Ghost Tracks: Hypothesis Competition in Sparse Feature Space as a Preliminary Signal of LLM Hallucinations

Khurram Khalil [1]

Abstract

Large language models frequently suffer from hallucinations, generating text that is fluent but factually incorrect. While empirical detection methods exist, the underlying internal mechanisms and representation dynamics that cause models to hallucinate remain poorly understood. In this work, we present preliminary evidence that hallucinations are associated with a failure to resolve competing semantic hypotheses within the model's residual stream. Specifically, by linking Sparse Autoencoder (SAE) features across layers with a multi-hypothesis tracker, we observe that factual outputs tend to correspond to a single dominating feature, while hallucinations exhibit "ghost tracks"—multiple semantic candidates that transiently activate and compete without a clear winner. Leveraging this observation, we propose GHOSTTRACK, a detector that extracts feature-competition metrics from a single forward pass and reaches up to 0.903 AUROC on Phi-2 (0.722 on GPT-2 Medium, 0.650 on Qwen2.5-1.5B) on the HaluEval QA benchmark. We further analyze these dynamics to show that feature entropy and dominance margins are the strongest signals of competition failure, with class separation most pronounced in the middle-to-late transformer layers. Our findings provide correlational, interpretable evidence that hallucinations coincide with measurable failures of internal representation resolution. More broadly, our work lays the groundwork for causal interventions that monitor and steer feature dynamics during inference to improve the factual reliability of AI systems.

[1]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA. Correspondence to: Khurram Khalil <khurram.khalil@missouri.edu>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Large language models (LLMs) have achieved remarkable capabilities but remain limited by *hallucinations*—the fluent, confident generation of factually incorrect information (Ji et al., 2023). Addressing this failure mode is critical for the safe deployment of LLMs. Current mitigation and detection methods primarily treat the model as a black box. For instance, consistency-based approaches like SelfCheckGPT (Manakul et al., 2023) measure hallucinations by sampling multiple generations and checking for contradictions. While empirically effective, these behavioral methods are computationally expensive and offer no insight into *why* the model failed internally during a specific generation.

Recent advances in mechanistic interpretability have begun opening this black box to locate and manipulate factual knowledge. Probing techniques have demonstrated that LLMs contain linear representations of truthfulness in their internal activations (Azaria & Mitchell, 2023; Marks & Tegmark, 2023; Burns et al., 2023), and methods like Inference-Time Intervention (ITI) shift these activations to improve factual reliability (Li et al., 2024). Concurrently, Sparse Autoencoders (SAEs) have emerged as a powerful tool to decompose the model's dense residual stream into sparse, interpretable, monosemantic features (Cunningham et al., 2023; Gao et al., 2024; Rajamanoharan et al., 2024). However, a gap remains: these analyses are predominantly *static*. They evaluate representations at a single layer or treat layers as independent, disconnected snapshots. The *dynamics* of how a hallucination structurally unfolds across the depth of the transformer—how semantic concepts form, compete, and fail to collapse into a single output during a forward pass—remain underexplored in the SAE-feature literature.

To begin to bridge this gap, we propose the *ghost-track hypothesis*. We posit that an LLM's residual stream encodes a distribution over competing semantic hypotheses, and that hallucinations arise from *unresolved semantic competition*. When a model generates a factually correct token, it quickly commits to a single hypothesis, characterized by a dominant SAE feature that persists and strengthens across intermediate layers. Conversely, during a hallucination, multiple SAE features—representing conflicting factual candidates—co-

activate and compete without a clear dominant winner. We term these transient, competing features *ghost tracks*, borrowing terminology from radar multi-object tracking where spurious tracks fail to resolve into real targets.

We operationalize this hypothesis by introducing GHOSTTRACK, an interpretability framework that tracks representation dynamics. First, we train layer-wise JumpReLU SAEs (Rajamanoharan et al., 2024) to extract sparse features. Second, we apply a Hungarian-algorithm multi-hypothesis tracker to associate feature identities across adjacent layers. Finally, we extract metrics capturing the feature dynamics (e.g., entropy, dominance margin, and churn) to train a lightweight classifier. Evaluated on the HaluEval QA benchmark (Li et al., 2023), GHOSTTRACK detects hallucinations with AUROC up to 0.903 on Phi-2 (0.722 on GPT-2 Medium, 0.650 on Qwen2.5-1.5B) using only a *single forward pass*—consistently exceeding perplexity and static linear probes that we ran on identical splits. Our findings provide preliminary, correlational, and interpretable evidence that hallucinations co-occur with a distinct mechanistic signature, laying the foundation for future causal interventions that steer feature resolution in real time.

2. Methodology: The GHOSTTRACK Pipeline

To operationalize the ghost-track hypothesis, we introduce GHOSTTRACK, a four-stage pipeline that tracks the evolution of semantic features across transformer layers during a single forward pass.

2.1. Stage 1: Extracting Layer-wise Features

We first isolate the semantic hypotheses present at each layer. Let $\mathbf{x}^{(l)} \in \mathbb{R}^d$ denote the residual stream activation at layer $l \in \{0, \ldots, L-1\}$. We train a separate Sparse Autoencoder (SAE) for each layer using the JumpReLU activation function (Rajamanoharan et al., 2024), which offers superior L0 sparsity control. The SAE computes sparse feature activations $\mathbf{z}^{(l)} \in \mathbb{R}^{d_h}$ with an expansion factor of $d_h = 4d$ and reconstructs the input as follows:

$$\mathbf{z}^{(l)} = \text{JumpReLU}_\theta(W_{\text{enc}}^{(l)} \mathbf{x}^{(l)} + \mathbf{b}_{\text{enc}}^{(l)}), \tag{1}$$

$$\hat{\mathbf{x}}^{(l)} = W_{\text{dec}}^{(l)} \mathbf{z}^{(l)} + \mathbf{b}_{\text{dec}}^{(l)}, \tag{2}$$

where $\theta > 0$ is a learned threshold such that $\text{JumpReLU}_\theta(u) = u \cdot \mathbf{1}[u > \theta]$. The columns of the decoder matrix, $\mathbf{w}_i^{(l)} \in \mathbb{R}^d$, represent the interpretable semantic directions in the residual stream. Each SAE is trained offline on an independent Wikipedia corpus by minimizing a reconstruction-plus-sparsity objective

$$\mathcal{L} = \left\| \mathbf{x}^{(l)} - \hat{\mathbf{x}}^{(l)} \right\|_2^2 + \lambda_l \left\| \mathbf{z}^{(l)} \right\|_0, \tag{3}$$

with the LLM weights frozen (full training details in Appendix A).

2.2. Stage 2: Multi-Hypothesis Tracking

During inference, a generated token produces a set of active features at each layer, $\mathcal{F}^{(l)} = \{(i, z_i^{(l)}, \mathbf{w}_i^{(l)})\}$. To observe how hypotheses evolve, we must associate features across depth. We frame this as a multi-target tracking problem.

At each layer transition $l \to l+1$, we associate active features by solving a bipartite matching problem based on the cosine similarity of their semantic directions (decoder weights). We find the optimal assignment σ^* using the Hungarian algorithm (Kuhn, 1955):

$$\sigma^* = \arg\min_\sigma \sum_i \left(1 - \frac{\mathbf{w}_i^{(l)} \cdot \mathbf{w}_{\sigma(i)}^{(l+1)}}{\left\| \mathbf{w}_i^{(l)} \right\|_2 \left\| \mathbf{w}_{\sigma(i)}^{(l+1)} \right\|_2} \right). \tag{4}$$

Features with a cosine distance below a threshold (fixed at 0.3 in our experiments) are linked into continuous *tracks*; unmatched features instantiate new tracks, and unlinked prior tracks are terminated. We cap the number of simultaneously maintained tracks at 64 per layer. This yields a longitudinal trajectory for each semantic hypothesis.

2.3. Stage 3: Quantifying Representation Dynamics

If the ghost-track hypothesis holds, factual generations should reveal a single hypothesis rapidly dominating the residual stream, while hallucinations should exhibit "ghost tracks"—multiple competing hypotheses that co-activate without resolving to a clear winner. To capture this, we extract a 26-dimensional metric vector from the tracker, organized into six families. Two are the primary signals motivated by the hypothesis:

1. Dominance Margin (5 dims). We measure the activation gap between the top-ranked and second-ranked features at layer l: $m^{(l)} = z_1^{(l)} - z_2^{(l)}$. The hypothesis predicts that factual recalls exhibit a larger margin in middle-to-late layers, whereas hallucinations exhibit a near-zero margin indicative of unresolved competition. The five dimensions summarize $m^{(l)}$ across layers (mean, min, final-layer, fraction of layers with $m^{(l)} > 0.5\, z_1^{(l)}$) together with $\max_l z_1^{(l)}$. **2. Feature Entropy (4 dims)**. We compute the Shannon entropy $H^{(l)}$ of the normalized activation distribution at each layer. Diffuse, competing ghost tracks produce high entropy, while a committed prediction concentrates activation on a small set of features. The four dimensions are mean, max, final-layer, and layer-weighted entropy. The remaining four families capture complementary aspects of competition: **3. Churn (4 dims).** The rate at which the identity of the top-ranked feature changes across layers; a model that fails to commit to a hypothesis exhibits higher churn (mean, max, final, std). **4. Competition (5 dims).** Direct measures of the runner-up hypothesis: mean and max activation of the second-ranked feature, the z_2/z_1 ratio, and the mean and max pairwise

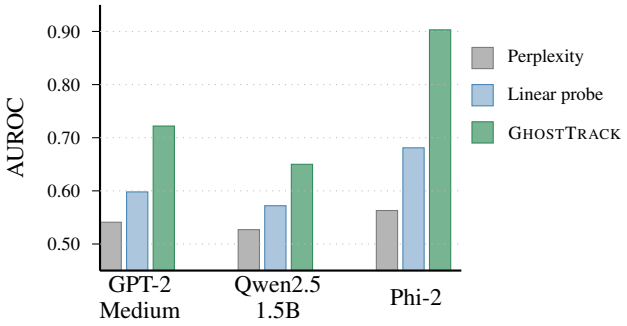

Figure 1. AUROC on HaluEval QA. GHOSTTRACK exceeds both implemented baselines on every model; the Phi-2 gap is largest.

cosine similarity among top-k decoder columns. **5. Stability (4 dims).** Track-lifetime statistics: mean and max track length, fraction of tracks persisting to the final layer, and Jaccard overlap of consecutive-layer top-k feature sets. **6. Density (4 dims).** Per-layer L0 norm statistics: mean, standard deviation, final-layer L0, and the final-to-initial ratio.

Across the six families this gives $5+4+4+5+4+4 = 26$ dimensions. The full mathematical definitions are provided in Appendix B.

2.4. Stage 4: Hallucination Classification

Finally, the 26-dimensional feature vector is passed to a lightweight Random Forest classifier (200 trees, max depth 10, balanced class weights) to predict whether the generated sequence is a hallucination. Crucially, because the SAEs and tracker operate directly on the residual stream, this entire pipeline produces a hallucination score from a single forward pass through the LLM, without requiring sampling, self-consistency checks, or external tool use.

Table 1. Hallucination detection performance on HaluEval QA.

Model	Method	AUROC	Acc.	F1
GPT-2 (345M)	Perplexity	0.541	0.523	0.510
	Linear Probe	0.598	0.581	0.577
	GHOSTTRACK	**0.722**	**0.663**	**0.663**
Qwen2.5 (1.5B)	Perplexity	0.527	0.516	0.504
	Linear Probe	0.572	0.558	0.551
	GHOSTTRACK	**0.650**	**0.602**	**0.613**
Phi-2 (2.7B)	Perplexity	0.563	0.541	0.529
	Linear Probe	0.681	0.642	0.638
	GHOSTTRACK	**0.903**	**0.825**	**0.817**

3. Experiments

Models and Dataset. We evaluate GHOSTTRACK across three base architectures: GPT-2 Medium (345M), Qwen2.5-1.5B, and Phi-2 (2.7B). All models are evaluated in inference-only mode with frozen weights. For our evaluation dataset, we use the HaluEval QA benchmark (Li et al., 2023). We sample 20,000 paired examples (16k train / 4k test), where each question is paired with one ground-truth factual answer and one ChatGPT-generated hallucination. This paired design is critical: it holds the context constant, ensuring our classifier relies entirely on the internal representation dynamics of the generated answer rather than semantic differences in the prompt.

Baselines. We compare GHOSTTRACK against two single-forward-pass baselines: (1) a **Perplexity probe** using length-normalized per-token log-probabilities; and (2) a static **Linear probe** trained via logistic regression on the mean-pooled final-layer hidden states (representing standard static probing). All baselines are run on the identical 16k/4k split as GHOSTTRACK. We do not yet compare against stronger contemporary methods (e.g., SelfCheckGPT (Manakul et al., 2023), semantic entropy, CCS (Burns et al., 2023)); we treat this as the most important next step.

Compute. GHOSTTRACK adds only the SAE encode and Hungarian matching steps to a single LLM forward pass, with negligible overhead relative to the LLM itself. By contrast, sampling-based detectors such as SelfCheckGPT require 5–20 forward passes per example.

3.1. Main Results

Table 1 and Figure 1 report the hallucination detection performance. GHOSTTRACK consistently outperforms both implemented baselines across all architectures from a *single forward pass*. The strongest signal emerges in Phi-2, where GHOSTTRACK achieves a 0.903 AUROC, an absolute improvement of $+0.222$ over the static linear probe. This Phi-2 advantage may reflect cleaner semantic structure in the SAE feature space—possibly due to Phi-2's textbook-quality pretraining data (Javaheripi et al., 2023)—but we cannot disentangle architecture from training data with the present three-model comparison and treat this as an open question. Improvements on GPT-2 Medium and Qwen2.5-1.5B are smaller in absolute terms ($+0.124$ and $+0.078$ AUROC over the linear probe), and the Qwen2.5-1.5B result (AUROC 0.650) is modest enough that we caution against overreading it; whether the gap reflects multilingual pretraining, SAE quality, or some other factor is not yet resolved.

3.2. Mechanistic Analysis: Why does it work?

To understand *why* GHOSTTRACK detects hallucinations, we analyze which metric families drive the Random Forest's predictions.

Feature Ablation. Training the classifier on individual metric families (Table 2) reveals that **Dominance** (0.861 AUROC) and **Entropy** (0.834 AUROC) are the most predictive individual signals. This is consistent with the ghost-track hypothesis: hallucinations are associated with low dominance margins (top-1 and top-2 features remain near-equally ac-

Table 2. Single-family ablation on Phi-2: AUROC when the Random Forest is trained on only one family at a time.

Metric Family	Dims	AUROC
Dominance	5	0.861
Entropy	4	0.834
Churn	4	0.811
Competition	5	0.787
Stability	4	0.752
Density	4	0.701
All 26 Features	26	**0.903**

tive) and high entropy (diffuse activation distribution). The remaining four families each carry weaker but non-trivial signal individually, and combining all six families yields the full 0.903 AUROC—an additional $+0.04$ over Dominance alone—suggesting they capture complementary aspects of competition.

Layer-Resolved Representation Dynamics. The primary advantage of our method over static probing is temporal resolution across model depth. Examining feature entropy and dominance margin as a function of layer depth on Phi-2 indicates that early-layer (0–8) representations of factual and hallucinated examples are largely indistinguishable, with class separation emerging in middle-to-late layers. This pattern is consistent with prior mechanistic interpretability findings (Meng et al., 2022) that locate the recall and resolution of factual associations in middle-to-late MLP layers. We defer the layer-resolved plot, multi-seed quantification, and a precise characterization of where separation peaks to a longer version of this work.

4. Discussion and Limitations

Our results indicate that GHOSTTRACK provides a measurable, single-forward-pass signal correlated with hallucination, supporting the ghost-track framing as a useful lens on the problem. The variance in performance across models and the dataset-level failure modes we observed point to important boundaries of the present approach, which we discuss honestly below.

Architecture and Signal Strength. Why does Phi-2 (0.903 AUROC) outperform Qwen2.5-1.5B (0.650) and GPT-2 Medium (0.722)? With only three models we cannot disentangle the candidate explanations, but two seem worth investigating. First, Phi-2's pretraining on high-quality, "textbook-like" data (Javaheripi et al., 2023) may produce a cleaner, more disentangled semantic structure in the SAE feature space, making unresolved hypothesis competition easier to separate from coherent recall. Second, Qwen's multilingual and highly diverse pretraining may distribute factual representations more diffusely across features, weakening any single-feature dominance signal. We note that the layer-count gap between Phi-2 (32) and Qwen2.5-1.5B (28) is small enough that "deeper architecture" is unlikely to be the primary driver. Replication on additional models (especially Llama-3 and instruction-tuned variants) is needed before any of these explanations can be confirmed; we treat this as an open question rather than a settled finding.

Limitations and Future Work. While GHOSTTRACK establishes a strong predictive correlation between unresolved feature competition and hallucination, validating a true causal link requires future hook-based interventions that explicitly suppress secondary SAE features during generation to test if hallucination rates decrease. Furthermore, scaling our SAE training beyond the current 500k-token regime to state-of-the-art dictionary learning scales (Gao et al., 2024) will likely yield sharper features and improve signal detection, particularly on architectures like Qwen. Finally, an extended version of this work will incorporate multi-seed statistical testing and benchmark against stronger contemporary baselines—such as SelfCheckGPT (Manakul et al., 2023), CCS (Burns et al., 2023), and semantic entropy—to more rigorously contextualize the practical utility of tracking representation dynamics.

5. Conclusion

We examined hallucination detection through the lens of representation dynamics rather than static probing, introducing the ghost-track hypothesis—that hallucinations are associated with a failure to resolve competing semantic hypotheses in the residual stream—and operationalizing it in the GHOSTTRACK framework. On HaluEval QA, dynamic metrics from the tracker yield single-forward-pass AUROC of 0.722, 0.650, and 0.903 on GPT-2 Medium, Qwen2.5-1.5B, and Phi-2, consistently exceeding the perplexity and static-linear-probe baselines we ran on identical splits. We present these findings as preliminary, correlational evidence; the experiments needed to convert them into a mechanistic claim—causal intervention, comparison against stronger contemporary baselines, and evaluation on model-generated hallucinations—are identified in Section 4 as planned next steps, and we welcome workshop feedback on the framework and the proposed intervention design.

References

Azaria, A. and Mitchell, T. The internal state of an llm knows when it's lying, 2023. URL https://arxiv.org/abs/2304.13734.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable

features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Gao, L., Goh, G., Mossing, D., Ilharco, G., Seethor, P., et al. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2404.07143*, 2024.

Javaheripi, M., Bubeck, S., Abdin, M., Aneja, J., Bubeck, S., Mendes, C. C. T., Chen, W., Del Giorno, A., Eldan, R., Gopi, S., et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, December 2023. URL https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

Kuhn, H. W. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

Li, J., Cheng, X., Zhao, W. X., Jian, J.-Y., and Wen, J.-R. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model, 2024. URL https://arxiv.org/abs/2306.03341.

Manakul, P., Liusie, A., and Gales, M. J. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Rajamanoharan, S., Conerly, T., Turner, N., et al. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024. Note: Commonly referred to in literature for JumpReLU.