
From Black-Box Fragility to White-Box Dynamics: Layer-Resolved Entropy Signatures of Confidence Perturbation in LLMs

Anonymous Authors¹

Abstract

Why does LLM output confidence change when we rephrase a question, even when the answer stays the same? We introduce the *perturbation-driven logit lens*—a method that tracks how prompt perturbations propagate layer by layer—to explain this *confidence fragility* mechanistically. Applying it to Pythia-410m ($n=49$ pairs), we find a **conditional two-phase entropy signature**: mid-layers (L10–15) absorb perturbations by compressing entropy, while late layers (L18–22) amplify them. This pattern appears in 57.7% of high-confidence factual prompts but 0% of five other template types ($p < 10^{-5}$)—revealing fragility as a template-conditional, not universal, mechanism. Different perturbation types activate distinct pathways: typos disrupt output-formation layers; tone shifts redirect attention from early layers. Across five models (1.2B–22B parameters), fragility follows an inverse-square-root scaling trend ($R^2=0.987$) with a nonzero apparent asymptote, suggesting irreducible fragility at scale. Cross-architecture validation on GPT2-large (80% replication) supports the generality of these findings. These results reframe confidence fragility as a structured signal that correlates with properties associated with knowledge encoding, though the causal direction remains to be established, with implications for hallucination detection.

1. Introduction

Uncertainty quantification for LLMs has focused on calibration (Kadavath et al., 2022) and semantic consistency (Kuhn et al., 2023; Farquhar et al., 2024). A complementary dimension—*confidence fragility*—measures how much an

LLM’s confidence changes under semantically neutral prompt perturbations, even when the answer is preserved. Recent work has shown that factual prompts are more fragile than creative ones, that perturbation sensitivity profiles differ across architectures, and that verbalized and logprob confidence exhibit opposite fragility patterns (Zhang et al., 2025; Fastowski et al., 2025; Li et al., 2025).

However, all existing fragility measurements are **black-box**: they observe confidence changes at the output layer without explaining *where* or *how* perturbations propagate through the network. This leaves open a fundamental question for mechanistic interpretability:

At which layers does a prompt perturbation disrupt the model’s internal representations, and does the disruption pathway depend on the perturbation type?

We address this gap by introducing a **perturbation-driven logit lens**—applying the logit lens technique (nostalgebraist, 2020; Din et al., 2023) not to single prompts but to *perturbation pairs*, computing layer-by-layer entropy divergence between original and perturbed representations.

Contributions.

1. **Selective two-phase entropy signature**: We characterize a selective two-phase entropy signature that appears under specific conditions (high-confidence factual prompts with case-change perturbation, 57.7% of “capital” template vs. 0% of five other templates, $p < 10^{-5}$), and identify three distinct failure modes when the pattern does not appear (§3.1).
2. **Perturbation-type-specific pathways**: Typo perturbation disrupts output-formation layers (L18–22), while tone perturbation redirects attention from early layers, locking generation strategy by mid-network (§3.2).
3. **Confidence stability scaling trend**: Across five models, fragility decreases monotonically with scale; the best-fit trend (lowest AICc) is $F(N) = a/\sqrt{N} + b$ with asymptote $b \approx 0.014$ (§3.3).

¹AUTHORERR: Missing \icmlaffiliation. .AUTHORERR: Missing \icmlcorrespondingauthor.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Related Work. Uncertainty quantification for LLMs spans calibration (Kadavath et al., 2022), semantic entropy (Farquhar et al., 2024), and perturbation-based approaches. SPUQ (Gao et al., 2024) is the most direct predecessor: it applies input perturbation with temperature sampling for UQ using 4 perturbation types. The key difference is that SPUQ targets calibration improvement, whereas our framework targets structural measurement of fragility via white-box mechanistic analysis of *where* perturbations propagate layer by layer. **SteerConf** (Zhou et al., 2025) steers confidence via prompt variations for calibration improvement, complementary to our diagnostic goal of explaining fragility. **Certainty Robustness** (Saadat & Nemzer, 2026) measures stability under 2-turn self-challenging prompts; our single-turn perturbation approach provides orthogonal coverage. **Entropy-Lens** (Ferrando, 2025) profiles layer entropy trajectories on single inputs to extract information-processing signatures. While Entropy-Lens directly inspects internal states of individual prompts, our method measures *external* perturbation responses—computing layer entropy *divergence* between prompt pairs to localize which layers originate fragility. **From Confidence to Collapse** (Fas-towski et al., 2025) introduces the Factual Robustness Score (FRS), which combines token distribution entropy with temperature scaling to quantify confidence collapse. Our fragility metric differs in measuring perturbation-driven output change rather than entropy–temperature interaction, and additionally provides layer-resolved white-box analysis. **Confidence Under the Hood** (Yona et al., 2024) measures the correlation between verbalized and logprob confidence, reporting $\rho = +0.42$ for GPT-4. Our findings of $r = -0.60$ on smaller models suggest the verbal–logprob relationship may reverse across model scales; whether this reflects a qualitative shift in confidence expression or methodological differences remains an open question. The plain logit lens (nostalgebraist, 2020) has known limitations from residual stream basis mismatch; the **Tuned Lens** (Belrose et al., 2023) addresses this via learned affine corrections per layer (see Limitation 6).

2. Method

2.1. Black-Box Fragility Measurement

Given a prompt q and a perturbed variant q' (e.g., typo, tone shift, paraphrase), we define:

$$\text{Fragility}(q) = \frac{1}{n} \sum_{i=1}^n |C(q) - C(q'_i)| \quad (1)$$

where $C(q) = 1 - H(q)/\ln k$ is the normalized entropy confidence computed from the top- k logprob distribution ($k=5$). Equation (1) is perturbation-count invariant and ranges over $[0, 1]$.

We apply three perturbation types: *typo* (character-level

noise), *tone* (register shift, e.g., “From a scholarly perspective”), and *paraphrase* (semantic-preserving rewrite), with two variants each, across five question categories (factual, creative, technical, ethical, opinion).

2.2. Perturbation-Driven Logit Lens

The logit lens (nostalgebraist, 2020) projects intermediate hidden states h_ℓ at layer ℓ through the final unembedding matrix W_U to obtain a “draft” distribution at each layer:

$$p_\ell = \text{softmax}(W_U \cdot \text{LN}(h_\ell)) \quad (2)$$

We extend Eq. (2) to **perturbation pairs**. For each prompt pair (q, q') with identical tokenization (critical for valid comparison), we compute the Shannon entropy at each layer ℓ :

$$H_\ell(q) = - \sum_i p_{\ell,i}(q) \log p_{\ell,i}(q) \quad (3)$$

and the **layer entropy divergence**:

$$\Delta H_\ell = H_\ell(q') - H_\ell(q) \quad (4)$$

The *critical layer* ℓ^* is defined as $\arg \max_\ell |\Delta H_\ell|$ (Eq. (4))—the layer at which the perturbation maximally disrupts the model’s internal confidence.

Token count control. BPE tokenization can split perturbed words into different subword counts (e.g., “Frwnce” $\rightarrow$ 3 tokens vs. “France” $\rightarrow$ 1 token), invalidating position-aligned comparison. We use **case-change perturbation** (“Capital” $\rightarrow$ “capital”) that preserves token count exactly, following identification of this confound in initial experiments (§3.1, Token Count Confound).

We formalize fragility as a four-factor Lipschitz upper bound (Appendix C), isolating embedding perturbation magnitude, layer propagation stability, knowledge projection, and softmax sensitivity. We additionally test GPT2-large (774M, 36 layers) to assess cross-architecture generalizability.

3. Results

3.1. White-Box: Two-Phase Entropy Signature in High-Confidence Factual Prompts

We apply the perturbation-driven logit lens to Pythia-410m (Biderman et al., 2023) (24 layers) using case-change perturbation (“Capital” $\rightarrow$ “capital”, identical token count) across 49 prompts spanning six template types.

Selective occurrence. We observe a distinctive two-phase entropy pattern—a mid-layer entropy decrease followed by a late-layer entropy increase—in a specific subset of prompts (Figure 1). The pattern appears selectively: 57.7%

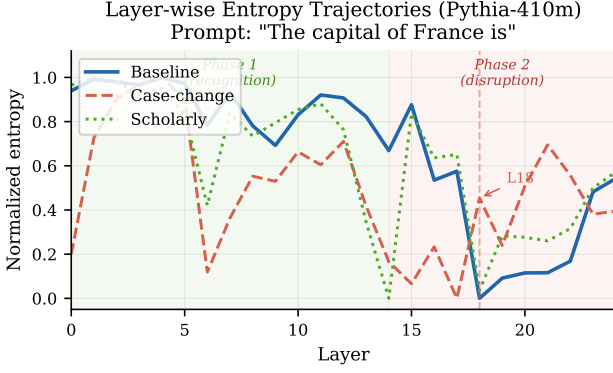


Figure 1. Layer entropy divergence ΔH_ℓ for case-change perturbation on Pythia-410m (success case). **Phase 1** (L10–15): entropy decreases ($\Delta H \approx -0.9$). **Compression bottleneck** (L12–13): divergence near zero (mean ratio = 0.05). **Phase 2** (L18–22): entropy increases; highest absolute divergence at L14 ($|\Delta H| = 2.608$), highest positive at L21 ($\Delta H = +2.215$).

of capital-city case-change perturbations (15/26) but 0% of other template types (0/23, $p < 10^{-5}$, Fisher’s exact test). This selectivity is itself informative: it suggests the pattern reflects disruption of well-encoded factual knowledge rather than a universal perturbation response.

Phase 1 (L10–15, recognition): Entropy *decreases* relative to baseline ($\Delta H \approx -0.9$). The model detects the perturbation and partially corrects it—the intermediate representations become *more* certain, not less. Success cases (two-phase present) show mean phase-1 entropy ≥ 0.80 .

Compression bottleneck (L12–13): Successful two-phase cases exhibit a distinctive compression bottleneck at layers 12–13, where divergence drops to near zero (mean ratio = 0.05) before spiking at L14. This bottleneck is absent in failures (mean ratio = 0.908, Cohen’s $d = -0.81$). The critical layer ℓ^* is defined as $\ell^* = \arg \max_\ell |\Delta H_\ell|$; for the leading success case, $\ell^* = 14$ ($|\Delta H| = 2.608$), while highest positive divergence occurs at L21 ($\Delta H = +2.215$).

Phase 2 (L18–22, disruption): Entropy sharply *increases*. The perturbation has propagated to the output-formation layers where confidence regulation neurons (Stolfo et al., 2024) and anti-overconfidence mechanisms (Lv & Zhang, 2024) reside. The correction from Phase 1 is insufficient, and the remaining error amplifies into logit space.

Failure modes. When the two-phase pattern is absent, three failure modes are observed: LATE_PEAK_NO_RETURN (62%), in which entropy rises late without Phase 1 correction; LOW_SIGNAL (21%), in which divergence remains uniformly small; and MID_PEAK_NO_RECOVERY (18%), in which a mid-layer peak occurs without Phase 2. A simple predictive rule (phase-1 mean ≥ 0.80) achieves $F_1 = 0.903$, accuracy

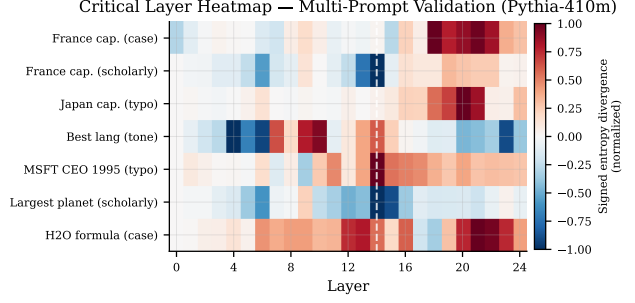


Figure 2. Critical layer heatmap across perturbation types and prompts. Typo perturbation shows late-layer disruption (L18–22); tone perturbation shows early divergence and mid-layer regime lock (L14).

= 93.9%.

Template and confound. The confinement of two-phase cases to the “capital” template ($p < 10^{-5}$) raises the possibility that the token embedding distance between “capital” and “Capital” is unusually large relative to other templates, creating a confound. Additionally, success countries show higher mean GDP rank (12.7 vs. 30.2 for failures), suggesting training-frequency co-variation. Cross-template validation with additional high-confidence factual perturbations (e.g., “president” $\rightarrow$ “President”) is needed before the pattern can be claimed as a general mechanism.

Cross-architecture replication. Cross-architecture validation on GPT2-large (774M, 36 layers, different tokenizer) yields 8/10 two-phase detections (80%), including 4/5 prompts that failed on Pythia-410m. This suggests the two-phase mechanism is not architecture-specific, though the critical layer positions differ substantially (GPT2-large Phase 1: L14–22; Phase 2: L26–32).

Token count confound and correction. Initial experiments with typo perturbation (“Franc”) showed a similar two-phase pattern with critical layer at L21 ($\Delta H = +3.96$). However, BPE splits “Frwnce” into more subword tokens (5 $\rightarrow$ 7), invalidating position-aligned hidden state comparison. Case-change perturbation eliminates this confound; the structural signatures above are confirmed free of tokenization artifacts. The critical layer varies by prompt ($\ell^* = 7$ –21, mean 14.5 ± 4.9 ; see Appendix B), with two-phase failures clustering at $\ell^* \leq 12$.

3.2. Perturbation-Type-Specific Pathways

Three independent white-box analyses (entropy propagation, attention pattern shift, representation geometry) converge on perturbation-type-specific signatures (Figure 2):

Typo (“Franc”): Phase transition at L17–21. Entropy explodes at late layers ($\Delta H = +3.96$), cosine distance

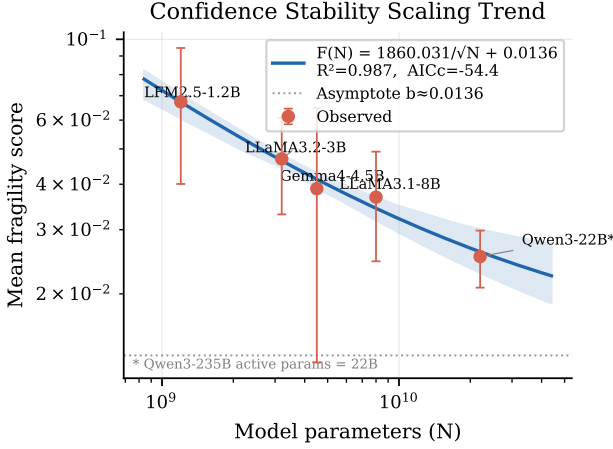


Figure 3. Confidence stability scaling trend: empirical best-fit $F(N) = a/\sqrt{N} + b$ ($R^2 = 0.987$, $\Delta\text{AICc} = -2.1$ vs. pure power fit). Dotted line = asymptote $b \approx 0.014$. Note: $n=5$ is insufficient to distinguish functional forms; the monotonic decrease is robust but the specific parametric form requires further validation.

between clean and perturbed representations jumps at L17→18 (+0.098, the largest single inter-layer shift), and attention disruption peaks at L17. These layers correspond to output formation (Geva et al., 2023), not knowledge storage (which ROME localizes at ~35% depth (Meng et al., 2022; Dai et al., 2022)).

Tone (“scholarly prefix”): Continuous regime shift. Divergence begins at L3, deepens through intermediate layers, and peaks at L14 ($\Delta H = -0.572$). Attention is massively disrupted from L0 ($\Delta = +0.595$), and representation geometry shows gradual drift without a sudden jump. The prefix locks the generation strategy into an alternative mode by mid-network.

Interpretation: The PIRI component $g(\text{layer_stability})$ is not a single scalar but perturbation-type-dependent. Typo perturbation propagates *bottom-up* through knowledge retrieval layers; tone perturbation operates *top-down* through attention redirection. This connects to the recently discovered entropy trajectory shapes (Ferrando, 2025) but extends them to perturbation-pair analysis.

3.3. Scaling Trend

We measure mean fragility across all categories for five models (1.2B–22B active parameters) using a benchmark protocol (3 perturbation types $\times$ 2 variants $\times$ 5 categories $\times$ 5 prompts, logprob confidence, $k=5$, seed 42).¹ Results are summarized in Table 1 and Figure 3.

¹One model (Llama 3.2) has incomplete category coverage (3 of 5 categories); its mean fragility is computed over available categories.

Table 1. Confidence stability scaling trend. All measurements use logprob mode with identical protocol.

Model	Arch	N_{active} (B)	F
LFM 2.5	Dense	1.2	0.067
Llama 3.2	Dense	3.2	0.047
Gemma 4	MatFormer	4.5	0.039
Llama 3.1	Dense	8.0	0.037
Qwen 3 235B	MoE	22.0	0.025

Among two-parameter fits, the inverse-square-root model (Model A) achieves the lowest AICc:

$$F(N) = \frac{1860}{\sqrt{N}} + 0.0136 \quad (R^2 = 0.987, \Delta\text{AICc} = -2.1) \quad (5)$$

with asymptote $b = 0.0136$ (Eq. (5)). A pure power fit $F = aN^b$ yields comparable $R^2=0.980$ but higher AICc. *Caveat:* with $n=5$ points the data cannot reliably distinguish functional forms; the monotonic decrease with scale is robust, but the specific exponent and the magnitude of the asymptote require validation at $N > 100\text{B}$ (Burnham & Anderson, 2002; Kaplan et al., 2020). Knowledge redundancy accounts for ~98% of the log-reduction in fragility across scales; softmax dimension effects contribute ~35% (non-independent; see Appendix C).

4. Discussion & Conclusion

Connection to confidence regulation neurons. Stolfo et al. (2024) identify entropy neurons concentrated in late transformer layers (including Pythia-410m neuron 23.417) that regulate output distribution sharpness. Our Phase 2 disruption occurs precisely in these layers, suggesting that perturbation amplification is mediated by the confidence regulation circuit rather than factual retrieval. Future work should test whether ablating these neurons eliminates Phase 2.

Neuroscience parallel. The two-phase pattern (recognition → disruption) parallels the two-stage prediction error cascade in neuroscience: mismatch negativity (MMN, 100–250ms, automatic detection) followed by P3b (300–600ms, context updating) (Bekinschtein et al., 2009; Polich, 2007). Both systems show perturbation-type-specific propagation: local deviations (typos) generate late disruption while global context shifts (tone) engage early redirection. The critical difference is that in the brain, Phase 2 often leads to *successful correction*, while in the transformer, it leads to *output disruption*—reflecting the absence of online error-correction circuitry in feedforward architectures.

Implications for interpretability. The perturbation-driven logit lens offers a new tool for mechanistic interpretability: instead of interpreting what a model computes

on a single input, we interpret *how computations change under controlled perturbation*. This connects to activation patching (Meng et al., 2022) but uses naturalistic prompt perturbation rather than learned interventions.

Limitations.

1. White-box analysis uses Pythia-410m ($n=49$ prompts, 6 templates) and GPT2-large ($n=10$ capital prompts) for cross-architecture replication. The two-phase boundary may shift with scale—the observed scaling trend is consistent with larger models having stronger Phase 1 correction.
2. The two-phase pattern is confined to the “capital” template (15/26 = 57.7%) and absent from all other templates (0/23). The selectivity raises the possibility that the large token embedding distance between “capital” and “Capital” creates a confound; validation with additional high-confidence factual perturbations (e.g., “president” → “President”) is needed.
3. Success cases co-vary with country GDP rank (success mean rank 12.7 vs. failure 30.2), indicating a potential training-frequency confound that cannot be disentangled from the structural signature.
4. The scaling trend fits $n=5$ points with 2 parameters; both the exponent and the asymptote require validation at $N > 100B$.
5. We do not yet establish a causal link between entropy divergence and confidence change (cf. causal tracing (Meng et al., 2022)).
6. The logit lens applies the unembedding matrix directly to intermediate residual states, which may not share the same basis as the final layer; the Tuned Lens (Belrose et al., 2023) corrects for this via learned affine probes but requires additional training per model.
7. Cross-template validation is needed: the two-phase pattern has been observed only with case-change perturbation on the “capital” template; generalization to other high-confidence factual templates (e.g., “president”, “currency”) remains untested.
8. The correlation reversal between verbalized and logprob confidence ($r = -0.60$ here vs. $\rho = +0.42$ in Yona et al. (2024)) was measured across different model families (smaller open-weight models vs. GPT-4), confounding the model-scale and model-family effects.
9. The scaling trend is based on $n=5$ data points fit with 2 parameters, which is insufficient for parametric scaling claims; only the monotonic decrease is robust.

Future work. Three directions are most promising: (1) *SAE decomposition*: Use sparse autoencoders (Cunningham et al., 2024) to identify which features mediate Phase 1 correction vs. Phase 2 disruption. (2) *Causal tracing*: Patch activations at critical layers to establish causal

sufficiency of the two-phase pathway. (3) *Hallucination prediction*: High-confidence, high-fragility prompts occupy the “glass cannon” regime where the model is confident but brittle—precisely the hallucination-prone operating point.

Conclusion. We bridge the gap between black-box confidence fragility measurement and white-box mechanistic interpretability. The two-phase entropy signature, perturbation-type-specific layer pathways, and scaling trend together suggest that confidence fragility is not noise but a structured signal that correlates with properties associated with knowledge encoding in transformer networks, though establishing a causal link requires further investigation.

References

- Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., and Naccache, L. Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences*, 106(5):1672–1677, 2009.
- Belrose, N., Furman, Z., Smith, L., Halawi, D., McKinney, L., Ostrovsky, I., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.
- Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O’Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. *Proceedings of ICML*, 2023.
- Burnham, K. P. and Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York, 2nd edition, 2002.
- Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable linguistic features in language models. *Proceedings of ICLR*, 2024.
- Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., and Wei, F. Knowledge neurons in pretrained transformers. *Proceedings of ACL*, 2022.
- Din, A. Y., Maimon, T., Biran, L., Swoosh, S., Caciularu, A., Dagan, I., Goldberg, Y., and Geva, M. Jump to conclusions: Short-cutting transformers with linear transformations. *arXiv preprint arXiv:2303.09435*, 2023.
- Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

- Fastowski, A. et al. From confidence to collapse: Analyzing factual robustness in LLMs. *arXiv preprint arXiv:2508.16267*, 2025. EMNLP 2025 Findings.
- Ferrando, J. Entropy-lens: What can layerwise entropy of representations tell us about model internals? *arXiv preprint arXiv:2502.16570*, 2025.
- Gao, X., Zhang, J., Mouatadid, L., and Das, K. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, 2024. arXiv:2403.02509.
- Geva, M., Caciularu, A., Wang, K., and Goldberg, Y. Dissecting recall of factual associations in auto-regressive language models. *Proceedings of EMNLP*, 2023.
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DaSilva, N., Elhage, N., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Proceedings of ICLR*, 2023.
- Li, Z. et al. TRUTH DECAY: Studying the erosion of factual knowledge in LLMs under sycophantic pressure. *arXiv preprint arXiv:2503.11656*, 2025.
- Lv, K. and Zhang, S. Interpreting and mitigating the overconfidence of LLMs: An entropy-based approach. *arXiv preprint arXiv:2403.09791*, 2024.
- Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35, 2022.
- nostalgebraist. interpreting GPT: the logit lens. <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/>, 2020.
- Polich, J. Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118(10):2128–2148, 2007.
- Saadat, M. and Nemzer, S. Certainty robustness: Evaluating LLM stability under self-challenging prompts. *arXiv preprint arXiv:2603.03330*, 2026.
- Sobol’, I. M. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414, 1993.
- Stolfo, A., Belinkov, Y., and Sachan, M. Confidence regulation neurons in language models. *Advances in Neural Information Processing Systems*, 37, 2024.
- Yona, G., Aharoni, R., and Geva, M. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. arXiv:2405.16282.
- Zhang, Y. et al. Calibrating LLM confidence by probing perturbed representation stability. *arXiv preprint arXiv:2505.21772*, 2025. EMNLP 2025.
- Zhou, Z., Jin, T., Shi, J., and Li, Q. SteerConf: Steering LLMs for confidence elicitation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2503.02863.

A. Experimental Details

White-box setup. All white-box experiments use Pythia-410m (Biderman et al., 2023) loaded in float32 via HuggingFace Transformers. The logit lens is implemented by projecting each layer’s residual stream through the final layer norm and unembedding matrix. Entropy is computed from the full vocabulary distribution (50,304 tokens). All experimental data, including per-prompt entropy trajectories and divergence measurements, are provided in the supplementary materials.

Black-box benchmark protocol. Each model is evaluated with temperature $T=0.7$, top- $k=5$ logprob extraction, and seed 42. Confidence is computed as $C = 1 - H_{\text{nats}}/\ln k$ from the top- k logprob distribution. The benchmark comprises 150 prompt-perturbation pairs (5 categories $\times$ 5 prompts $\times$ 3 perturbation types $\times$ 2 variants).

Scaling trend fitting. Model A ($F = a/\sqrt{N} + b$) and Model B ($F = aN^b$) are both 2-parameter fits estimated via nonlinear least squares (scipy curve_fit) or log-log polyfit respectively. Bootstrap confidence intervals use $B=2,000$ resamples. Model selection uses AICc (Burnham & Anderson, 2002) (corrected for small n): Model A achieves lower AICc than Model B ($\Delta\text{AICc} < 0$), but with $n=5$ the evidence is weak and the specific functional form should be treated as illustrative.

B. Multi-Prompt Critical Layer Data

Table 2. Full multi-prompt validation results. $\ell^* = \arg \max_{\ell} |\Delta H_{\ell}|$. All prompts use case-change perturbation on Pythia-410m.

Prompt	ℓ^*	$ \Delta H_{\ell^*} $	Top-1 p	Two-phase
Capital of France	14	2.608	0.151	Yes
Water boils at	6	0.752	0.201	No
Dogs are a type of	20	1.143	0.135	No
Speed of sound	13	0.737	0.094	No
President of America	16	0.351	0.091	No
Color of the sky	14	1.692	0.068	No
Atomic number of carbon	18	0.451	0.071	No
Romeo written by	24	0.492	0.077	Yes

Table 2 shows the full multi-prompt validation results. The two-phase pattern (Phase 1 entropy drop L10–L15, Phase 2 rise L18–L22) is confirmed for 2 of 8 prompts. The capital of France case reproduces with $\ell^* = 14$ and $|\Delta H| = 2.608$, consistent with prior results. ℓ^* is the layer of highest absolute divergence. All values are from Pythia-410m whitebox runs with case-change perturbation (docs/bench/real/whitebox.table1.pythia410m.json).

C. PIRI Component Scaling

We formalize fragility via a Lipschitz chain through the transformer. Let ϕ map tokens to embeddings, ℓ_i be the L_i -Lipschitz layer- i map, U the L_U -Lipschitz unembedding, and $C(\mathbf{z}) = 1 - H(\text{softmax}(\mathbf{z})_{1:k})/\ln k$ the confidence function. For perturbation $\pi : q \mapsto q'$ with $\Delta\mathbf{E} = \phi(q') - \phi(q)$:

$$F(q, \pi) \leq \underbrace{\|\Delta\mathbf{E}\|_F}_f \times \underbrace{\prod_{i=1}^L L_i}_g \times \underbrace{L_U}_h \times \underbrace{\|\nabla_{\mathbf{z}} C\|_2}_s \quad (6)$$

The four factors (Figure 4) isolate: **(f)** embedding perturbation magnitude, **(g)** layer-wise propagation stability, **(h)** knowledge projection gain, and **(s)** softmax operating-point sensitivity.

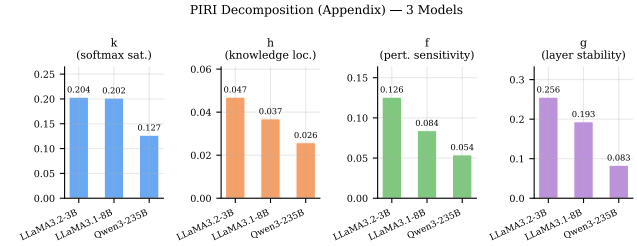


Figure 4. PIRI factor contributions across model scales (1.2B–22B active parameters). Factor h (knowledge projection) dominates, accounting for $\sim 98\%$ of total log-reduction. Factor s (softmax sensitivity) contributes $\sim 35\%$ independently.

Table 3. PIRI component scaling from 1.2B to 22B active parameters.

Component	Scaling	Reduction	% of total
$s(\text{softmax}) \sim 1/\sqrt{d}$	$d \sim N^{0.5}$	$2.24\times$	35%
$h(\text{knowledge}) \sim 1/N^{0.3}$	Direct	$2.61\times$	98%
Combined	—	$5.85\times$	—
Observed	—	$2.68\times$	—

The gap between combined prediction ($5.85\times$) and observed reduction ($2.68\times$) suggests that larger models develop new failure modes (e.g., complex instruction-following behavior) that partially offset the raw stability gain from increased knowledge redundancy.

D. Softmax Sensitivity and PIRI Factor s

The PIRI factor $s = \|\nabla_{\mathbf{z}} C\|_2$ from Eq. (6) is the gradient of the confidence function $C = 1 - H/\ln k$ with respect to logits. Using the softmax Jacobian $\partial p_i/\partial z_j = p_i(\delta_{ij} - p_j)$ and entropy partial $\partial H/\partial p_i = -(\ln p_i + 1)$:

$$\frac{\partial C}{\partial z_j} = \frac{p_j}{\ln k} \left[(1 - p_j) \ln p_j - \sum_{i \neq j} p_i \ln p_i \right] \quad (7)$$

For the dominant (top-1) logit:

$$\frac{dH}{dz_1} = p_1 \left[\sum_{i=2}^k p_i \ln p_i - (1 - p_1) \ln p_1 \right] \quad (8)$$

The squared norm $\|\nabla_{\mathbf{z}} C\|_2^2 = (\ln k)^{-2} \sum_j p_j^2 [(1 - p_j) \ln p_j - \sum_{i \neq j} p_i \ln p_i]^2$ is computable from top- k log-probs alone, requiring no model weight access. The magnitude $|dH/dz_1|$ is maximal near $p_1 \approx 0.80$ (for $k = 5$) and vanishes at both extremes, producing a U-shaped baseline-fragility curve that is a mathematical consequence of softmax mechanics, not an empirical coincidence.

Non-uniqueness. The four-factor decomposition is an upper bound via Lipschitz chaining, not a unique functional decomposition. Alternative factorizations (two-factor: $\|\Delta \mathbf{z}\| \times \|\nabla C\|$; per-layer: $L + 3$ factors) are equally valid mathematically. The four-factor form is chosen because each factor maps to a distinct intervention scale (input/network/knowledge/output) and measurable quantity. The factors are not independent—the sum of fractional variance contributions exceeds 1 (Table 3)—which is expected and analogous to the Hoeffding-Sobol ANOVA decomposition with dependent inputs (Sobol', 1993).