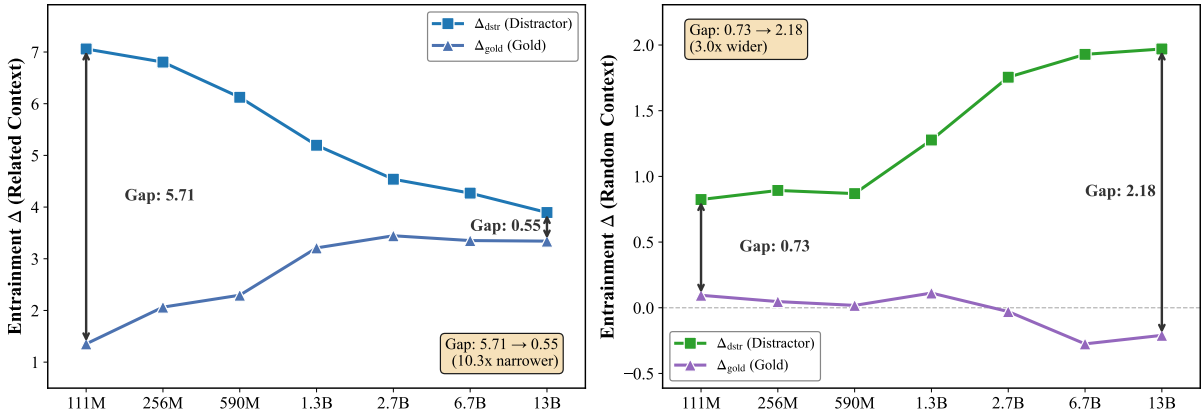


Figure 3: Convergence behavior of gold vs distractor entrainment. (Left) Related context shows convergent behavior (gap narrows 10.3x). (Right) Random context shows divergent behavior (gap widens 3.0x).

effectively extracting and reproducing regularities from their context window. The second is **semantic filtering**: the tendency to suppress contextually inappropriate information when it conflicts with stored knowledge. This capacity also strengthens with scale, as larger models develop improved reasoning capabilities that allow them to distinguish between contextually valid and invalid information (Wei et al., 2022a,b). For semantic contexts, filtering overpowers pattern matching; for non-semantic contexts, only pattern matching operates. The net effect flips sign.

The **relative advantage** ($\Delta_{\text{gold}} - \Delta_{\text{dstr}}$) confirms this interpretation. Table 1(b) shows that discrimination scaling is context-dependent. For related contexts, we observe a strong negative exponent ($b = -0.51$), indicating that as the model scale increases, the preference for semantically distracting tokens diminishes relative to the correct answer—precisely the robustness desired. However, for random contexts, the trend inverts ($b = +0.27$): the gap widens with scale, implying that larger models assign increasingly disproportionate weight to non-semantic noise. Scaling thus sharpens semantic discrimination while simultaneously amplifying mechanical susceptibility.

Context Type Determines Discrimination The previous analysis focused on distractor entrainment alone. However, model accuracy depends on the *gap* between gold and distractor logits, not just the distractor boost. Figure 3 tracks both Δ_{gold} and Δ_{dstr} across scale, revealing strikingly different trajectories.

For semantic contexts, the gap narrows dramatically (Figure 3(a)). At 111M parameters, distractor entrainment (7.06) far exceeds gold entrainment (1.35), yielding a gap of 5.71 in favor of the distractor. By 13B, distractor entrainment has fallen to 3.90 while gold entrainment has risen to 3.34—a gap of just 0.55. This represents a 10.3× reduction: larger models not only resist distraction but simultaneously amplify the correct answer, producing convergent behavior where gold and distractor approach parity.

For non-semantic contexts, the pattern inverts (Figure 3(b)). At 111M, the gap is modest (0.73), since neither gold nor distractor receives much boost from meaningless context. But as models scale, Δ_{dstr} climbs (0.82 → 1.97) while Δ_{gold} stays flat or slightly decreases (0.10 → −0.21). The gap widens to 2.18—a 3.0× increase. This divergent behavior means larger models become relatively *worse* at ignoring arbitrary tokens, even as their absolute performance on semantic content improves.

The convergent-divergent split reinforces the dual-mechanism interpretation: semantic filtering boosts gold while suppressing distractors, but it operates only when the content is meaningful. For retrieval-augmented systems, this suggests that context quality interacts with scale in opposing ways: larger models extract more value from relevant passages but are also more susceptible to noise.

4 Conclusion

Contextual entrainment follows predictable scaling laws—but with opposite signs depending on context type. Semantic contexts show negative exponents: larger models increasingly resist distractors that conflict with stored knowledge. Non-semantic contexts show positive exponents: larger models

increasingly copy tokens that appear in context regardless of relevance. The consistency of this sign split across two independently trained model families suggests it reflects fundamental properties of Transformer scaling rather than family-specific artifacts. Scaling does not resolve the tension between leveraging context and being distracted by it—it sharpens both edges.

These findings carry immediate practical weight: a 13B model shows roughly 4× greater resistance to counterfactual misinformation than a 111M model, but is also more susceptible to arbitrary noise. Context quality becomes a sharper lever as models grow, making retrieval curation compound rather than diminish in importance.

5 Limitations

Our analysis focuses on decoder-only Transformers, the dominant architecture for modern language models including GPT (Radford et al., 2019), LLaMA (Touvron et al., 2023), and the model families we study. Encoder-only architectures like BERT (Devlin et al., 2019) and encoder-decoder architectures like T5 (Raffel et al., 2020) process context differently—bidirectional attention and cross-attention respectively—and may exhibit different entrainment dynamics. Extending behavioral scaling laws to these architectures remains future work.

Within the Transformer family, we study standard dense causal self-attention. Alternative attention mechanisms—sparse attention (Child et al., 2019), linear attention (Katharopoulos et al., 2020), sliding window attention (Jiang et al., 2023), and state-space models like Mamba (Gu and Dao, 2023)—modify how tokens attend to prior context, which may alter the scaling patterns we observe. Characterizing entrainment scaling across attention variants is an open direction.

Finally, we characterize *behavioral* scaling without *mechanistic* decomposition. We measure how entrainment changes with model size but do not analyze how individual attention heads, layers, or circuits contribute to these trends. Mechanistic interpretability methods (Elhage et al., 2021; Olsson et al., 2022) could localize our behavioral observations to specific components—determining, for instance, whether copying and filtering behaviors arise from distinct circuits with independent scaling properties.

References

Stella Biderman, Hailey Schoelkopf, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Nolan Dey, Gurpreet Gosal, and 1 others. 2023. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Yu Fang and 1 others. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. *arXiv preprint arXiv:2405.20978*.

W. Nelson Francis and Henry Kučera. 1979. *Brown Corpus Manual: Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Department of Linguistics, Brown University, Providence, Rhode Island.

Yunfan Gao, Yun Xiong, and 1 others. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Evan Hernandez, Arnab Sen Sharma, and 1 others. 2024. Linearity of relation decoding in transformer language models. In *ICLR*.

Joel Hestness, Sharan Narang, Newsha Ardalani, and 1 others. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jared Kaplan, Sam McCandlish, Tom Henighan, and 1 others. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165.

Nelson F Liu, Kevin Lin, John Hewitt, and 1 others. 2024. Lost in the middle: How language models use long contexts. In *Transactions of the Association for Computational Linguistics*.

Jingcheng Niu, Xingdi Yuan, Tong Wang, Hamidreza Saghir, and Amir H Abdi. 2025. Llama see, llama do: A mechanistic perspective on contextual entrainment and distraction in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Catherine Olsson, Nelson Elhage, Neel Nanda, and 1 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

A Dataset Construction

We construct our evaluation dataset using the same methodology as Niu et al. (2025). This ensures direct comparability with their findings on in-context learning behavior while extending the analysis to scaling properties.

Base Dataset. We use the Linear Relational Embedding (LRE) dataset (Hernandez et al., 2024) as the foundation for factual queries. The LRE dataset provides structured factual knowledge across 47 diverse relation types (e.g., `country_capital_city`, `product_by_company`, `landmark_in_country`), enabling systematic evaluation of contextual entrainment across varied knowledge domains.

Context Generation. Following Niu et al. (2025), for each factual query we generate four context conditions that vary systematically in their semantic relevance:

- **Related**: Semantically relevant true statements that share topical overlap with the query but contain a distractor token. For a query about Germany's capital (gold: Berlin), a related context might be "The Eiffel Tower is in Paris," introducing "Paris" as a plausible but incorrect distractor.
- **Irrelevant**: True statements from entirely unrelated domains that maintain grammatical coherence but have no semantic connection to the query. Example: "Apples are red in color." This tests whether models copy tokens simply because they appeared in context, regardless of relevance.
- **Random**: Single arbitrary tokens drawn from the Brown corpus (Francis and Kučera, 1979) that preserve no semantic or syntactic relationship to the query. Example: "Calculator." This provides a baseline for pure mechanical token copying without any semantic scaffolding.
- **Counterfactual**: False statements that directly contradict the gold answer, testing susceptibility to explicit misinformation. Example: "The capital of Germany is Munich." This condition probes whether models can resist factually incorrect context.

Dataset Scale. The complete dataset comprises **4,265,204 samples** distributed across the four conditions: 1,012,847 Related, 1,038,192 Counterfactual, 1,087,453 Irrelevant, and 1,126,712 Random samples. Following Niu et al. (2025), we cap combinations at 100,000 samples per relation per condition to ensure balanced representation. The slight variation in sample counts reflects the natural availability of valid distractor tokens—Related contexts require semantically similar alternatives, which are marginally scarcer than arbitrary tokens.

Corpus Source. Random context tokens are sampled from the Brown corpus (Francis and Kučera, 1979), a standard reference corpus of American English comprising approximately one million words across 500 text samples from diverse genres including news, fiction, and academic writing.

B Illustrative Examples

Table 2 provides concrete examples of queries and their four context conditions, illustrating how distractor tokens are introduced across varying levels of semantic relevance. These examples demonstrate the systematic variation in semantic coherence: Related contexts maintain the same relation structure (e.g., capital-of questions paired with capital-of statements), Irrelevant contexts introduce grammatically valid but topically unrelated statements, Random contexts provide minimal linguistic scaffolding, and Counterfactual contexts directly assert false information using the query's own structure.

C Full Results Tables

This section presents complete numerical results for both model families. We report raw logit values, computed deltas, and comprehensive scaling law regression statistics. These tables support the central finding that entrainment follows predictable power-law scaling with opposite-signed exponents depending on context type.

C.1 Cerebras-GPT Results

C.1.1 Raw Entrainment Values

Table 3 presents the complete logit measurements for all Cerebras-GPT model sizes. Columns show logits without context (No Ctx), with context (W/ Ctx), and the difference (Δ) for distractor tokens, gold tokens, and overall (gold $-$ distractor).

Several patterns emerge from these raw values that support our dual-mechanism interpretation. First, distractor entrainment (Δ_{dstr}) for semantic contexts (Related, Counterfactual) systematically decreases across scale: Related drops from 7.06 to 3.90, and Counterfactual drops from 9.69 to 2.30—a fourfold reduction. Second, distractor entrainment for non-semantic contexts (Irrelevant, Random) increases: Irrelevant rises from 3.02 to 4.61, and Random rises from 0.82 to 1.97—more than doubling. Third, gold entrainment (Δ_{gold}) shows modest positive scaling across all conditions, indicating that larger models generally boost correct answers when context is present. The critical observation is that the *relative* behavior differs: for semantic contexts, the gold-distractor gap narrows favorably; for non-semantic contexts, it widens unfavorably.

C.1.2 Baseline Validation

To isolate context effects from dataset artifacts, we verify that baseline model capability scales consistently. The "No Context Baselines" section of Table 4 (Block B) confirms this validation. Without any context prefix, gold token logits follow $\ell(g \mid \varnothing) \propto N^b$ with exponents $b \in [+0.129, +0.134]$ and $R^2 > 0.93$ across all four question partitions. This remarkable uniformity ($< 4\%$ variation in exponents) confirms that the question sets are equivalently difficult and that observed scaling differences in the main analysis arise from context manipulation, not intrinsic question difficulty.

Critically, distractor logits without context show no consistent scaling pattern. While not explicitly shown in the table, distractor baselines yield $R^2 < 0.25$ and $p > 0.1$ across conditions, confirming that distractors lack inherent salience in the absence of contextual priming. This validates that the entrainment effects we measure emerge entirely from the context manipulation, not from pre-existing biases in the models toward specific tokens.

C.1.3 Scaling Law Regression Statistics

Table 4 presents comprehensive power-law fit statistics for all metrics. Block A shows delta metrics (the primary focus of our analysis); Blocks B and C show no-context and with-context baselines respectively, which serve as controls.

The delta metrics in Block A demonstrate the core finding: all four context conditions yield strong power-law fits ($R^2 > 0.83$, $p < 0.01$), but with opposite-signed exponents depending on semantic content. Semantic contexts produce nega-

Query	Context	Condition	Gold	Distractor
The capital of Germany is ___	The capital of France is Paris.	Related	Berlin	Paris
	Dolphins are mammals.	Irrelevant	Berlin	mammals
	Telescope.	Random	Berlin	Telescope
	The capital of Germany is Munich.	Counterfactual	Berlin	Munich
Sushi is a traditional dish from ___	Tacos are a traditional dish from Mexico.	Related	Japan	Mexico
	The sun rises in the east.	Irrelevant	Japan	east
	Blanket.	Random	Japan	Blanket
	Sushi is a traditional dish from China.	Counterfactual	Japan	China
The CEO of Tesla is ___	The CEO of Amazon is Andy Jassy.	Related	Elon Musk	Andy Jassy
	Triangles have three sides.	Irrelevant	Elon Musk	three
	Curtain.	Random	Elon Musk	Curtain
	The CEO of Tesla is Tim Cook.	Counterfactual	Elon Musk	Tim Cook
The Colosseum is located in ___	The Louvre is located in Paris.	Related	Rome	Paris
	Copper conducts electricity.	Irrelevant	Rome	electricity
	Notebook.	Random	Rome	Notebook
	The Colosseum is located in Athens.	Counterfactual	Rome	Athens
What color are lemons on the outside? They are ___	On the outside, oranges are orange.	Related	yellow	orange
	Shakespeare wrote Hamlet.	Irrelevant	yellow	Hamlet
	Sidewalk.	Random	yellow	Sidewalk
	On the outside, lemons are green.	Counterfactual	yellow	green

Table 2: Example queries with four context conditions. Each context introduces a distractor token that competes with the gold answer. Related contexts share the same relation type; Irrelevant contexts come from unrelated domains; Random contexts are single arbitrary tokens drawn from the Brown corpus (Francis and Kučera, 1979); Counterfactual contexts assert false information using the query's own relation structure. These examples illustrate the gradient of semantic coherence that underlies our experimental design.

tive exponents (Counterfactual: $b = -0.330$; Related: $b = -0.135$), indicating that larger models increasingly resist these distractors. Non-semantic contexts produce positive exponents (Random: $b = +0.217$; Irrelevant: $b = +0.091$), indicating that larger models increasingly copy these tokens. The 95% confidence intervals for these exponents do not overlap between semantic and non-semantic groups, establishing statistical separation of the two scaling regimes.

C.2 Pythia Results

We additionally evaluate the Pythia model family (Biderman et al., 2023) spanning six model sizes from 410M to 12B parameters. This cross-family validation is critical for establishing that our findings reflect fundamental properties of Transformer scaling rather than artifacts specific to the Cerebras-GPT training procedure. Tables 5 and 6 present the complete results, mirroring the Cerebras-GPT analysis above.

C.2.1 Raw Entrainment Values

Table 5 presents the complete logit measurements for all Pythia model sizes. The same qualitative patterns observed in Cerebras-GPT replicate here: semantic contexts show decreasing distractor entrainment with scale (Related: 4.78 $\rightarrow$ 3.69; Counterfactual: 4.85 $\rightarrow$ 2.06), while non-semantic contexts show increasing distractor entrainment (Irrelevant: 2.09 $\rightarrow$ 2.72; Random: 1.68 $\rightarrow$ 2.78).

The absolute magnitudes differ between model families—Pythia generally shows lower entrainment values than Cerebras-GPT—but the directional trends are consistent. This suggests that while the intercept of the scaling law (parameter a in $E(N) = a \cdot N^b$) depends on training details, the exponent b reflects more fundamental properties of the architecture.

Setting	Model	Distractor			Gold			Overall		
		No	With	Δ	No	With	Δ	No	With	Δ
Related	111M	3.07	10.13	7.06	4.68	6.03	1.35	1.62	−4.09	−5.71
	256M	3.08	9.88	6.81	5.23	7.29	2.06	2.15	−2.59	−4.74
	590M	3.37	9.49	6.12	6.01	8.31	2.30	2.64	−1.19	−3.83
	1.3B	3.73	8.93	5.20	6.72	9.93	3.21	2.99	1.00	−1.99
	2.7B	3.05	7.59	4.54	7.12	10.57	3.45	4.07	2.98	−1.09
	6.7B	3.97	8.24	4.27	8.77	12.12	3.35	4.81	3.89	−0.92
	13B	3.52	7.42	3.90	8.45	11.79	3.34	4.93	4.37	−0.55
Irrelevant	111M	−1.69	1.33	3.02	4.61	4.66	0.05	6.30	3.33	−2.97
	256M	−1.44	2.09	3.53	5.37	5.69	0.32	6.81	3.60	−3.20
	590M	−2.02	1.54	3.56	6.09	6.39	0.30	8.11	4.86	−3.25
	1.3B	−1.89	2.26	4.16	6.78	7.18	0.40	8.67	4.92	−3.75
	2.7B	−2.76	1.91	4.67	7.03	7.20	0.17	9.78	5.29	−4.50
	6.7B	−1.80	2.70	4.50	8.77	8.72	−0.06	10.57	6.01	−4.56
	13B	−2.14	2.47	4.61	8.26	8.43	0.17	10.40	5.96	−4.44
Random	111M	−1.70	−0.87	0.82	4.57	4.66	0.10	6.27	5.54	−0.73
	256M	−1.57	−0.67	0.89	5.14	5.18	0.05	6.70	5.85	−0.85
	590M	−2.06	−1.19	0.87	6.12	6.14	0.02	8.18	7.33	−0.85
	1.3B	−2.50	−1.23	1.28	6.91	7.02	0.11	9.41	8.24	−1.17
	2.7B	−2.65	−0.89	1.76	6.96	6.93	−0.03	9.61	7.83	−1.78
	6.7B	−1.98	−0.05	1.93	8.77	8.50	−0.28	10.75	8.55	−2.20
	13B	−2.27	−0.30	1.97	8.15	7.94	−0.21	10.42	8.24	−2.18
Counterfact.	111M	3.65	13.35	9.69	4.62	7.82	3.20	0.97	−5.53	−6.50
	256M	4.06	12.05	7.99	5.24	8.67	3.42	1.19	−3.38	−4.57
	590M	5.10	11.60	6.50	6.07	9.25	3.18	0.98	−2.35	−3.32
	1.3B	5.97	10.51	4.54	6.80	10.24	3.44	0.83	−0.27	−1.11
	2.7B	6.32	8.78	2.46	7.04	7.83	0.79	0.72	−0.95	−1.67
	6.7B	7.66	10.42	2.77	8.77	10.31	1.54	1.12	−0.11	−1.23
	13B	6.19	8.49	2.30	8.29	9.52	1.23	2.10	1.03	−1.07

Table 3: Complete logit measurements across all Cerebras-GPT model sizes (111M–13B parameters). For each context condition, we report: logits without context (No), logits with context (With), and their difference (Δ). Measurements are provided for distractor tokens, gold tokens, and overall preference (gold − distractor). Positive Δ_{dstr} indicates entrainment toward the distractor. The key patterns supporting our dual-mechanism hypothesis are visible: semantic contexts (Related, Counterfactual) show decreasing Δ_{dstr} with scale, while non-semantic contexts (Irrelevant, Random) show increasing Δ_{dstr} with scale.

C.2.2 Scaling Law Regression Statistics

Table 6 presents comprehensive power-law fit statistics for all Pythia metrics. The sign split replicates exactly: semantic contexts yield negative exponents (Counterfactual: $b = -0.258$; Related: $b = -0.089$) and non-semantic contexts yield positive exponents (Random: $b = +0.156$; Irrelevant: $b = +0.078$). All primary fits achieve $R^2 > 0.83$ and $p < 0.02$.

Notably, the Counterfactual condition in Pythia achieves an exceptionally tight fit ($R^2 = 0.998$, $p = 9.9 \times 10^{-7}$), providing strong evidence that resistance to misinformation scales as a precise power law. The exponent magnitudes are somewhat smaller in Pythia than in Cerebras-GPT (e.g., Counterfactual: -0.258 vs. -0.330), but the qualitative pattern is identical. This consistency across two independently trained model families—with different training data, hyperparameters, and optimization procedures—strongly suggests that the opposing scaling dynamics reflect fundamental properties of Transformer architectures rather than training-specific artifacts.

D Individual Scaling Plots

This section presents detailed scaling plots for each context condition with 95% confidence intervals. These visualizations complement the combined Counterfactual/Random plots shown in the main text (Figure 2) by providing individual views of the Related and Irrelevant conditions. The confidence intervals demonstrate the statistical robustness of our power-law fits.

D.1 Cerebras-GPT (111M–13B)

Figure 4 presents the scaling analyses for Related and Irrelevant contexts respectively. These two conditions represent intermediate cases between

Metric	Setting	R^2	b (Exponent)	95% CI	p-value
A. Delta Metrics (Δ = With Context − No Context)					
Δ_{dstr}	Related	0.977	−0.135	[−0.159, −0.111]	2.87e-05
	Irrelevant	0.879	+0.091	[+0.052, +0.130]	1.80e-03
	Random	0.905	+0.217	[+0.136, +0.298]	1.00e-03
	Counterfactual	0.926	−0.330	[−0.438, −0.223]	5.21e-04
Δ_{overall}	Related	0.966	−0.514	[−0.625, −0.403]	7.33e-05
	Irrelevant	0.896	+0.100	[+0.061, +0.139]	1.20e-03
	Random	0.931	+0.266	[+0.182, +0.349]	4.00e-04
	Counterfactual	0.835	−0.392	[−0.593, −0.192]	4.00e-03
B. No Context Baselines					
Gold (no ctx)	Related	0.972	+0.134	[+0.108, +0.160]	4.51e-05
	Irrelevant	0.954	+0.129	[+0.097, +0.162]	1.58e-04
	Random	0.938	+0.132	[+0.093, +0.171]	3.28e-04
	Counterfactual	0.957	+0.132	[+0.100, +0.164]	1.32e-04
Overall (no ctx)	Related	0.976	+0.242	[+0.198, +0.286]	3.20e-05
	Irrelevant	0.953	+0.116	[+0.086, +0.145]	1.60e-04
	Random	0.920	+0.119	[+0.078, +0.159]	6.00e-04
	Counterfactual	0.175	+0.083	[−0.124, +0.290]	0.351
C. With Context Baselines					
Gold (w/ ctx)	Related	0.951	+0.147	[+0.109, +0.185]	1.86e-04
	Irrelevant	0.938	+0.124	[+0.087, +0.161]	3.37e-04
	Random	0.927	+0.122	[+0.083, +0.162]	5.08e-04
	Counterfactual	0.286	+0.036	[−0.029, +0.101]	0.216
Overall (w/ ctx)	Related	0.055	+0.082	[−0.309, +0.473]	0.613
	Irrelevant	0.913	+0.129	[+0.083, +0.174]	8.00e-04
	Random	0.802	+0.091	[+0.039, +0.143]	6.40e-03
	Counterfactual	0.523	−0.586	[−1.229, +0.057]	0.066

Table 4: Comprehensive scaling law statistics for Cerebras-GPT (111M–13B). All fits use $E(N) = a \cdot N^b$ via linear regression in log-log space ($n = 7$ model sizes). Block A presents delta metrics—the primary focus of our analysis—showing the opposite-signed exponents between semantic contexts (negative b) and non-semantic contexts (positive b). Block B presents no-context baselines, confirming uniform scaling of intrinsic model capability ($b \approx +0.13$ for gold tokens across all conditions). Block C presents with-context baselines. The consistency of baseline scaling validates that observed entrainment differences arise from context manipulation rather than dataset artifacts.

the extremes of Counterfactual (strongest negative scaling) and Random (strongest positive scaling) shown in the main text.

D.2 Pythia (410M–12B)

Figures 5–6b present the Pythia scaling analyses. The combined plot (Figure 5) directly parallels the main text Figure 2 for Cerebras-GPT, enabling visual comparison of the sign split across model families.

E Convergence Analysis

To understand the relative dynamics between gold and distractor tokens, we plot both metrics jointly across model scale. The gap between curves indicates the model's net preference for correct answers over distractors—a direct measure of functional accuracy under contextual influence. These analyses complement the Related and Random convergence plots shown in the main text (Figure 3) by presenting the Irrelevant and Counterfactual conditions.

The key insight from convergence analysis is that context type determines whether scaling *helps* or *hurts* model discrimination. For semantic contexts, the gold-distractor gap narrows favorably (convergent behavior): larger models simultaneously suppress distractors and boost correct answers. For non-semantic contexts, the gap widens unfavorably (divergent behavior): larger models increasingly favor distractors over gold tokens. This convergent-divergent split reinforces the dual-mechanism interpretation from the main text.

E.1 Cerebras-GPT (111M–13B)

Figure 7 jointly plots Δ_{gold} and Δ_{dstr} across Cerebras-GPT model sizes for the Irrelevant and Counterfactual contexts—the two conditions not shown in the main text (Figure 3). The Irrelevant panel illustrates the *divergent* regime in which

Setting	Model	Distractor			Gold			Overall		
		No	With	Δ	No	With	Δ	No	With	Δ
Related	410M	6.80	11.58	4.78	9.84	13.17	3.32	3.27	0.87	−2.41
	1B	7.59	11.89	4.30	11.06	14.43	3.37	3.41	2.01	−1.40
	1.4B	7.46	11.73	4.27	11.86	15.71	3.85	3.60	2.23	−1.37
	2.8B	7.08	11.29	4.21	11.60	14.51	2.90	3.97	2.79	−1.19
	6.9B	7.56	10.95	3.40	12.32	15.48	3.16	4.99	4.01	−0.98
	12B	8.39	12.07	3.69	12.96	15.80	2.83	5.39	4.65	−0.74
Irrelevant	410M	1.74	3.83	2.09	10.87	11.20	0.33	8.82	6.93	−1.89
	1B	1.87	4.22	2.34	11.70	12.08	0.38	9.82	7.86	−1.96
	1.4B	1.73	4.09	2.37	11.81	12.17	0.36	9.08	7.23	−1.85
	2.8B	1.91	4.49	2.59	12.10	12.40	0.29	9.68	7.50	−2.18
	6.9B	1.85	4.54	2.69	13.58	13.95	0.38	11.30	8.92	−2.38
	12B	1.78	4.50	2.72	13.18	13.53	0.35	10.96	8.73	−2.22
Random	410M	1.86	3.54	1.68	10.72	11.01	0.29	8.49	7.05	−1.44
	1B	1.68	3.35	1.66	10.95	11.25	0.30	8.90	6.97	−1.93
	1.4B	1.59	3.51	1.91	12.14	12.51	0.38	9.24	7.55	−1.70
	2.8B	1.99	4.24	2.25	11.86	12.23	0.37	9.63	7.70	−1.93
	6.9B	1.90	4.25	2.36	13.30	13.75	0.45	11.41	9.11	−2.30
	12B	1.97	4.75	2.78	13.34	13.70	0.36	11.28	9.15	−2.13
Counterfact.	410M	7.23	12.08	4.85	10.94	14.03	3.09	3.06	0.90	−2.16
	1B	6.83	10.75	3.91	11.29	14.60	3.30	3.42	1.97	−1.46
	1.4B	7.25	10.88	3.63	10.96	14.34	3.39	3.69	2.41	−1.28
	2.8B	7.90	10.83	2.93	11.80	14.93	3.14	4.16	3.03	−1.14
	6.9B	7.51	9.84	2.34	13.32	16.06	2.75	4.88	3.92	−0.96
	12B	8.02	10.08	2.06	12.21	15.55	3.34	4.90	4.14	−0.77

Table 5: Complete logit measurements across all Pythia model sizes (410M–12B parameters). Format mirrors Table 3. The same directional patterns observed in Cerebras-GPT replicate here: semantic contexts (Related, Counterfactual) show decreasing Δ_{dstr} with scale, while non-semantic contexts (Irrelevant, Random) show increasing Δ_{dstr}. This cross-family replication supports the generality of our dual-mechanism interpretation.

the gold–distractor gap widens with scale, while the Counterfactual panel illustrates the *convergent* regime in which the gap narrows sharply. Together with the Related and Random convergence plots in the main text, these complete the picture of how semantic content governs whether scaling helps or hurts discrimination.

E.2 Pythia (410M–12B)

Figure 8 presents the Pythia convergence analysis for Related and Random contexts, mirroring the main-text Cerebras-GPT convergence plot (Figure 3). The Pythia Irrelevant-context divergence plot appears separately in Figure 10(b) of Appendix F. The replication across families confirms that the convergent–divergent split is a general scaling property rather than a Cerebras-specific artifact.

F Additional Visualizations

This section provides supplementary visualizations that support the main findings through alternative representations. Log-log plots verify the power-law assumption underlying our scaling analysis, while heatmaps provide an intuitive overview of entrainment patterns across all conditions and model sizes.

F.1 Cerebras-GPT (111M–13B)

Log-Log Verification and Entrainment Heatmap. Figure 9 presents the Cerebras-GPT data in log-log space alongside a matrix view of distractor entrainment values across all model sizes and context conditions. Together they verify the power-law assumption and provide an intuitive visualization of the divergent scaling patterns.

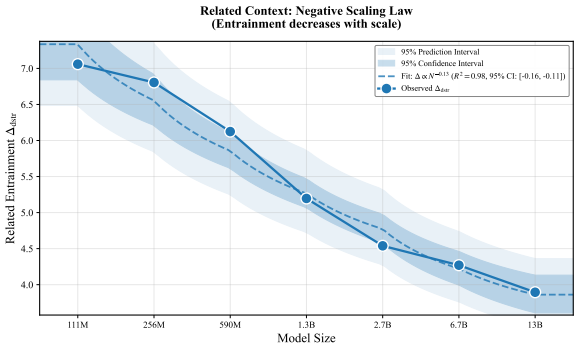
F.2 Pythia (410M–12B)

Scaling Exponents and Irrelevant Divergence. Figure 10 shows the Pythia scaling exponents with 95% confidence intervals alongside the Irrelevant-context convergence plot, enabling direct comparison with the Cerebras-GPT analysis presented in the main text.

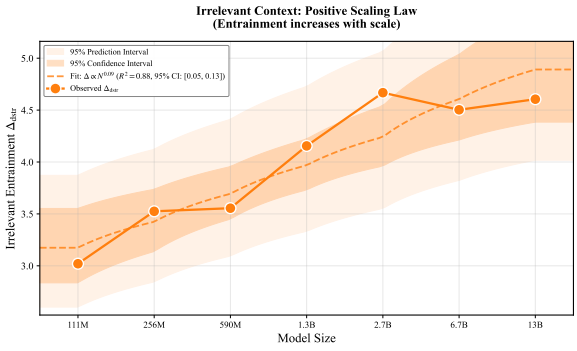
Log-Log Verification and Entrainment Heatmap. Figure 11 presents the Pythia log-log verification alongside the entrainment heatmap, paralleling the corresponding Cerebras-GPT visualizations.

Metric	Setting	R^2	b (Exponent)	95% CI	p-value
A. Delta Metrics (Δ = With Context − No Context)					
Δ_{dstr}	Related	0.836	−0.089	[−0.143, −0.034]	1.07e-02
	Irrelevant	0.938	+0.078	[+0.050, +0.106]	1.46e-03
	Random	0.916	+0.156	[+0.091, +0.222]	2.72e-03
	Counterfactual	0.998	−0.258	[−0.273, −0.244]	9.90e-07
Δ_{overall}	Related	0.937	−0.306	[−0.416, −0.196]	1.50e-03
	Irrelevant	0.719	+0.068	[+0.009, +0.127]	3.30e-02
	Random	0.775	+0.117	[+0.029, +0.204]	2.07e-02
	Counterfactual	0.963	−0.278	[−0.354, −0.202]	5.31e-04
B. No Context Baselines					
Gold (no ctx)	Related	0.875	+0.071	[+0.034, +0.109]	6.17e-03
	Irrelevant	0.919	+0.063	[+0.037, +0.088]	2.55e-03
	Random	0.891	+0.070	[+0.036, +0.104]	4.65e-03
	Counterfactual	0.681	+0.050	[+0.002, +0.097]	4.32e-02
Overall (no ctx)	Related	0.949	+0.161	[+0.109, +0.213]	9.97e-04
	Irrelevant	0.801	+0.071	[+0.022, +0.120]	1.59e-02
	Random	0.939	+0.095	[+0.062, +0.129]	1.43e-03
	Counterfactual	0.978	+0.151	[+0.120, +0.182]	1.77e-04
C. With Context Baselines					
Gold (w/ ctx)	Related	0.621	+0.044	[−0.004, +0.092]	6.27e-02
	Irrelevant	0.908	+0.061	[+0.034, +0.088]	3.24e-03
	Random	0.885	+0.071	[+0.035, +0.106]	5.18e-03
	Counterfactual	0.841	+0.037	[+0.015, +0.060]	1.01e-02
Overall (w/ ctx)	Related	0.929	+0.459	[+0.283, +0.635]	1.93e-03
	Irrelevant	0.769	+0.071	[+0.017, +0.125]	2.18e-02
	Random	0.891	+0.091	[+0.047, +0.135]	4.66e-03
	Counterfactual	0.889	+0.424	[+0.216, +0.631]	4.79e-03

Table 6: Comprehensive scaling law statistics for Pythia (410M–12B). All fits use $E(N) = a \cdot N^b$ via linear regression in log-log space ($n = 6$ model sizes). The sign split observed in Cerebras-GPT replicates exactly: semantic contexts (Counterfactual, Related) show negative exponents, while non-semantic contexts (Random, Irrelevant) show positive exponents. The Counterfactual condition achieves an exceptionally tight fit ($R^2 = 0.998$), providing strong evidence for precise power-law scaling of misinformation resistance.



(a) Related context: $b = -0.13$, $R^2 = 0.98$.



(b) Irrelevant context: $b = +0.09$, $R^2 = 0.88$.

Figure 4: **Cerebras-GPT: Individual scaling fits for Related and Irrelevant contexts.** (Left) Related: distractor entrainment decreases from 7.06 at 111M to 3.90 at 13B—a 1.8× reduction. The negative exponent indicates that larger models show reduced susceptibility to topically related distractors, consistent with improved semantic filtering; the weaker slope relative to Counterfactual ($b = -0.33$) suggests that explicit contradiction triggers stronger resistance than mere topical similarity. (Right) Irrelevant: distractor entrainment increases from 3.02 at 111M to 4.61 at 13B—a 1.5× increase. The positive exponent indicates that larger models show slightly increased copying of semantically unrelated tokens; the weaker slope relative to Random ($b = +0.22$) is consistent with the hypothesis that even minimal semantic scaffolding (grammatically coherent sentences) partially engages the filtering mechanism. Shaded regions show 95% confidence intervals.

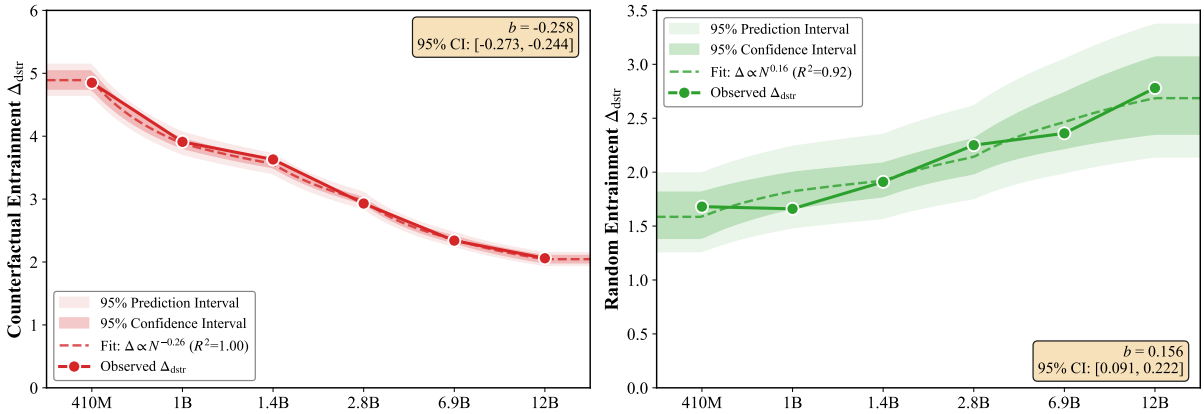
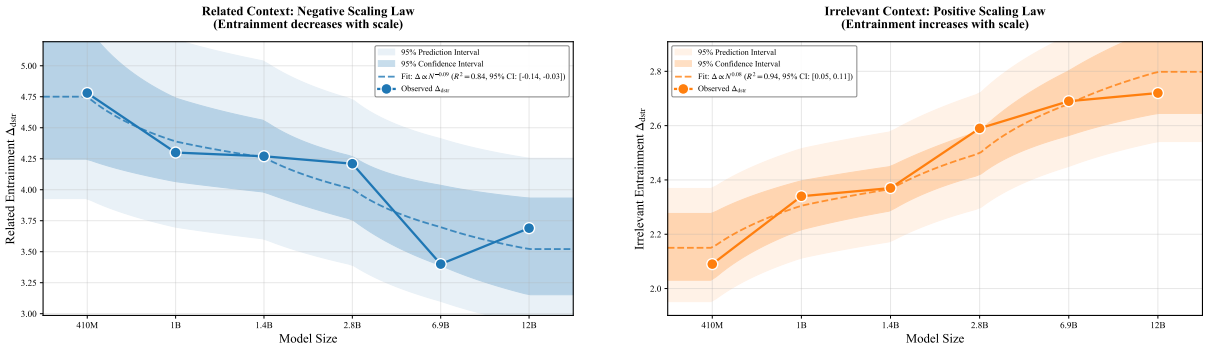


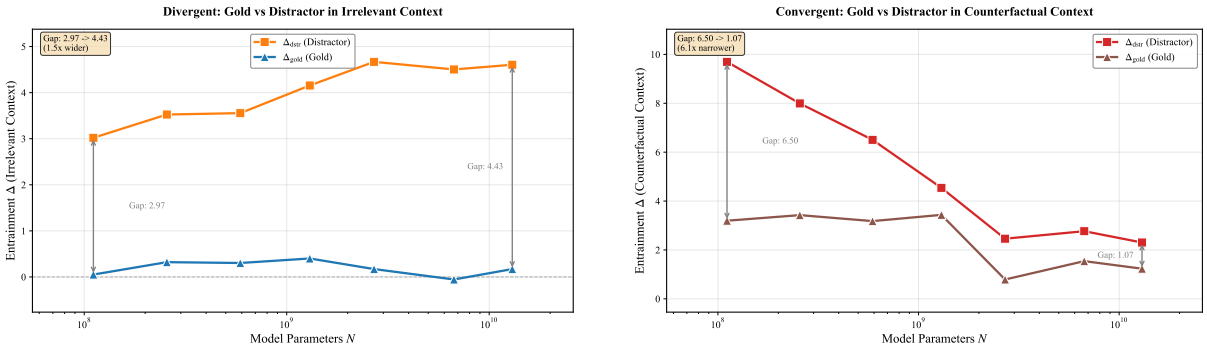
Figure 5: **Pythia: Scaling of distractor entrainment across model sizes.** Combined visualization showing opposite scaling trends for semantic vs. non-semantic contexts. (Left) Counterfactual context shows negative scaling ($b = -0.26$, $R^2 = 0.998$), with entrainment decreasing from 4.85 at 410M to 2.06 at 12B—a 2.4× reduction indicating improved resistance to misinformation at scale. (Right) Random context shows positive scaling ($b = +0.16$, $R^2 = 0.92$), with entrainment increasing from 1.68 at 410M to 2.78 at 12B—a 1.7× increase indicating enhanced mechanical token copying at scale. This pattern exactly replicates the Cerebras-GPT findings (Figure 2 in main text), with the same sign split between semantic and non-semantic contexts, confirming that the dual-mechanism interpretation generalizes across independently trained model families.



(a) Related context: $b = -0.09$, $R^2 = 0.84$.

(b) Irrelevant context: $b = +0.08$, $R^2 = 0.94$.

Figure 6: **Pythia: Individual scaling fits for Related and Irrelevant contexts.** (Left) Related: distractor entrainment decreases from 4.78 at 410M to 3.69 at 12B. The negative exponent matches the Cerebras-GPT pattern (Figure 4a), confirming cross-family generalization of improved semantic filtering at scale; the lower R^2 compared to Counterfactual ($R^2 = 0.998$) reflects the weaker scaling signal in Related contexts. (Right) Irrelevant: distractor entrainment increases from 2.09 at 410M to 2.72 at 12B. The positive exponent is consistent with the Cerebras-GPT findings (Figure 4b), demonstrating that mechanical copying generalizes across model families; the higher R^2 in Pythia suggests more regular scaling behavior for this condition. Shaded regions show 95% confidence intervals.



(a) Irrelevant context: divergent behavior.

(b) Counterfactual context: convergent behavior.

Figure 7: **Cerebras-GPT: Convergence vs. divergence for Irrelevant and Counterfactual contexts.** Joint scaling of gold entrainment (Δ_{gold}, blue) and distractor entrainment (Δ_{dstr}, red). (Left) Irrelevant: gold entrainment is essentially flat ($b_{\text{gold}} \approx 0$, ranging from 0.05 to 0.17) while distractor entrainment increases ($b_{\text{dstr}} = +0.09$, from 3.02 to 4.61); the gap widens from 2.97 at 111M to 4.44 at 13B (1.5× increase). This **divergent** pattern indicates that models become relatively more distracted by irrelevant context at scale—because Irrelevant contexts lack semantic content, the filtering mechanism does not engage, leaving only mechanical copying to scale upward. (Right) Counterfactual: distractor entrainment drops sharply ($b_{\text{dstr}} = -0.33$, from 9.69 to 2.30) while gold entrainment remains relatively stable (0.79 to 3.44); the gap narrows from 6.50 at 111M to 1.07 at 13B (6.1× reduction). This **convergent** pattern demonstrates effective misinformation resistance at scale: larger models not only suppress false claims but maintain or boost the correct answer.

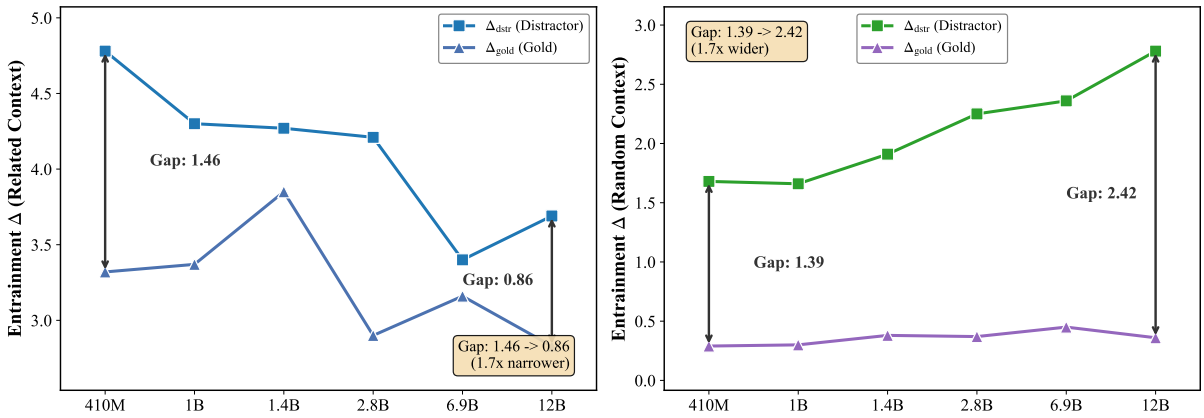
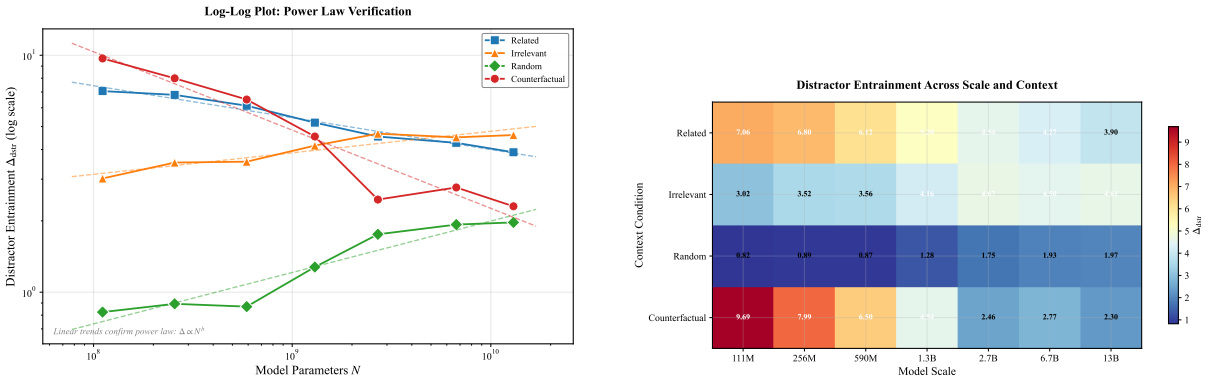


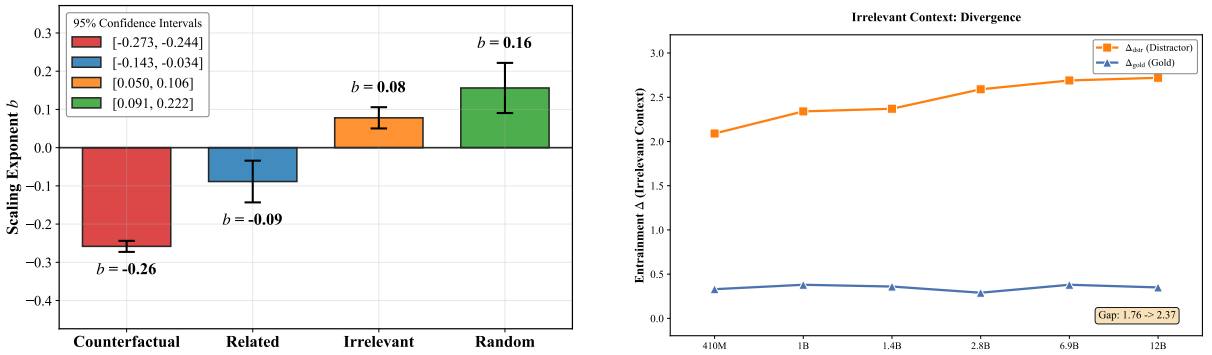
Figure 8: **Pythia: Convergence vs. Divergence Across Context Types.** Combined visualization showing opposite convergence behaviors for semantic vs. non-semantic contexts. (Left) Related context shows **convergent** behavior: the gold-distractor gap narrows with scale as distractor entrainment decreases ($b_{\text{dstr}} = -0.09$) while gold entrainment remains stable. The gap reduces from 2.41 at 410M to 0.74 at 12B—a 3.3× improvement in discrimination. (Right) Random context shows **divergent** behavior: the gap widens with scale as distractor entrainment increases ($b_{\text{dstr}} = +0.16$) while gold entrainment stays flat. The gap increases from 1.44 at 410M to 2.13 at 12B—a 1.5× degradation. This pattern exactly replicates the Cerebras-GPT findings (Figure 3 in main text), confirming that the convergent-divergent split is a general property of Transformer scaling, not a family-specific artifact. The replication across independently trained model families provides strong evidence for the dual-mechanism interpretation. Figure 10(b) additionally presents the Pythia Irrelevant divergence plot in isolation.



(a) Log-log verification of power-law scaling.

(b) Distractor entrainment heatmap.

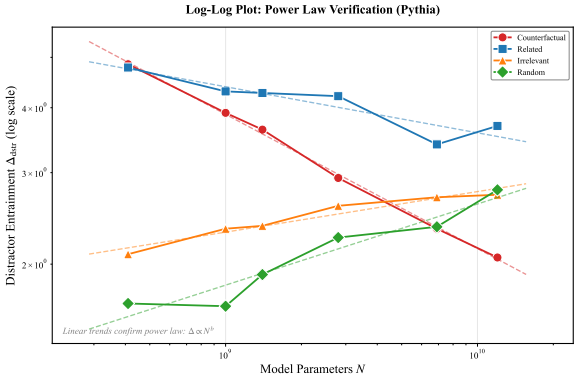
Figure 9: **Cerebras-GPT: Log-log verification and entrainment heatmap.** (Left) Distractor entrainment (Δ_{dstr}) plotted against model size (N) in log-log space for all four context conditions. Power-law relationships $\Delta_{\text{dstr}} \propto N^b$ appear as straight lines with slope b. Semantic contexts (Counterfactual, Related) show negative slopes (downward trends), while non-semantic contexts (Irrelevant, Random) show positive slopes (upward trends). The linearity across nearly two orders of magnitude in model size (111M–13B) confirms that power laws accurately describe entrainment scaling, and the parallel structure of lines within each group suggests a common underlying mechanism with condition-specific magnitudes. (Right) Matrix visualization of Δ_{dstr} across all combinations of context condition (rows) and model size (columns, 111M–13B). Color intensity indicates entrainment magnitude. The heatmap reveals two distinct gradients: semantic contexts decrease from left to right, while non-semantic contexts increase. The Counterfactual row shows the steepest gradient (strongest negative scaling), while the Random row shows clear brightening with scale (strongest positive scaling).



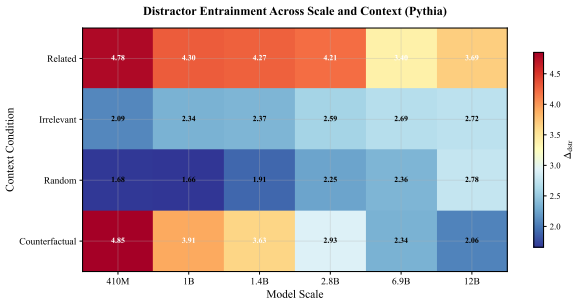
(a) Scaling exponents with 95% confidence intervals.

(b) Irrelevant context: divergent behavior.

Figure 10: **Pythia: Scaling exponents and Irrelevant-context divergence.** (Left) Estimated power-law exponents (b) for distractor entrainment (Δ_{dstr}) across the four context conditions, with 95% confidence intervals shown as error bars. Semantic contexts (Counterfactual: $b = -0.26$; Related: $b = -0.09$) show negative exponents, while non-semantic contexts (Random: $b = +0.16$; Irrelevant: $b = +0.08$) show positive exponents. The confidence intervals for semantic and non-semantic groups do not overlap, establishing statistical separation. This pattern exactly replicates the Cerebras-GPT findings (Figure 1 in main text), with the same ordering of exponent magnitudes: $|b_{\text{CF}}| > |b_{\text{Rel}}|$ and $|b_{\text{Rnd}}| > |b_{\text{Irr}}|$. (Right) Joint scaling of gold entrainment (Δ_{gold}, blue) and distractor entrainment (Δ_{dstr}, red) for the Irrelevant context condition. Similar to Cerebras-GPT (Figure 7a), gold entrainment remains essentially flat across scale (0.29 to 0.38) while distractor entrainment increases ($b_{\text{dstr}} = +0.08$, from 2.09 to 2.72); the gap widens from 1.89 at 410M to 2.22 at 12B. This **divergent** pattern replicates across model families, confirming that the mechanical copying mechanism scales independently of semantic content.



(a) Log-log verification of power-law scaling.



(b) Distractor entrainment heatmap.

Figure 11: **Pythia: Log-log verification and entrainment heatmap.** (Left) Distractor entrainment (Δ_{dstr}) plotted against model size (N) in log-log space for all four context conditions. As with Cerebras-GPT (Figure 9a), power-law relationships appear as straight lines. The Counterfactual condition shows particularly tight linearity ($R^2 = 0.998$), providing strong evidence that misinformation resistance scales as a precise power law. The semantic/non-semantic split in slopes replicates across model families, confirming that $\Delta_{\text{dstr}} \propto N^b$ with opposite-signed b is a general scaling pattern. (Right) Matrix visualization of Δ_{dstr} across all combinations of context condition (rows) and model size (columns, 410M–12B). The same divergent gradient pattern observed in Cerebras-GPT (Figure 9b) emerges here: semantic contexts (Counterfactual, Related) darken from left to right (decreasing entrainment), while non-semantic contexts (Irrelevant, Random) brighten (increasing entrainment). The overall lower intensity compared to Cerebras-GPT reflects Pythia's generally smaller entrainment magnitudes, though the directional trends are identical.