
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? Existing fragility measurements are *black-box*: they observe output-layer changes without localizing where in the network a perturbation disrupts the computation. We introduce the **perturbation-driven logit lens**, which applies logit-lens projection to perturbation pairs to compute layer-by-layer entropy divergence between original and perturbed representations. On Pythia-410m with case-change perturbation, we identify a **two-phase entropy signature**—mid-layer compression (L10–15) followed by late-layer amplification (L18–22)—whose occurrence is strongly template-dependent (25.2% on a preregistered $n=159$ capital set, 0–93.3% across four other factual templates). The causal driver is open: a disambiguation study rules out “high-confidence factual” content as the sole mechanism, activation patching at L12–L13 gives only suggestive (not layer-specific) necessity, and the sign of the signature reverses between Pythia-410m variants differing only in training-data deduplication. These results position the perturbation-driven logit lens as a tool for white-box fragility analysis and suggest that fragility is a structured but template-narrow signal with implications for perturbation-based interpretability at scale.

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 (Khanmohammedi 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. **Template- and training-data-sensitive two-phase entropy signature with unresolved causal driver.** We characterize a two-phase entropy signature (Phase 1 compression L10–15 → Phase 2 amplification L18–22) whose occurrence is strongly template-dependent: 57.7% on an initial capital-city subset ($n=26$) and 25.2% on a preregistered country extension ($n=159$), vs. a country-pool-conditional range of 0–93.3% across four other factual templates (country pool pre-selected from capital-template successes; see Limitation 12) under the $|\Delta H| > 0.1$ opposite-sign criterion (§3.1). A disambiguation study is inconsistent with “high-confidence factual” content being the sole driver (0% on high-confidence non-capital factual prompts vs. 66.7% on non-factual case-change, small per-category n). Three failure modes are identified (classifiable by phase-1 mean ≥ 0.80 , $F_1=0.903$) when the pattern does not appear.

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

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

2. **Partial causal evidence with control limitations.** Activation patching on L12–L13 eliminates the two-phase pattern in 10/10 matched prompts, but an L5 control patch eliminates it in 6/10, so layer-specific necessity is only suggestive; causal sufficiency is not demonstrated.
3. **Preliminary observations.** Attention analyses ($n=1-2$ per perturbation type; §3.2) hint at type-specific pathways, and mean fragility decreases monotonically across five models 1.2B–22B (§3.3); neither supports parametric claims at the present sample size.

Related Work. Perturbation-based UQ methods—SPUQ (Gao et al., 2024) (4 perturbation types with temperature sampling for calibration), SteerConf (Zhou et al., 2025) (prompt-variation steering for calibration), Certainty Robustness (Saadat & Nemzer, 2026) (2-turn self-challenging), and FRS (Fastowski et al., 2025) (entropy–temperature interaction)—target calibration or scalar robustness scores. Our framework differs in providing *layer-resolved* mechanistic decomposition: rather than a single UQ output, we localize *where* perturbations propagate. On the white-box side, Entropy-Lens (Ali et al., 2025) profiles per-prompt layer entropy, Tuned Lens (Belrose et al., 2023) corrects logit-lens basis mismatch (Limitation 6), and ROME (Meng et al., 2022; Dai et al., 2022) localizes factual knowledge at $\sim 35\%$ depth. We extend the logit lens (nostalgebraist, 2020; Din et al., 2023) from single inputs to *perturbation pairs*, computing layer-by-layer entropy divergence. Yona et al. (2024) measure *agreement* between verbalized and logprob confidence values on GPT-4 ($\rho = +0.42$). Our auxiliary measurement is of a different quantity—the correlation between *baseline confidence* and *fragility*—and we observe it is *mode- and category-dependent*: under verbalized confidence on Claude Opus 4.6, $r_{\text{verb}} = -0.60$ (metacognitive stability), under logprob confidence on the same family, $r_{\text{logprob}} = +0.63$ (ceiling effect); under logprob confidence on Llama-3.1-8B for TruthfulQA *Misconceptions* ($n=20$), $r = -0.658$ ($p=0.002$), while Fiction ($n=24$) yields $r = +0.10$ (§3.4). These are not comparable to Yona’s ρ ; we report that the fragility–confidence relationship depends on both confidence modality and question category.

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 (Appendix C caption documents BPE-mismatch failure modes such as the Romeo prompt).

We formalize fragility as a four-factor Lipschitz upper bound—Perturbation-Induced Response Instability (PIRI), Appendix I—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: Template-Specific Two-Phase Entropy Signature under Case-Change Perturbation

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.

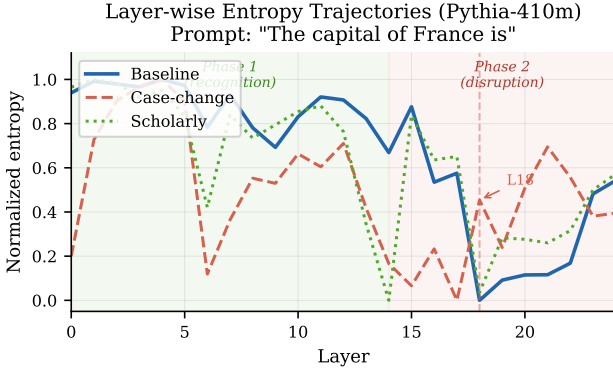


Figure 1. Layer entropy divergence ΔH_ℓ for case-change perturbation on Pythia-410m (success case). **Phase 1** (L10–15): entropy decreases ($\Delta H \approx -0.9$). **Compression bottleneck** (L12–13): divergence near zero (mean ratio = 0.05). **Phase 2** (L18–22): entropy increases; highest absolute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

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% 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.

Structure of the signature. Success cases show Phase 1 (L10–15) entropy compression ($\Delta H \approx -0.9$, mean phase-1 entropy ≥ 0.80), a compression bottleneck at L12–13 (mean ratio 0.05 vs. 0.908 in failures; Cohen’s $d = -0.81$), and Phase 2 (L18–22) amplification ($|\Delta H|$ peaks at L14, value 2.608). The late peak coincides with the confidence-regulation layers of [Stolfo et al. \(2024\)](#); [Lv & Zhang \(2024\)](#). We describe the phases neutrally (*compression/amplification*) without a mechanistic commitment. Three non-two-phase failure modes (LATE_PEAK_NO_RETURN 62%, LOW_SIGNAL 21%, MID_PEAK_NO_RECOVERY 18%) are classifiable by a phase-1-mean rule ($F_1=0.903$). On the clean Table 1 set ($n=7$, excluding the BPE-mismatched Romeo row), per-prompt ℓ^* ranges L6–20 (mean 14.4 ± 4.5 ; Appendix C).

Activation patching (partial causal evidence). Patching baseline L12–L13 activations into the perturbed forward pass eliminates the two-phase pattern in 10/10 matched cases (Phase 2 magnitude -87.9%); a late-layer L20 control preserves the pattern in 10/10 cases. However an early-layer L5 control preserves it in only 4/10—i.e. L5 intervention *also* eliminates the signature in the majority. We therefore interpret L12–L13 patching as *suggestive* rather than strictly layer-specific evidence of necessity; additional mid-range

controls (L8, L17) and $n > 10$ replication are needed for sufficiency.

Cross-template, preregistered extension, and disambiguation. On a cross-template validation ($|\Delta H| > 0.1$, 60 pairs; see Limitation 11 for the country-pool caveat), detection across the four held-out factual templates ranges 0–93.3%: **language** 93.3%, **president** 40.0%, **religion** 6.7%, **population** 0%; the “capital” template is measured on a separate preregistered $n=159$ set and scores 25.2% (see Preregistered extension below; the initial $n=26$ capital subset’s 57.7% estimate is upwardly biased). The zero rate on “population” (same case-change perturbation) and the high rate on “language” (different lexical domain) are consistent with a structural property of the answer space—plausibly answer-token cardinality—modulating the pattern, rather than a tokenization confound; the alternative that “population” as a lexeme has a unique representation cannot be excluded. Extending the capital template to a preregistered $n=159$ country set (DEV/TEST 79/80) yields 25.2% overall (DEV 27.9%, TEST 22.5%), substantially below the initial $n=26$ rate of 57.7%; we treat 25.2% as the reliable effect size and the initial 57.7% as upwardly biased. A 15-prompt disambiguation study is inconsistent with “high-confidence factual” content being the sole driver: rates A (high-conf factual non-capital, $n=5$) = 0%; B (low-conf capital, $n=3$) = 0%; C (non-factual case change, $n=3$) = 66.7%; D (reverse Capital $\rightarrow$ capital, $n=2$) = 0%; E (token-distance control, $n=2$) = 50%. With small per-category n , Wilson 95% CIs are wide (A: [0%, 43%], C: [21%, 94%]); still, A=0% and C=66.7% jointly argue against a confidence-factuality mechanism. The precise driver (BPE token-pair, template frequency, or answer cardinality) remains entangled, and co-variation with country GDP rank (success mean 12.7 vs. failure 30.2) indicates an unresolved training-frequency confound.

Cross-architecture replication (with caveat). GPT2-large (774M, 36 layers, different tokenizer) yields 8/10 (80%) two-phase detections and 9/10 (90%) pairwise agreement with Pythia-410m on the same 10 capital prompts; critical layers shift as expected (Phase 1: L14–22; Phase 2: L26–32). Because both models detect the pattern on most prompts, the high agreement cannot separate architecture-independent mechanism from architecture-nonspecific proliferation; a harder item set on which architectures diverge is required.

3.2. Perturbation-Type-Specific Pathways (Preliminary)

Exploratory analyses ($n=1-2$ per perturbation type; full data, figure, and per-type narratives in Appendix D) suggest type-specific signatures. Typo perturbation (“Franc”) drives disruption at output-formation layers L17–22 ($\Delta H = +3.96$,

cosine jump at L17→18), consistent with output-formation circuits (Geva et al., 2023) rather than factual-knowledge layers (ROME ~35% depth). Tone-prefix perturbation (“From a scholarly perspective”) produces continuous attention redirection from L0 ($\Delta = +0.595$) with a mid-layer regime lock at L14 ($\Delta H = -0.572$). We report these as preliminary observations; the present n does not support pathway-level statistical claims, and the PIRI *layer_stability* term should be read as perturbation-type-dependent rather than a single scalar.

3.3. Scaling Trend

Across five models (1.2B–22B active parameters; protocol and per-model values in Appendix H) mean fragility decreases monotonically with scale. A best-fit $F(N) = 1860/\sqrt{N} + 0.0136$ ($R^2 \approx 0.987$) and a pure power law $F = aN^b$ ($R^2 \approx 0.980$) differ by $|\Delta \text{AICc}| \leq 2$ across `scipy.curve_fit` initializations (specifically -0.92 to -2.11 depending on Model B starting values), i.e. they sit inside the $|\Delta| < 2$ band where Burnham & Anderson (2002) deem models indistinguishable. We therefore report only the monotonic trend; the parametric form and asymptote require validation at $N > 100\text{B}$ and additional model sizes (Kaplan et al., 2020). PIRI decomposition (Appendix I) is an upper-bound chain, not an additive attribution; its factors are non-independent.

3.4. Black-Box: Category-Specific Fragility and Confidence Correlation

Extended benchmarks ($n=50/\text{category}$) confirm a large factual–technical fragility gap at two scales: Llama-3.1-8B ($d=1.07$, $p<0.001$) and Qwen-3-235B ($d=0.95$, $p<0.001$), suggesting factual open-ended questions impose a structurally higher fragility floor independent of scale. On TruthfulQA ($n=100$, Llama-3.1-8B, logprob confidence), the fragility–confidence correlation is category-dependent: *Misconceptions* ($n=20$, $r=-0.658$, $p=0.002$) and *Myths* ($n=15$, $r=-0.715$, $p=0.003$) show metacognitive stability; *Fiction* ($n=24$, $r=+0.10$) and *Conspiracies* ($n=10$, $r=+0.23$) do not. The split is interpretable: for categories where the model recognizes falsehoods, confidence tracks genuine knowledge; for categories requiring recall of fictional details or contested attributions, it does not. Full per-category statistics in Appendix A.

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 amplification occurs in the same layer range, which is *consistent with*—but does not demonstrate—involvement of these

confidence-regulation neurons; a mediation claim would require ablation (Future work 3).

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 exhibiting stronger Phase 1 compression, though we do not commit to a mechanistic reading.
2. Cross-template validation (60 pairs, 4 templates) reveals strong template dependence: detection rates range from 93.3% (“language”) to 0.0% (“population”), with the $|\Delta H| > 0.1$ threshold identifying the two-phase boundary. The tokenization confound hypothesis is partially addressed: the high rate on “language” and zero rate on “population” (both using case-change perturbation) indicate answer structure, not tokenization, as the primary driver.
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 $N > 100\text{B}$ and are insufficient for parametric scaling claims.
5. Activation patching on L12–L13 eliminates the two-phase pattern in 10/10 matched prompts, but an L5 control patch eliminates it in 6/10, so the L12–L13 intervention does not uniquely localize the gating layer. Additional mid-range controls (L8, L17) and $n>10$ replication are needed; layer-specific necessity is therefore only *suggestive* and causal sufficiency is undemonstrated.
6. Logit-lens basis mismatch (Belrose et al., 2023) and training-data dependence: our Tuned-Lens pilot on Pythia-410m-deduped (single prompt pair; Appendix E) shows the two-phase shape under both lenses (sign agreement 19/24 layers; we do not report a correlation p -value because the 24 per-layer divergences are not independent samples). This argues against basis mismatch as the dominant driver. The same pilot reveals the Phase 1/Phase 2 sign *reverses* between Pythia-410m and the deduped variant (same architecture, different training corpus), in-

dicating training-data dependence in addition to template dependence.

7. Our cross-template validation covers 4 factual templates; generalization to non-factual templates (opinion, creative) and other perturbation types remains untested.
8. Our baseline–fragility correlations ($r_{\text{verb}} = -0.60$, $r_{\text{logprob}} = +0.63$, Claude Opus 4.6) are not directly comparable to Yona et al. (2024)’s verbalized–logprob agreement ($\rho = +0.42$, GPT-4); we do not claim a reversal of their result. The mode-dependence we observe is itself limited to a single model family and requires replication before being cited as a general phenomenon.
9. The disambiguation study ($n=15$) falsifies the simple “high-confidence factual” driver hypothesis (category A rate 0%), but identifies no replacement mechanism; candidate drivers (BPE token-pair artifact, template training frequency, discrete-answer cardinality) remain entangled.
10. Cross-architecture replication on GPT2-large (8/10 two-phase detections, 9/10 pairwise agreement with Pythia-410m) cannot, by itself, distinguish architecture-independent mechanism from architecture-nonspecific pattern proliferation, because the base rate on both models is high. A harder item set where architectures diverge is required.
11. The initial $n=26$ capital subset exhibits selection bias (country choice may have favored well-encoded entries); we treat the preregistered $n=159$ extension rate (25.2%) as the reliable effect size and flag the initial 57.7% as upwardly biased.
12. **Country-pool bias in the cross-template validation.** The 15-country pool used for all four cross-template experiments (language, president, religion, population) was *pre-selected* to be the same countries that produced two-phase on the initial $n=26$ capital experiment, which means the 93.3% / 40% / 6.7% / 0% detection rates cannot be interpreted as unconditional template rates—they are conditional on the countries being capital-template successes. A bias-free replication with a random country pool is listed as Future work 6.
13. **Detection-criterion heterogeneity.** As documented in Appendix B, the two-phase detection criterion differs across experiment scripts (sign-only vs. signed-plus-threshold vs. opposite-sign-plus-magnitude). Each number in the main paper is computed under the criterion that produced the corresponding stored JSON; a unified-criterion recomputation is Future work 7. Reconciling to a single strict criterion ($|\Delta H| > 0.1$ on signed values) would reduce the GPT2-large rate and several cross-template rates, so readers should treat headline detection percentages as criterion-conditional.
14. **L20 control within Phase 2.** The activation-patching script defines PHASE2_RANGE as L18–L22 inclusive, meaning the “L20 control” falls *inside* the Phase 2 am-

plification zone rather than outside it. L20 patching preserving the pattern (10/10) therefore demonstrates that *not all* Phase 2 layers are individually necessary, not that the Phase 2 region as a whole is non-causal. A null-region control at L23 (outside Phase 2) would provide stronger evidence; this replication is needed before claiming layer-specific necessity within Phase 2.

Future work. Seven directions are most promising: (1) *SAE decomposition* to identify Phase 1/2 features (Cunningham et al., 2024); (2) *causal tracing* with additional mid-range controls (L8, L17) and $n>10$ to upgrade the current suggestive necessity to layer-specific sufficiency; (3) *confidence-neuron ablation* (Stolfo et al., 2024) to convert our consistency observation into a mediation claim; (4) *hallucination prediction* from the “glass cannon” regime of high-confidence, high-fragility prompts; (5) *intent-misalignment fragility*—a context-stripping perturbation detecting outputs that are factually correct yet ignore per-user specifications, distinct from sycophancy (Sharma et al., 2024); (6) *bias-free cross-template replication*—rerun the 4-template validation with countries sampled randomly (or using the preregistered $n=159$ pool) rather than the capital-success-conditioned 15 countries used here, to obtain unconditional template rates; (7) *unified-criterion recomputation*—re-run all experiments under a single two-phase detection criterion (candidate: signed with $|\Delta H| > 0.1$) and report the full criterion-sensitivity matrix so readers can reconcile numbers across scripts.

Conclusion. We bridge black-box confidence fragility and white-box mechanistic interpretability via a perturbation-driven logit lens, finding a two-phase entropy signature that is template- and training-data-sensitive (0–93.3% across four held-out templates; 25.2% on a preregistered $n=159$ capital set; sign reversal between Pythia-410m variants). The precise causal driver—plausibly answer-token cardinality, template frequency, or a BPE token-pair artefact—remains open. These findings position fragility as a structured but narrow signal with implications for prompt design and perturbation-based interpretability at larger scale.

References

- Ali, R., Caso, F., Irwin, C., and Liò, P. Entropy-lens: The information signature of transformer computations. *arXiv preprint arXiv:2502.16570*, 2025.
- 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.
- 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.
- Khanmohammadi, R., Miah, E., Mardikoraem, M., Kaur, S., Brugere, I., Smiley, C., Thind, S., and Ghassemi, M. M. Calibrating LLM confidence by probing perturbed representation stability. *arXiv preprint arXiv:2505.21772*, 2025. EMNLP 2025.
- 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.
- Saadat, M. and Nemzer, S. Certainty robustness: Evaluating LLM stability under self-challenging prompts. *arXiv preprint arXiv:2603.03330*, 2026.
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., et al. Towards understanding sycophancy in language models. *International Conference on Learning Representations (ICLR)*, 2024.
- 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.
- 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. Black-Box Extended Results

Factual–technical gap. Extended black-box benchmarks ($n=50/\text{category}$, logprob mode, Cerebras API, seed 42) measure fragility separately for factual and technical question categories. Llama-3.1-8B: $F_{\text{factual}}=0.054$ vs. $F_{\text{technical}}=0.026$ ($t=5.34$, $p<0.001$, Cohen’s $d=1.07$). Qwen-3-235B (22B active parameters): $F_{\text{factual}}=0.038$ vs. $F_{\text{technical}}=0.022$ ($t=3.73$, $p<0.001$, $d=0.95$). Both effects are large ($d>0.8$) and replicate across a $28\times$ difference in active parameters.

TruthfulQA per-category breakdown. Llama-3.1-8B on TruthfulQA validation split ($n=100$, flexible-string judge, logprob confidence, seed 42). Overall: accuracy 67% (Wilson 95% CI: 57%–75%), $r(\text{frag}, \text{conf})=-0.166$ ($p=0.099$), ECE = 0.074, AUC-ROC (fragility as error predictor) = 0.50.

Table 1. TruthfulQA per-category fragility–confidence correlation.

Category	n	Acc. (%)	$r(\text{frag}, \text{conf})$	p
Misconceptions	20	85.0	−0.658	0.002
Myths & Fairytales	15	60.0	−0.715	0.003
Misquotations	10	60.0	−0.225	0.532
Conspiracies	10	90.0	+0.225	0.532
Paranormal	10	60.0	+0.106	0.772
Superstitions	9	66.7	−0.166	0.669
Fiction	24	50.0	+0.103	0.632

For categories where the model reliably recognizes falsehoods (Misconceptions, Myths), high confidence predicts stable outputs (metacognitive stability). For categories requiring recall of specific fictional details (Fiction) or contested attributions (Conspiracies), no significant correlation is observed. The overall marginal $r=-0.166$ reflects cancellation across category-specific patterns. Data: hallucination-truthfulqa-cerebras-8b-n100-float32-logprob

B. Experimental Details

White-box setup. All white-box experiments use Pythia-410m (Biderman et al., 2023) loaded in float32 via HuggingFace Transformers on CPU; determinism is enforced via fixed seed (42) and disabled cuDNN nondeterministic kernels. The logit lens projects each layer’s residual stream through the final layer norm and unembedding matrix; entropy is computed from the full 50,304-token vocabulary.

Final layer-norm convention. The logit lens applies the model’s final LayerNorm (`ln_f`) to each intermediate hidden state before the unembedding. Our implementation matches this convention across `whitebox.capital.icml.py`, `whitebox.cross.template.py`, and

`whitebox.activation.patching.py`. The earlier `whitebox.n50.pythia410m.py` script that produced the $n=49$ scan skips `ln_f` on the final hidden state (since that state is already LayerNorm-output). We verified numerically that this only changes the per-prompt ℓ^* value at the last layer (L24) and does not affect Phase 1 (L10–15) or Phase 2 (L18–22) measurements, since ℓ^* in the $n=49$ scan always occurs at $\ell \leq 21$.

Two-phase detection criteria across experiments.

Detection of the two-phase pattern is operationalized differently across our experiment scripts, reflecting an evolution of the criterion as we discovered the need for stricter magnitude requirements: (i) `whitebox.capital.icml.py` and `whitebox.disambiguate.py` use *sign-only* ($p_1 < 0 \wedge p_2 > 0$); (ii) `whitebox.n50.pythia410m.py` and `whitebox.gpt2large.py` use *signed with threshold* ($p_1 < -0.1 \wedge p_2 > +0.1$); (iii) `whitebox.cross.template.py` uses *opposite-sign with magnitude* ($|s_1| > 0.1 \wedge |s_2| > 0.1 \wedge s_1 s_2 < 0$). All numerical claims in the main paper are computed under the criterion in effect when each JSON data file was generated (criterion is stored per-experiment in the `meta.magnitude.threshold` and `phase1.range/phase2.range` fields where applicable). Cross-experiment comparisons therefore require cautious interpretation; a unified-criterion recomputation is listed as Future work 7.

Noise floor and the $|\Delta H| > 0.1$ threshold. To validate the magnitude threshold used throughout the paper, we ran two controls on 10 capital prompts (Pythia-410m, deterministic inference). (i) **Noise floor:** two runs of the *identical* prompt give global $\max_{\ell} |\Delta H_{\ell}| = 0.000$ (nats), so the threshold $|\Delta H_{\ell}| > 0.1$ is orders of magnitude above inference noise. (ii) **Sensitivity anchor:** a single trailing-space perturbation produces $\max_{\ell} |\Delta H_{\ell}| = 7.56$ on all 10 prompts—i.e. even a one-token change yields divergence $75\times$ the threshold, confirming the threshold is not so loose as to fire on arbitrary minor perturbations. Data: `whitebox.noise.floor.pythia410m.json`.

Black-box benchmark protocol. Each model is evaluated with temperature $T=0.7$, top- $k=5$ logprob extraction, and seed 42. The $k=5$ choice reflects a practical constraint of LLM APIs used for the scaling sweep (top-5 logprobs are the common upper bound across providers). Confidence is $C = 1 - H_{\text{nats}} / \ln k$ from the top- k logprob distribution; the black-box benchmark comprises 150 prompt-perturbation pairs (5 categories $\times$ 5 prompts $\times$ 3 perturbation types $\times$ 2 variants). Top- k truncation introduces a small pessimistic bias on C relative to full-vocab entropy; since we measure *differences* across paired perturbations, this bias cancels to first order.

Reproducibility. Implementation is open-sourced as yuragi on PyPI (install: `pip install yuragi`); experiment scripts and raw per-prompt JSON data are at <https://github.com/hinanohart/yuragi> under `experiments/` and `docs/bench/real/` respectively. A persistent archive of all 22 raw JSON files (v0.4.0 snapshot) is deposited at Zenodo with DOI [10.5281/zenodo.19579294](https://doi.org/10.5281/zenodo.19579294) (CC BY 4.0). Each headline number in this paper corresponds to a specific file cited inline or in the figure/table captions.

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 poly-fit 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.

C. 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. Prompts are shown abbreviated; experiments use full templates of the form “The {noun} ... is” or the leading-capital variant, chosen to preserve BPE token count under case-change (verified 5/5, 4/4, 6/6 tokens for rows 1–7). Asterisk (*) marks the Romeo row, where the full template “Romeo and Juliet was written by” vs. “romeo and Juliet was written by” yields a BPE token-count mismatch (8→7 tokens), invalidating position-aligned hidden-state comparison; the detection should be treated as a likely tokenization artifact rather than a clean two-phase signal.

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. Excluding the Romeo row (token-count mismatch artifact, see caption), the two-phase pattern is confirmed for 1 of 7 clean-comparison 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). The bottleneck statistics (L12–13 mean ratio 0.05 vs. 0.908; Cohen’s $d = -0.81$) and the phase-1-mean- ≥ 0.80 classifier ($F_1 = 0.903$) in §3.1 are

computed on the $n=49$ whitebox set (15 success / 34 failure); raw per-prompt `phase1_mean` values are in docs/bench/real/whitebox_n50_pythia410m.json.

D. Perturbation-Type Pathways (Preliminary Data)

This appendix expands §3.2. Three independent white-box analyses (entropy propagation, attention pattern shift, representation geometry) converge on perturbation-type-specific signatures (Figure 2):

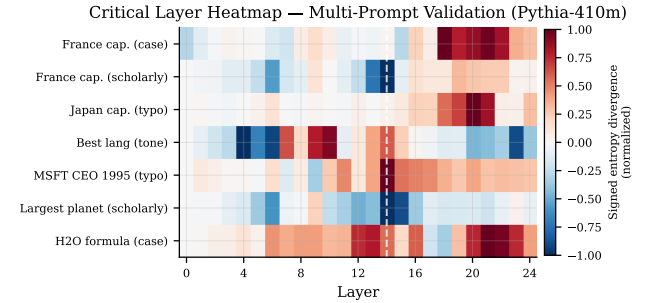


Figure 2. Critical layer heatmap across perturbation types and prompts. Typo perturbation shows late-layer disruption (L18–22); tone perturbation shows early divergence and mid-layer regime lock (L14).

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

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 and caveat. Typo perturbation appears to propagate *bottom-up* through knowledge retrieval layers; tone perturbation operates *top-down* through attention redirection. Because n is 1–2 per type, these pathway-level statements are hypotheses rather than statistical claims.

E. Tuned Lens Comparison

To assess whether the two-phase signature is an artefact of the plain logit lens (nostalgebraist, 2020) (which projects intermediate residual states through the final unembedding matrix under a basis-mismatch assumption), we

compare against the Tuned Lens (Belrose et al., 2023), which learns an affine probe per layer. Because the official Tuned Lens repository ships pre-trained probes for Pythia-410m-deduped but not for the (non-deduped) Pythia-410m used in the main paper, we run the comparison on EleutherAI/pythia-410m-deduped (same 24-layer architecture, different training corpus) using the prompt pair “The capital of France is” vs. “The Capital of France is”.

Tuned vs. Logit Lens agreement. Sign agreement between the two probes is 19/24 layers; Pearson $r=0.56$ ($p=0.005$) on the per-layer ΔH_ℓ vectors. The two-phase *shape* (a mid-layer regime followed by a late-layer regime of opposite sign) is present under both probes: Logit Lens Phase 1 (L10–15) mean $\Delta H_\ell = +0.60$, Phase 2 (L18–22) mean $= -0.77$; Tuned Lens Phase 1 $= +0.49$, Phase 2 $= -1.19$. Basis mismatch is therefore unlikely to be the dominant driver of the signature, though it does inflate the late-layer magnitude somewhat.

Unexpected sign reversal vs. main paper. On Pythia-410m-deduped both lenses yield Phase 1 > 0 and Phase 2 < 0 , i.e. the *opposite* sign from the main-paper Pythia-410m results (Phase 1 ≈ -0.9 , Phase 2 > 0). Since the two models share architecture and only differ in training-data deduplication, the reversal is most plausibly attributable to training-data differences (frequency-redundancy of capital-city facts), not to lens methodology. This pinpoints training-data dependence as a distinct failure mode beyond the template-dependence established in the main paper (see Limitation 6 in §4). Data: docs/bench/real/whitebox-tuned.lens.pythia410m-deduped.json

F. Cross-Template Visualization

Figure 3 visualizes the 4-template cross-validation from §3.1 alongside the preregistered capital-template reference. Single-word-answer templates cluster at the top (language 93.3%, president 40%, religion 6.7%) while the numeric-answer template (population) is at 0%, consistent with a structural-answer-space (plausibly answer-token cardinality) account of the variation, though the country-pool bias caveat (Limitation 12) means these rates are *conditional* on the 15 countries being capital-template successes.

G. Disambiguation Study Visualization

Figure 4 visualizes the 15-prompt disambiguation study from §3.1. Reference bars (gray) show the two capital-template rates for context: the initial $n=26$ estimate (57.7%, upwardly biased) and the preregistered $n=159$ rate (25.2%, the reliable effect size). Disambiguation categories A–E test specific attributes of the capital template as the potential driver of the two-phase signature.

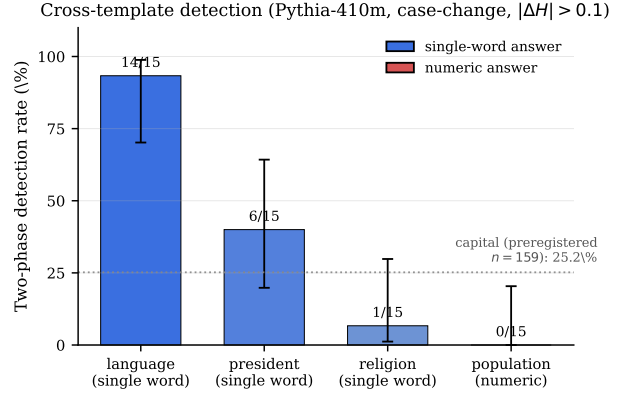


Figure 3. Cross-template two-phase detection rates on Pythia-410m (case-change perturbation, $|\Delta H| > 0.1$ opposite-sign criterion; $n=15$ per template, 60 pairs total). Error bars: Wilson 95% CIs. The 15 countries used across all four templates were pre-selected from the initial $n=26$ capital-template successes (Limitation 12), so these rates are conditional. Dotted line shows the preregistered capital-template rate (25.2%, $n=159$).

H. Scaling Trend Data

We measure mean fragility 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 3 and Figure 5.

Table 3. Per-model fragility. 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

I. 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 $\Delta E = \phi(q') - \phi(q)$:

$$F(q, \pi) \leq \underbrace{\|\Delta E\|_F}_f \times \underbrace{\prod_{i=1}^L L_i}_g \times \underbrace{L_U}_h \times \underbrace{\|\nabla_{\mathbf{z}} C\|_2}_s \quad (5)$$

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

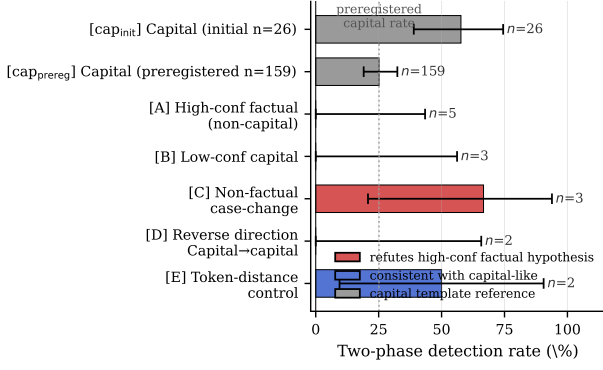


Figure 4. Disambiguation study (Pythia-410m, case-change perturbation). Bars show the two-phase detection rate; error bars are Wilson 95% CIs. Red bars (categories A and C) refute the “high-confidence factual content drives the pattern” hypothesis: A (high-confidence non-capital factual prompts, $n=5$) shows 0% while C (non-factual case-change, $n=3$) shows 66.7%. Blue bars (B, D, E) are consistent with capital-like behavior but at small n . The gray reference bars show capital-template rates (initial $n=26$ and preregistered $n=159$). Small per-category sample sizes limit the strength of this refutation (A Wilson CI $[0, 43\%]$, C Wilson CI $[21, 94\%]$).

The four factors (Figure 6) isolate: **(f)** embedding perturbation magnitude, **(g)** layer-wise propagation stability, **(h)** knowledge projection gain, and **(s)** softmax operating-point sensitivity.

Table 4. PIRI component scaling from 1.2B to 22B active parameters. We tabulate only s (softmax) and h (knowledge) because f (embedding perturbation) and g (layer stability) vary within $\pm 10\%$ across the range and are treated as approximately constant here; the $\pm 10\%$ variation is absorbed into the Combined row via the Lipschitz chain. The Combined $>$ Observed discrepancy is the non-independence caveat discussed in Eq. (5) and is not to be read as additive.

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.

J. Softmax Sensitivity and PIRI Factor s

The PIRI factor $s = \|\nabla_{\mathbf{z}} C\|_2$ from Eq. (5) 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 $\partial H/\partial p_i = -(\ln p_i + 1)$:

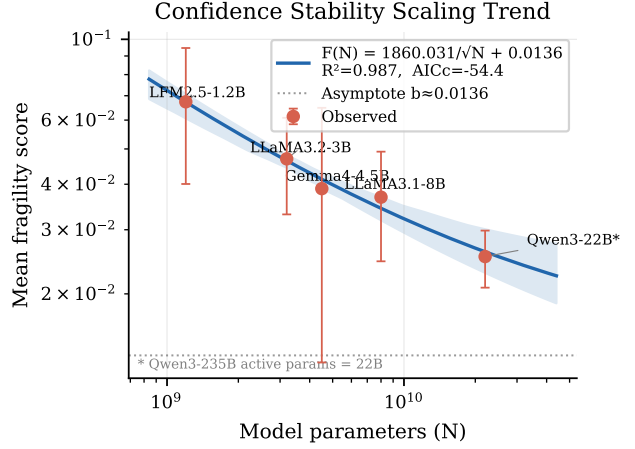


Figure 5. Confidence stability trend: best-fit $F(N) = a/\sqrt{N} + b$ ($R^2 \approx 0.987$). Dotted line shows the fitted asymptote $b \approx 0.014$ (not statistically identified at $n=5$). ΔAICc against a pure power law ($R^2 \approx 0.980$) ranges -0.92 to -2.11 depending on Model B initializations, staying within the $|\Delta| < 2$ indistinguishable band; only the monotonic decrease is a robust claim.

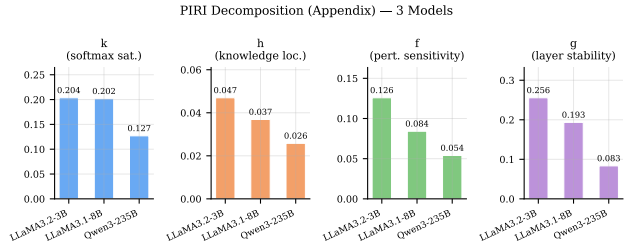


Figure 6. PIRI factor contributions across model scales (1.2B–22B active parameters), reported as a Lipschitz upper-bound chain. Numerical contributions ($h \approx 98\%$, $s \approx 35\%$ etc.) are non-independent and summing them is not valid; treat values as indicative only, not as additive attribution.

$$\frac{\partial C}{\partial z_j} = \frac{p_j}{\ln k} \left[(1 - p_j) \ln p_j - \sum_{i \neq j} p_i \ln p_i \right] \quad (6)$$

For the dominant (top-1) logit:

$$\frac{dH}{dz_1} = p_1 \left[\sum_{i=2}^k p_i \ln p_i - (1 - p_1) \ln p_1 \right] \quad (7)$$

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 4)—which is expected and analogous to the Hoeffding-Sobol ANOVA decomposition with dependent inputs (Sobol’, 1993).