Semantic Thermodynamics of Transformer Architectures: A Framework for Understanding Hallucination Constraints

Zulqarnain Ali *

Posted Date: 27 February 2026

doi: 10.20944/preprints202602.1962.v1

Keywords: transformers; hallucinations; information theory; semantic uncertainty; Fano's inequality; retrieval-augmented generation

 doi:10.20944/preprints202602.1962.v1

Article

Semantic Thermodynamics of Transformer Architectures: A Framework for Understanding Hallucination Constraints

Zulqarnain Ali

The Islamia University of Bahawalpur; zulqar445ali@gmail.com

Abstract

We develop *Semantic Thermodynamics*, an information-theoretic framework for analyzing hallucinations in transformer systems under finite resources. The central object is mutual information between latent facts and model outputs, together with Fano-style lower bounds on semantic error. We clarify the stochastic assumptions required for non-degenerate information measures, distinguish true data-generating uncertainty from model-implied uncertainty, and replace unsupported hard capacity formulas with explicit capacity surrogates tied to precision, context budget, and effective representational rank. Under standard identification assumptions, we derive a baseline bound

$$H_R \geq \max\left\{0,\, 1 - \frac{I(F;Y)+1}{\log M}\right\},$$

where H_R is hallucination rate, F is the latent semantic fact, Y is model output, and M is semantic cardinality. We also provide a distribution-dependent variant and a bottleneck-aware extension for retrieval-augmented generation. This paper contributes a mathematically consistent formulation, a tighter assumptions section, and concrete empirical protocols for estimation and falsification.

Keywords: transformers; hallucinations; information theory; semantic uncertainty; Fano's inequality; retrieval-augmented generation

---

1. Introduction

Hallucination remains a central failure mode of modern language models: outputs can be fluent and high-confidence while semantically incorrect. Mitigation methods including alignment, finetuning, and retrieval improve performance but do not eliminate failures. This pattern motivates a complementary question: beyond optimization quality, what lower bounds arise from finite information-processing resources?

This paper proposes an information-theoretic treatment of that question. We model semantic prediction as transmission of latent facts through a bounded inference pipeline. We then use mutual information and Fano-type converse bounds to characterize unavoidable error regimes.

Our goal is not to claim that architecture alone determines hallucinations. Rather, we isolate one layer of explanation: when effective semantic information throughput is below task complexity, nonzero semantic error is unavoidable even with idealized training. Throughout, "thermodynamics" is used as a metaphorical organizing label for resource-constrained information processing; the formal results in this paper are information-theoretic converses rather than a physical thermodynamic formalism.

 doi:10.20944/preprints202602.1962.v1

1.1. Contributions

This paper makes the following contributions:

1. It standardizes mutual information notation and removes degenerate uses of $I(F;Y \mid X)$ under deterministic decoding.
2. It separates true semantic distributions from model-induced distributions to avoid circular definitions.
3. It replaces inconsistent capacity formulas with explicit, assumption-dependent surrogates.
4. It weakens strict layerwise-loss claims to DPI-based monotonicity, with strict decrease only under explicit stochastic perturbations.
5. It adds distribution-dependent and bottleneck-aware bound variants, including a non-additive RAG analysis.
6. It expands related work to include semantic rate-distortion, memory-limited hallucination frontiers, and f-divergence minimax lower bounds.

2. Related Work

2.1. Information-Theoretic and Semantic Foundations

Our framework uses modern information-theoretic analyses of language models and transformers, with emphasis on semantic capacity, compression, and distortion [8,9,14]. We focus on contemporary (2020+) references and empirical-theoretical bridges that are directly actionable for current LLM systems. In autoregressive settings, directed-information viewpoints can complement $I(F;Y)$ when feedback and tokenwise causality are central [14].

2.2. Theory of Hallucinations and Transformer Limits

Recent analyses formalize hallucination incentives and uncertainty calibration in modern training/evaluation pipelines [4]. Complementary memory-theoretic results show high-confidence hallucination can be an unavoidable consequence of finite storage in sparse-fact regimes [10]. New 2026 studies also characterize multi-turn hallucination detection and relation-structure predictors of hallucination [11,12]. Transformer-specific analyses of attention pathologies and rank effects provide architectural mechanisms compatible with information bottlenecks [5–7]. Layerwise interpretability/probing work (e.g., tuned-lens analyses) complements our use of task-level semantic MI by exposing non-monotonic internal dynamics even when end-to-end converse bounds remain valid [16]. Recent layerwise hallucination studies and latent-risk detectors further support probing intermediate states for predictive risk signals [18,19].

2.3. Lower-Bound Methodology

Although we use Fano-style bounds as the primary instrument, we position them as baseline converse statements, not universally tight bounds. In practice, we pair these bounds with modern alignment and optimization results to evaluate real-system tightness [13]. Alternative inequalities relating error, entropy, and total variation can refine tightness analysis in some regimes [15]. Beyond average 0–1 error, interactive/Fano-generalized and f-divergence-based converses can target tail-risk functionals, including CVaR-style hallucination severity bounds [20]. We view these as natural extensions of the present framework.

3. Theoretical Framework

3.1. Semantic Prediction Setup

Let X denote an input prompt, $F \in \mathcal{F}$ the latent fact to be recovered, and Y the model output (token sequence). We assume an evaluator map ϕ that converts Y to a discrete fact estimate $\hat{F} = \phi(Y)$ for benchmarked tasks. In closed-world QA, ϕ is the canonical answer-mapping function (exact match or alias table). In open-ended generation, ϕ is instantiated by a pipeline: claim extraction, canonicalization, semantic clustering, and verifier scoring.

 doi:10.20944/preprints202602.1962.v1

Definition 1 (Hallucination Rate). *For fact-identification tasks with evaluator ϕ, hallucination rate is*

$$H_R := \Pr\big[\hat{F} \neq F\big]. \tag{1}$$

This definition is task-level and does not require token-level exact match.

3.2. True vs Model-Induced Uncertainty

Definition 2 (True Semantic Conditional Entropy). *For the data-generating distribution $p^\star(f \mid x)$,*

$$H^\star(F \mid X = x) := -\sum_f p^\star(f \mid x) \log p^\star(f \mid x). \tag{2}$$

Definition 3 (Model-Induced Surrogate Entropy). *Given calibrated model scores $q_\theta(f \mid x)$,*

$$\widetilde{H}_\theta(F \mid X = x) := -\sum_f q_\theta(f \mid x) \log q_\theta(f \mid x). \tag{3}$$

All converse bounds in this paper are written in terms of true quantities ($p^\star$). Model-induced quantities are used only as estimators/surrogates in experiments.

3.3. Mutual Information Convention

We use unconditional $I(F;Y)$ as the primary quantity. This avoids the deterministic-decoder degeneracy of $I(F;Y \mid X)$, which may collapse to zero when $Y = g(X)$.

Definition 4 (Semantic Information Throughput).

$$I_S := I(F;Y). \tag{4}$$

For stochastic decoding or randomized pipelines, conditional variants can be introduced with explicit noise variables. We model stochasticity as

$$Y = g_\theta(X, U), \qquad U \sim p(u), \tag{5}$$

where U captures decoding randomness (e.g., temperature sampling), retrieval randomness, or quantization noise. Deterministic decoding is the special case where U is constant. For blockwise analyses we similarly use

$$H^{(\ell+1)} = T_\ell(H^{(\ell)}, \xi_\ell), \tag{6}$$

with ξ_ℓ capturing stochastic perturbations (quantization noise, dropout-like perturbations, retrieval jitter, or randomized decoding controls).

3.4. Units and Normalization

All logarithms are base 2, so entropy and mutual information are in bits. We report hallucination bounds using $I(F;Y)$ measured per output sequence (or per prompt-response pair). Capacity surrogates are normalized to the same unit; if per-token surrogates are used in experiments, they must be multiplied by effective output length before being compared to sequence-level bounds.

4. Three Principles (Assumption-Dependent)

4.1. Principle 1: Capacity-Constrained Semantic Throughput

Principle 1 (Capacity-Constrained Information). *Under a fixed inference pipeline with finite numerical precision, finite context budget, and bounded effective representational rank, semantic throughput satisfies*

$$I(F;Y) \leq C_{surr}. \tag{7}$$

 doi:10.20944/preprints202602.1962.v1

We model C_{surr} as a surrogate envelope:

$$C_{\text{surr}} := \min\{C_{\text{param}}, C_{\text{ctx}}, C_{\text{arith}}\}, \tag{8}$$

with example components

$$C_{\text{param}} = \alpha\, b\, P_{\text{eff}}, \tag{9}$$
$$C_{\text{ctx}} = \beta\, b\, L_{\text{eff}}, \tag{10}$$
$$C_{\text{arith}} = \gamma\, b\, r_{\text{attn}}. \tag{11}$$

Here b is precision (bits), P_{eff} effective parameter budget, L_{eff} effective usable context, and r_{attn} an effective rank/degree-of-freedom proxy. Constants (α, β, γ) are task- and implementation-dependent and must be estimated empirically.

This replaces unsupported claims such as universal "$\log d$ bits per head" limits.

$$\text{(Assumption A1)} \quad F \to X \to Y, \qquad \text{(A2)} \quad I(F;Y) \le C_{\text{surr}}, \tag{12}$$

with (A2) treated as a falsifiable modeling assumption rather than a universal law. We additionally enforce minimal structural constraints for interpretability:

$$C_{\text{surr}} \ge 0, \quad \frac{\partial C_{\text{surr}}}{\partial b}, \frac{\partial C_{\text{surr}}}{\partial P_{\text{eff}}}, \frac{\partial C_{\text{surr}}}{\partial L_{\text{eff}}}, \frac{\partial C_{\text{surr}}}{\partial r_{\text{attn}}} \ge 0, \quad I(F;Y) \le H(F). \tag{13}$$

Hence non-RAG envelopes use $\min\{C_{\text{surr}}, H(F)\}$.

4.2. When A1 Fails

The Markov assumption $F \to X \to Y$ can fail in closed-book settings where model parameters encode prior correlations with F, or in tool-augmented pipelines with extra latent state. A broader graphical model is

$$F \to (X, Z) \to Y, \tag{14}$$

where Z captures latent priors, memory, or tool outputs not represented in X. In this case, the same Fano logic applies with $I(F;Y)$ measured under the expanded data-generating process; empirically, this requires logging/estimating variables contributing to Z.

4.3. Principle 2: Layerwise Non-Increase Under Markovization

Principle 2 (DPI Layer Monotonicity)**.** *If hidden states satisfy $F \to H^{(\ell)} \to H^{(\ell+1)}$, then*

$$I(F; H^{(\ell+1)}) \le I(F; H^{(\ell)}). \tag{15}$$

Strict decrease requires additional lossy mechanisms (e.g., stochastic quantization, dropout, or explicit compression).

Residual paths and layer normalization can violate naive layerwise Markov assumptions if treated componentwise. Our recommendation is to analyze complete transformer blocks as stochastic maps, then apply DPI at the block level. This statement is weaker but technically robust relative to asserting fixed positive decrements $\Delta_\ell > 0$ at every layer. Sufficient conditions for block-level Markovization are: (i) each block can be written as $H^{(\ell+1)} = T_\ell(H^{(\ell)}, \xi_\ell)$ with independent block noise ξ_ℓ, and (ii) no external side-channel state is injected into T_ℓ beyond $H^{(\ell)}$. Autoregressive KV caching can violate (ii) unless the cache state is included in the block state variable.

 doi:10.20944/preprints202602.1962.v1

4.4. Principle 3: Diminishing Returns as a Testable Hypothesis

Principle 3 (Sublinear Gains Hypothesis). *Across families of models and tasks, the empirical gain in semantic throughput often follows*

$$\frac{\partial I(F;Y)}{\partial C_{surr}} \propto C_{surr}^{-\delta}, \qquad \delta > 0, \tag{16}$$

but this is a hypothesis to be tested, not a theorem.

We treat this hypothesis as falsified for a model family if, after controlling for optimization budget and data mixture, fitted slopes fail to remain negative over the tested C_{surr} range or if cross-family estimates of δ are unstable under pre-registered robustness checks. Pre-registration should include model families, compute budgets, decoding policies, exclusion rules, and planned regressions before any slope fitting.

5. Bounds on Hallucination Rates

5.1. Baseline Fano Bound

Let $M = |\mathcal{F}|$ and $\hat{F} = \phi(Y)$.

Theorem 4 (Fano Hallucination Bound). *For finite $M \geq 2$,*

$$H_R \geq 1 - \frac{I(F;\hat{F}) + 1}{\log M} \geq 1 - \frac{I(F;Y) + 1}{\log M}, \tag{17}$$

where logs are base 2.

Proof. Fano gives

$$H(F \mid \hat{F}) \leq h_2(H_R) + H_R \log(M-1) \leq 1 + H_R \log M. \tag{18}$$

Using $H(F \mid \hat{F}) = H(F) - I(F;\hat{F})$ and $H(F) \leq \log M$ yields

$$H_R \geq 1 - \frac{I(F;\hat{F}) + 1}{\log M}. \tag{19}$$

Finally, $I(F;\hat{F}) \leq I(F;Y)$ by data processing through ϕ. □

5.2. Distribution-Dependent Variant

For non-uniform F, keeping $H(F)$ explicit gives

$$H_R \geq \frac{H(F) - I(F;Y) - 1}{\log(M-1)}. \tag{20}$$

This can be materially less pessimistic than replacing $H(F)$ by $\log M$.

Using the full multiclass Fano form also yields

$$H(F \mid \hat{F}) \leq h_2(H_R) + H_R \log(M-1), \tag{21}$$

which is preferable when M is large and H_R is not close to 0 or 1.

5.3. Capacity-Constrained Corollary

Corollary 1 (Capacity Envelope Bound). *If $I(F;Y) \leq \min\{C_{surr}, H(F)\}$, then*

$$H_R \geq \max\left\{0, 1 - \frac{\min\{C_{surr}, H(F)\} + 1}{\log M}\right\}. \tag{22}$$

 doi:10.20944/preprints202602.1962.v1

The bound is informative only when $\min\{C_{\text{surr}}, H(F)\} < \log M - 1$.

$$\text{Equivalent informativity condition:} \quad I(F;Y) < \log M - 1. \tag{23}$$

6. Empirical Validation Protocols

We outline protocols aimed at *estimable* quantities.

6.1. Protocol A: Closed-World Scaling with Controlled M

Construct synthetic QA datasets with known fact set size M. Measure H_R and estimate $I(F;\hat{F})$ from confusion matrices:

$$\widehat{I}(F;\hat{F}) = \widehat{H}(F) - \widehat{H}(F \mid \hat{F}). \tag{24}$$

Use bootstrap confidence intervals over prompts.

6.2. Protocol B: Precision and Context Stress Tests

For fixed task family, vary quantization level and usable context. Track changes in H_R, calibration error, and $\widehat{I}(F;\hat{F})$ to fit surrogate constants (α, β, γ). We use reproducible proxies:

1. $P_{\text{eff}}(\epsilon)$: smallest retained parameter count (via magnitude pruning) that keeps task score within ϵ of baseline;
2. $L_{\text{eff}}(\epsilon)$: largest removable context span whose ablation changes task score by at most ϵ;
3. r_{attn}: participation-ratio effective rank of attention maps, averaged across heads/layers.

Because L_{eff} and r_{attn} can be collinear, fits should report variance inflation diagnostics and regularized sensitivity analyses. To reduce confounding, fit α, β, γ using staged ablations (one factor perturbed at a time), then validate with joint perturbations and held-out tasks.

6.3. Protocol C: Layerwise Probing with Explicit Noise Controls

Probe $I(F;H^{(\ell)})$ across layers using probe classifiers or variational MI estimators. Include deterministic and stochastic inference settings separately to test when strict layerwise decrease appears. To reduce estimator-induced artifacts, we recommend at least two estimator families (e.g., InfoNCE-style lower bounds and matrix-based Rényi estimators) and synthetic controls with known MI for calibration. Report disagreement bands across estimators rather than a single curve. For selective prediction analyses, report risk-coverage curves and compare against strong abstention baselines (entropy thresholding, verifier confidence, and latent-risk probes).

6.4. Measurement Notes

For free-form generation, map outputs to fact states via a verifier and report verifier reliability. Calibrated likelihood-based surrogates should be reported with bias/variance diagnostics and sensitivity to calibration drift. Sensitivity analysis should include recalibration ablations (temperature scaling / isotonic variants) and report the resulting spread in $\widehat{I}(F;\hat{F})$ and bound slack; conclusions are considered robust only if qualitative trends are stable across these recalibrations. For the verifier pipeline, report inter-annotator agreement (or adjudication consistency) on a sampled subset and propagate verifier uncertainty to $\widehat{I}(F;\hat{F})$ confidence intervals.

Concrete estimators used in the proposed protocol design are:

1. plugin/confusion-matrix estimator for $\widehat{I}(F;\hat{F})$ in closed-world settings;
2. cross-entropy-based upper/lower surrogates after calibration;
3. variational MI lower bounds for representation-level analyses.

Uncertainty is quantified with nonparametric bootstrap confidence intervals over prompts.

 doi:10.20944/preprints202602.1962.v1

6.5. Pilot Synthetic Experiment Blueprint

To make feasibility explicit, we recommend a pilot experiment with $M \in \{16, 32, 64, 128\}$ and precision $b \in \{16, 8, 4\}$. For each (M, b):

1. generate balanced synthetic QA data with known latent fact F;
2. measure H_R, $\widehat{I}(F; \hat{F})$, and calibration error;
3. fit (α, β, γ) in C_{surr} and test whether observed points satisfy the predicted lower envelope.

This pilot does not prove the framework, but it directly tests whether predicted monotonic trends hold.

6.6. Semantic Cardinality for Open-Ended Generation

For open-ended tasks with variable fact sets, we replace raw M with an effective cardinality:

$$M_{\text{eff}}(x) := \exp\big(H(F \mid X = x)\big), \qquad \bar{M}_{\text{eff}} := \exp\big(\mathbb{E}_x[H(F \mid X = x)]\big). \tag{25}$$

This captures correlation and prompt-dependent fact support. In practice, $H(F \mid X = x)$ is estimated from verifier-induced fact labels with uncertainty intervals. The open-ended estimator uses: (i) claim induction from outputs, (ii) canonical entity-relation normalization, (iii) clustering of semantically equivalent claims, and (iv) verifier-calibrated assignment to fact clusters.

7. Extension to Retrieval-Augmented Generation

Additive capacity claims for RAG are generally optimistic. A safer decomposition uses chain rule:

$$I(F; Y, R) = I(F; R) + I(F; Y \mid R). \tag{26}$$

Hence,

$$I(F; Y, R) \le \min\{H(F),\ I(F; R) + C_{\text{tr}|R},\ C_{\text{ctx}}^{\text{RAG}}\}. \tag{27}$$

Proposition 1 (RAG Bottleneck Envelope)**.** *Define*

$$C_{eff}^{RAG} := \min\{H(F),\ I(F; R) + C_{tr|R},\ C_{ctx}^{RAG}\}. \tag{28}$$

Then

$$H_R \ge \max\left\{0, 1 - \frac{C_{eff}^{RAG} + 1}{\log M}\right\}. \tag{29}$$

This form captures retrieval quality limits, conditional transformer processing limits, and context-window bottlenecks simultaneously. For practical estimation, $I(F; R)$ is proxied by retrieval quality signals (recall@k, evidence coverage, and relevance calibration), while $C_{\text{tr}|R}$ is proxied by context-utilization diagnostics (attention-to-evidence mass, citation faithfulness, and answer sensitivity to evidence ablation). For ecosystem context and evaluation practices, see recent RAG surveys [17]. A practical calibration recipe is to construct synthetic retrieval tasks where F is known and retrieval corruption is controlled; estimate empirical relationships between retrieval metrics and $I(F; R)$, then transfer the calibrated map to real tasks with uncertainty bands.

Proposition 2 (Selective Prediction Extension)**.** *Let $q(x) \in \{0, 1\}$ indicate answer/abstain decisions with coverage $\kappa = \mathbb{E}[q(X)]$. Define conditional hallucination rate $H_R^{(\kappa)} = \Pr\big[\hat{F} \ne F \mid q(X) = 1\big]$. Then any lower bound on unconditional error induces a coverage-risk trade-off:*

$$\Pr\big[\hat{F} \ne F\big] = \kappa\, H_R^{(\kappa)} \ge \max\left\{0, 1 - \frac{I(F; Y) + 1}{\log M}\right\}, \tag{30}$$

so abstention can reduce conditional error only by reducing coverage.

 doi:10.20944/preprints202602.1962.v1

8. Discussion

The framework supports several practical conclusions.

First, lower bounds are most useful when paired with *task entropy* estimates; worst-case $\log M$ can be too loose for skewed domains.

Second, resource scaling should be bottleneck-aware: increasing one resource axis (e.g., parameters) cannot guarantee lower hallucination if context-use efficiency or retrieval precision is limiting.

Third, evaluation design matters: if benchmarks reward confident guessing more than calibrated abstention, realized hallucination can exceed information-theoretic baselines [4].

Limitations remain. Fact extraction maps can introduce measurement noise; semantic cardinality is difficult for open-ended generation; and surrogate capacities require calibration rather than closed-form universality. In API-constrained deployments where stable token log-probabilities are unavailable, practitioners should rely on confusion-matrix estimators, verifier-based surrogates, and retrieval-side observables, and report the additional uncertainty induced by these proxies. This framework is complementary to generalization-bound analyses (e.g., PAC-Bayes/Rademacher viewpoints): those upper-bound learnability and excess risk, while our converse perspective lower-bounds irreducible error under semantic information bottlenecks [21–23].

9. Conclusions

This paper presents a mathematically consistent information-theoretic account of hallucination constraints in transformer systems. The key contributions are: (i) non-degenerate MI definitions, (ii) explicit distinction between true and model-implied uncertainty, (iii) assumption-dependent capacity surrogates in place of unsupported universal formulas, (iv) DPI-consistent layerwise statements, and (v) bottleneck-aware RAG bounds. Together, these results improve rigor, falsifiability, and empirical relevance while preserving the core thesis: finite semantic information throughput implies irreducible hallucination risk on sufficiently complex tasks.

Acknowledgments: The author acknowledges discussions in the broader information theory and machine learning communities that informed the framing and technical presentation of this work.

References

1. J. Kaplan *et al.*, "Scaling Laws for Neural Language Models," arXiv:2001.08361, 2020.
2. J. Hoffmann *et al.*, "Training Compute-Optimal Large Language Models," arXiv:2203.15556, 2022.
3. P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv:2005.11401, 2020.
4. A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why Language Models Hallucinate," arXiv:2509.04664, 2025.
5. T. Nait Saada, A. Naderi, and J. Tanner, "Mind the Gap: a Spectral Analysis of Rank Collapse and Signal Propagation in Attention Layers," arXiv:2410.07799, 2024.
6. D. Heo and H. Choi, "Generalized Probabilistic Attention Mechanism in Transformers," arXiv:2410.15578, 2024.
7. M. Burger, S. Kabri, Y. Korolev, T. Roith, and L. Weigand, "Analysis of Mean-Field Models Arising from Self-Attention Dynamics in Transformer Architectures with Layer Normalization," arXiv:2501.03096, 2025.
8. Y.-Q. Zhao, Z.-M. Ma, G. Y. Li, S. Yuan, T. Ye, and C. Zhou, "Semantic Rate-Distortion Theory with Applications," arXiv:2509.10061, 2025.
9. A. de Andrade, A. Harell, and I. V. Bajic, "Rate-Distortion Optimization for Transformer Inference," arXiv:2601.22002, 2026.
10. A. Guo and J. Li, "Hallucination is a Consequence of Space-Optimality: A Rate-Distortion Theorem for Membership Testing," arXiv:2602.00906, 2026.
11. V. Rathore, S. Aneesh, and H. Singh, "Temporal Graph Network: Hallucination Detection in Multi-Turn Conversation," arXiv:2601.03051, 2026.
12. Y. Lu, Y. Liu, and H. Schütze, "Relational Linearity is a Predictor of Hallucinations," arXiv:2601.11429, 2026.
13. P. L. Chen, X. Li, X. Chen, and T. Lin, "Reward-free Alignment for Conflicting Objectives," arXiv:2602.02495, 2026.

 doi:10.20944/preprints202602.1962.v1

14. B. Bai, "Forget BIT, It is All about TOKEN: Towards Semantic Information Theory for LLMs," arXiv:2511.01202, 2025.
15. O. Rioul, "The Interplay between Error, Total Variation, Alpha-Entropy and Guessing: Fano and Pinsker Direct and Reverse Inequalities," *Entropy*, vol. 25, no. 7, 978, 2023.
16. N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, and S. Biderman, "Eliciting Latent Predictions from Transformers with the Tuned Lens," arXiv:2303.08112, 2023.
17. Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv:2312.10997, 2023.
18. "Layerwise Hallucination Dynamics in LLMs," arXiv:2403.20009, 2024.
19. "HALT: Hallucination Assessment via Latent Testing," arXiv:2601.14210, 2026.
20. "Interactive Fano and f-Divergence Converses for Tail-Risk Bounds," arXiv:2601.12027, 2026.
21. "Mutual Information and Recoverability for Understanding in LLMs," arXiv:2505.23790, 2025.
22. "A Unified Framework for Hallucination Detection and Mitigation," arXiv:2507.22915, 2025.
23. "Representation Evolution and Fine-Tuning Effects in Transformers," arXiv:2210.12696, 2022.