HALLUSAE: DETECTING HALLUCINATIONS IN LARGE LANGUAGE MODELS VIA SPARSE AUTO-ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) are powerful and widely adopted, but their practical impact is limited by the well-known hallucination phenomenon. While recent hallucination detection methods have made notable progress, we find most of them overlook the dynamic nature and underlying mechanisms of it. To address this gap, we propose **HalluSAE**, a phase transition-inspired framework that models hallucination as a critical shift in the model's latent dynamics. By modeling the generation process as a trajectory through a potential energy landscape, HalluSAE identifies critical transition zones and attributes factual errors to specific high-energy sparse features. Our approach consists of three stages: (1) Potential Energy Empowered Phase Zone Localization via sparse autoencoders and a geometric potential energy metric; (2) Hallucination-related Sparse Feature Attribution using contrastive logit attribution; and (3) Probing-based Causal Hallucination Detection through linear probes on disentangled features. Extensive experiments on Gemma-2-9B, Llama-3.1-8B-Base, and Qwen3-8B-Base with independently developed SAE families demonstrate that HalluSAE achieves state-of-the-art hallucination detection performance and that its three-stage transition signature generalizes across model architectures.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated remarkable emergent abilities Wei et al. (2022), but still suffer from the well-known issue of hallucination, where the model generates content that is plausible yet incorrect Ji et al. (2023). This issue limits the use of LLMs in critical domains such as healthcare Singhal et al. (2023) and law Dahl et al. (2024), raising concerns about their reliability. Therefore, understanding the underlying causes of hallucination Elhage et al. (2021); Gao et al. (2025) and conducting effective hallucination detection has become a key research focus.

Mainstream hallucination detection methods are typically divided into black-box and white-box approaches.

Black-box methods, like SelfCheck-GPT Manakul et al. (2023) and CoVe Dhuliawala et al. (2024), rely on output consistency across multiple model runs, but are computationally intensive and lack interpretability. White-box methods instead leverage internal states, using heuristics such as perplexity Huang et al. (2025b); Friel & Sanyal (2023), attention entropy Li et al. (2025), or activation statistics. More advanced approaches, like SAPALMA Azaria & Mitchell (2023) and ITI Li et al. (2023b), use linear probes on

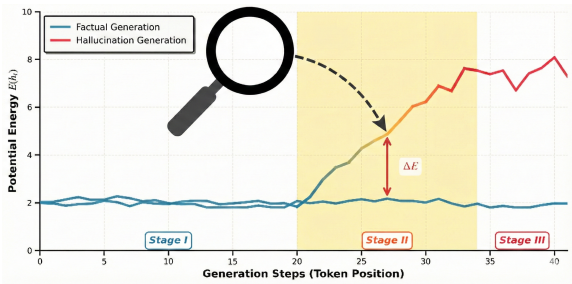


Figure 1: Illustration of Phase Transition in LLM's reasoning trajectories. The trajectories in potential energy space reveal three phases: early stability (Phase I), critical transition (Phase II, yellow highlight), and sustained error plateau (Phase III). Factual generation (blue) maintains low energy throughout, while hallucination (gradient color) undergoes abrupt energy increase (ΔE) during the transition zone, entering a persistent high-energy state.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the residual stream, while others model errors as constraint satisfaction in attention heads Yuksekgonul et al. (2024). Recent studies Sriramanan et al. (2024) have shown that internal-state methods offer significant speed advantages over black-box approaches. Although existing hallucination detection methods have achieved remarkable results, we find that they primarily analyze hallucination from the perspective of static feature representations or final outputs. Consequently, they do not fully consider the dynamic nature of hallucination as it unfolds during the generation process, and often overlook important information about how hallucinations develop over time. Moreover, focusing only on static features makes it easy for individual neurons to mix together many unrelated concepts Elhage et al. (2021), which can negatively impact detection performance.

Therefore, this paper proposes a physics-inspired perspective to bridge this gap. Rather than interpreting hallucinations as isolated surface errors, we posit that they reflect a critical phase transition within the model's latent dynamics (see Fig. 1). By conceptualizing the generation process as a trajectory over a potential energy landscape, we reveal a distinctive mechanism: factual errors emerge when internal representations, propelled by specific high-energy sparse features, abruptly transition from a low-energy "truth attractor" to a persistent, unstable high-energy plateau. This dynamic viewpoint motivates a key insight: *by monitoring the temporal evolution of latent potential energy in LLM, one can potentially localize phase transition zones that serve as reliable indicators for hallucination detection.*

To this end, we propose a systematic detection framework, **HalluSAE**, grounded in the phase transition perspective. Our approach follows a coarse-to-fine pipeline comprising three stages: (1) *Potential Energy Empowered Phase Zone Localization*, where we leverage the Gemma Scope Sparse Autoencoder (SAE) to disentangle the residual stream and introduce a geometric potential energy metric, allowing us to efficiently localize "phase transition zones" by identifying layers with exponential energy growth and narrowing the search space to a few critical transition points; (2) *Hallucination-related Sparse Feature Attribution*, in which we apply Contrastive Direct Logit Attribution (DLA) within these high-energy regions to quantify the contributions of individual sparse features to incorrect outputs, thereby precisely isolating "hallucination-inducing features" responsible for specific error patterns; and (3) *Probing-based Causal Hallucination Detection*, where we train linear probes on the selected sparse features to enable accurate and efficient inference-time hallucination detection, establishing a causal link between internal model dynamics and output errors. Collectively, HalluSAE offers a targeted, interpretable, and scalable solution for hallucination detection by explicitly modeling the dynamic processes underlying factual errors in large language models.

To assess the effectiveness and generalizability of our approach, we conduct experiments on Gemma-2-9B, Llama-3.1-8B-Base, and Qwen3-8B-Base using both in-distribution (HaluEval) and out-of-distribution (TriviaQA) benchmarks, where our method consistently achieves superior hallucination detection performance over existing baselines. The main contributions of this paper are summarized as follows:

- We introduce HalluSAE, a framework that approaches hallucination from a phase transition perspective, enabling dynamic and process-aware hallucination detection rather than relying solely on static representations.
- We propose Potential Energy Empowered Phase Zone Localization, Hallucination-related Sparse Feature Attribution, and Probing-based Causal Hallucination Detection in HalluSAE to precisely identify and attribute hallucination-inducing features.
- We conduct comprehensive experiments with three independently developed model–SAE pairs, which demonstrate that our method achieves new state-of-the-art results on in-distribution (HaluEval) tasks and out-of-distribution (TriviaQA) benchmarks.

2 Related Work

Hallucination in Large Language Models. Despite the impressive emergent abilities of LLMs Wei et al. (2022), hallucinations remain a persistent challenge, especially in high-stakes domains such as healthcare Singhal et al. (2023) and law Dahl et al. (2024). Recent taxonomies categorize these errors as input-conflicting or fact-conflicting, attributing them to data divergence or the compounding of early generation mistakes Zhang et al. (2025b); Huang et al. (2025a). However, while these classifications are well-studied, the mechanistic understanding of *how* and *why* hallucinations dy-

namically arise during inference is still lacking Orgad et al. (2025); Zhang et al. (2025c). Most existing methods treat hallucination as a static phenomenon based on final output or representations Ji et al. (2023); Huang et al. (2025a); in contrast, this work analyzes hallucination from a dynamical perspective for its detection.

Hallucination Detection Methods. Existing hallucination detection strategies fall into black-box and white-box categories. Black-box methods rely on output consistency checks Manakul et al. (2023); Dhuliawala et al. (2024), external LLM judges Benkirane et al. (2024), or Bayesian sequential estimation Wang et al. (2023a), but are often computationally costly and lack interpretability. Zero-shot attention methods like AGSER Liu et al. (2025) also require multiple inference passes, increasing latency. White-box methods leverage internal states. Early work by Xu et al. Xu et al. (2023) pioneered model introspection in NMT. Recently, HARP Hu et al. (2026) and GSP Noël (2025) utilized SVD projections and graph spectral energy, respectively, to detect errors. However, these dense or topological approaches often suffer from neuron polysemanticity Elhage et al. (2021), making it difficult to isolate causal features. In contrast, we take a sparse and dynamic perspective, utilizing SAEs to precisely localize hallucination causes.

Sparse Autoencoders and Their Application. Sparse Autoencoders (SAEs) have become a key interpretability tool, decomposing activations into human-interpretable features Bricken et al. (2023); Huben et al. (2024); Lieberum et al. (2024), with applications in circuit discovery Marks et al. (2025) and privacy Frikha et al. (2025). In hallucination research, SAFE Abdaljalil et al. (2025), SSL Hua et al. (2025), SAVE Park et al. (2026), and RAGLens Xiong et al. (2026) employ SAEs for RAG-specific tasks, but rely on static activations and task-specific selection. We instead introduce a geometric potential energy framework that tracks temporal dynamics and localizes causal features—the first to leverage dynamic SAE geometry for general factual hallucination detection. Open SAE suites now cover more models: Llama Scope for Llama-3.1-8B-Base and Qwen-Scope for Qwen3-8B-Base He et al. (2024); OpenMOSS-Team and Fudan NLP (2024); Deng et al. (2026); Qwen Team (2026). This paper employs these independently developed suites for hallucination detection.

3 Pre-analysis on Hallucination From the Phase Transition View

As discussed above, this work analyzes hallucination in LLMs from a phase transition perspective and proposes a hallucination detection framework based on this view. Before introducing our HalluSAE method, we first present the key analytical tools and some preliminary findings. Specifically, to analyze hallucination in LLMs, we primarily leverage the two following tools: . Detailed introductions to Sparse Autoencoders (SAEs) and Geometric Potential Energy (GPE) are provided in Appendix A.

Using both analytical tools, as shown in Fig. 2, we first conduct a series of controlled experiments to pre-analyze the internal dynamics of LLM reasoning. Our experimental results illustrated in Fig. 3 reveal several key patterns. First, layer-wise energy analysis (Fig. 3a) shows that the energy difference between hallucination and factual samples is negligible in the early layers, increases steadily in the middle layers, and surges sharply in the late layers—exhibiting a clear phase transition Meng et al. (2022); Geva et al. (2021). Statistical analysis (Cohen's $d = 1.64$, $p < 0.001$) further confirms that hallucinations arise within specific layer intervals, rather than being uniformly distributed across the network. Second, at the feature level (Fig. 3b), we observe that only a small subset of sparse fea-

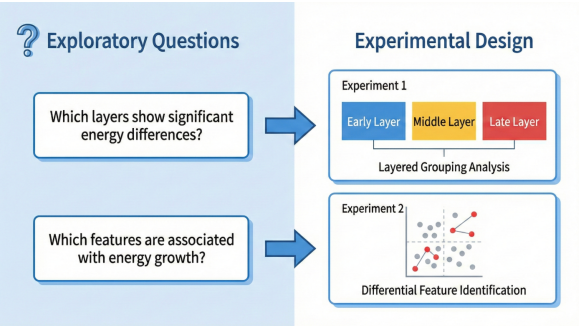


Figure 2: Illustration of the Exploratory Experimental Design. We conduct two complementary experiments: Exp 1 investigates layer-wise energy distribution by dividing 42 layers into Early/Middle/Late groups and comparing GPE differences; Exp 2 identifies microscopic feature-level contributions by analyzing differential features that exhibit high activation in hallucination samples but low activation in factual samples.

tures—those highly activated in hallucination cases—are responsible for the majority of the energy increase. This heavy-tailed distribution indicates that hallucinations are driven by a few high-energy features, rather than by diffuse, low-level noise. In summary, our analysis yields two main findings:

- **Finding 1**: Hallucination emerges as a dynamic process during generation, with clear phase transitions occurring at specific layer intervals rather than manifesting as a static property throughout the network.
- **Finding 2**: Hallucination is predominantly driven by a small subset of high-impact sparse activations, rather than by broad or diffuse noise, highlighting the critical role of a few key features in triggering the transition to erroneous outputs.

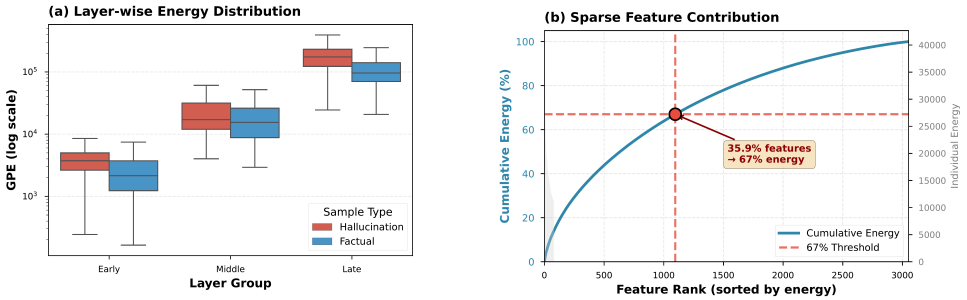


Figure 3: Experimental Results on Hallucination Dynamics. (a) Layer-wise energy distribution (grouped analysis). Grouped box plots reveal that hallucination samples' GPE exhibits significant escalation from Early to Late layer groups (***$p < 0.001$). Error bars indicate 95% confidence intervals (bootstrap, n=200 per group). (b) Sparse feature contribution analysis. Cumulative energy contribution curve demonstrates that a small fraction of differential features account for the majority of total energy increase, exhibiting a highly skewed distribution. Gray bars represent individual feature energy contributions.

These two findings motivate our approach: *by dynamically monitoring potential energy changes in the sparse feature space, we can effectively detect phase transitions that indicate hallucination.* Next, we introduce our motivated approach HalluSAE in detail.

4 METHOD

Given an input prompt, our goal is to predict whether the model's output will contain hallucinations before generation is complete—by analyzing activations from selected layers inside the LLM. Motivated by our earlier findings, we propose HalluSAE, a three-stage framework grounded in the dynamics of sparse feature activations and phase transitions.

HalluSAE operates as follows: (1) Potential Energy Empowered Phase Zone Localization leverages sparse autoencoders and a geometric potential energy metric to identify critical layer intervals where phase transitions indicative of hallucination occur. (2) Hallucination-related Sparse Feature Attribution uses contrastive logit attribution to pinpoint a small set of sparse features most responsible for hallucination within these layers. (3) Probing-based Causal Hallucination Detection conducts linear probing on disentangled features to detect and validate causal relationships between feature activation patterns and hallucination outcomes. In the following subsections, we detail the methodology and implementation of each stage.

STAGE I: POTENTIAL ENERGY EMPOWERED PHASE ZONE LOCALIZATION

The primary goal of this module is to identify which layers within the model are most susceptible to hallucination, by detecting phase changes through differences in Geometric Potential Energy (GPE) between hallucination and factual samples. By quantifying how the GPE gap evolves across layers, we aim to pinpoint the specific regions in the network where the internal state shifts most sharply toward hallucination. The operations are as follows:

For each layer ℓ, we compute the average GPE difference as

$$\Delta E_\ell = \mathbb{E}_{x\in\mathcal{H}}[E_\ell(x)] - \mathbb{E}_{x\in\mathcal{F}}[E_\ell(x)], \tag{1}$$

where $\mathcal{H}$ and $\mathcal{F}$ denote the sets of hallucination and factual samples, respectively. To characterize the dynamics of this difference, we define the relative growth rate

$$\gamma_\ell = \frac{\Delta E_\ell - \Delta E_{\ell-1}}{\Delta E_{\ell-1}} \times 100\%. \tag{2}$$

A stable or oscillatory γ_ℓ suggests the model remains in a consistent state, while a sequence of sustained positive growth indicates a transition toward hallucination.

To robustly identify the onset of this transition, we locate the layer interval $[\ell_{\text{start}}, \ell_{\text{end}}]$ starting from the earliest layer ℓ_{start} that exhibits three consecutive positive growth rates,

$$\ell_{\text{start}} = \min\left\{\ell \mid \gamma_\ell > 0, \gamma_{\ell+1} > 0, \gamma_{\ell+2} > 0\right\}. \tag{3}$$

We then define the endpoint ℓ_{end} as the layer where ΔE_ℓ reaches its post-onset maximum,

$$\ell_{\text{end}} = \arg\max_{\ell\geq\ell_{\text{start}}} \Delta E_\ell. \tag{4}$$

This process allows us to automatically detect the "phase transition zone"—the interval in which the model's internal state shifts most dramatically toward hallucination—by analyzing potential energy dynamics. The identified zone provides a principled basis for focusing subsequent fine-grained analysis on the most vulnerable layers for hallucination emergence.

Stage II: Hallucination-related Sparse Feature Attribution

After localizing the layers most vulnerable to hallucination, the core challenge is to understand which internal mechanisms within these layers are responsible for the phenomenon. Due to the polysemanticity and entanglement present in raw model activations, direct analysis is often unreliable. We therefore employ the sparse and overcomplete representations provided by the SAE as a principled basis for disentangling and attributing semantic effects in the network. The interpretability and sparsity of SAE features allow for precise tracing of how specific activations influence model outputs.

Specifically, within the identified phase transition layers, we aim to isolate the subset of SAE features that causally drive hallucination Elhage et al. (2021); Wang et al. (2023b). For each feature i, we define its direct logit attribution to a target token t as

$$\mathcal{A}_i^{(t)} = \left(s_i\, \mathbf{w}_{\text{dec}}^{(i)}\right)^\top \mathbf{W}_{\text{U}}[:, t], \tag{5}$$

where s_i is the feature activation, $\mathbf{w}_{\text{dec}}^{(i)}$ is the SAE decoder direction, and $\mathbf{W}_{\text{U}}$ is the unembedding matrix. This quantifies the linear pathway from a sparse feature to its impact on the model's output space.

Yet, hallucination generation is inherently a contrastive process: it depends not only on increasing the likelihood of the incorrect token but also on suppressing the correct one Li et al. (2023b). To capture this, we introduce the Contrastive Direct Logit Attribution (C-DLA), defined as

$$\mathcal{C}_i = s_i \cdot \left(\mathbf{w}_{\text{dec}}^{(i)}\right)^\top \left(\mathbf{W}_{\text{U}}[:, t_{\text{wrong}}] - \mathbf{W}_{\text{U}}[:, t_{\text{correct}}]\right). \tag{6}$$

Here, the vector $\Delta\mathbf{v} = \mathbf{W}_{\text{U}}[:, t_{\text{wrong}}] - \mathbf{W}_{\text{U}}[:, t_{\text{correct}}]$ represents the semantic direction from the correct answer to the hallucinated answer in output space. A positive value of $\mathcal{C}_i$ indicates that feature i both elevates the hallucinated token and suppresses the factual one, thus promoting hallucination.

For each prompt, we compute $\mathcal{C}_i$ for all features in the critical layers, rank them by magnitude, and select the Top-100 features as the critical set. This procedure yields a compact collection of features that are not merely correlated with hallucination, but mechanistically responsible for shifting model outputs away from factuality. These identified features serve as the analytical and causal substrate for detection and intervention in the next stage.

Model	SAE	Type	Width	Stable	Transition	Escape/Plateau
Gemma-2-9B	Gemma Scope	JumpReLU	131K	L0–22	L23–35	L36–41
Llama-3.1-8B-Base	Llama Scope	JumpReLU	131K	L0–22	L23–28	L29–31
Qwen3-8B-Base	Qwen-Scope	TopK	65K	L0–15	L16–23	L24–35

Table 1: Model, SAE, and Phase-Zone Summary. Layer indices are zero-based; the transferable object is the three-stage structure rather than a fixed absolute layer range.

Method	Gemma-2-9B		Llama-3.1-8B-Base		Qwen3-8B-Base	
	ID	OOD	ID	OOD	ID	OOD
LN-Entropy	54.93	70.94	62.62	73.62	61.61	65.74
Semantic Entropy	62.77	71.85	56.66	74.65	61.96	69.54
Lexical Similarity	64.25	65.40	56.21	67.77	60.70	62.83
SelfCheckGPT	63.78	63.53	66.87	68.15	65.15	64.09
EigenScore	60.53	61.51	58.11	56.28	58.72	55.98
MHAD	64.18	61.01	59.38	63.72	63.58	65.22
SAPLMA	79.14	64.33	69.09	58.95	68.24	66.37
HaloScope	84.37	78.94	72.46	66.39	82.30	76.38
HalluSAE	**92.86**	**80.44**	**88.67**	**80.05**	**90.81**	**79.32**

Table 2: Hallucination Detection AUC Across Three Model Families. All values are percentages. ID denotes HaluEval and OOD denotes TriviaQA. Best results are highlighted in gray.

Stage III: Probing-based Causal Hallucination Detection

Building on the identification of critical SAE features within the phase transition zone, the final step is to leverage these features for hallucination detection. The objective is to determine, for a given prompt and its candidate output, whether the activation pattern across this feature set reliably signals the presence of hallucination.

In particular, for each sample x, let $\mathbf{s}_{\mathcal{T}}(x) \in \mathbb{R}^{|\mathcal{T}|}$ denote the concatenated activations of the selected feature set $\mathcal{T}$ across all layers in the identified zone. Rather than relying on granular error categorization, we adopt a unified detection strategy. We construct the critical feature set $\mathcal{T}$ by selecting the top-ranked sparse features based on their average C-DLA magnitude across the entire training dataset. A single logistic regression Probe is then trained on these high-impact features Alain & Bengio (2017) to distinguish between hallucination and factual generations. This holistic approach leverages the semantic disentanglement of SAEs, enabling a compact feature set to capture diverse hallucination patterns—ranging from numerical errors to entity substitutions—without requiring explicit type-specific engineering.

The detection model is formulated as an ℓ_1-regularized logistic regression, optimizing

$$\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + \exp(-y_i \, \boldsymbol{\theta}^\top \mathbf{s}_{\mathcal{T}}(x_i))\right) + \lambda \|\boldsymbol{\theta}\|_1, \tag{7}$$

where $y_i \in \{+1, -1\}$ is the hallucination label, and λ is selected by cross-validation. This approach yields a sparse, interpretable classifier in which nonzero coefficients directly indicate the contributions of individual features to hallucination prediction.

During inference, for any input, the model extracts the activation pattern across $\mathcal{T}$ and outputs a hallucination probability. Because the cached SAE projection and linear probe are applied during the same autoregressive pass, the current token can be flagged after the selected internal layers are evaluated but before that token is decoded and before the complete response is generated.

5 Experiments

Experimental Setup

Datasets. We evaluate our method on two complementary benchmarks. For in-distribution (ID) evaluation, we use the HaluEval dataset Li et al. (2023a), consisting of 1,800 samples (900 hal-

lucination, 900 factual) with 80% allocated for training (1,440 samples) and 20% for testing (360 samples). All samples undergo dual annotation: automatic verification by GPT-4o followed by human expert review. This work focuses specifically on *fact-conflicting hallucinations*: linguistically plausible generations that contradict stable, externally verifiable facts. The annotation protocol operationalizes this scope through entity verification, numerical precision, contradiction, and relevance checks; temporally changing claims are outside our evaluation scope. The dataset is carefully balanced across generation length, error types (50% numerical, 50% entity), and question difficulty. For out-of-distribution (OOD) evaluation, we construct a TriviaQA Joshi et al. (2017) test set containing 4,000 samples (2,000 hallucination, 2,000 factual) to assess zero-shot cross-domain generalization. Compared to HaluEval, TriviaQA features more colloquial phrasing, broader knowledge coverage spanning geography, history, people, and organizations, and greater diversity in question formats.

Baselines. We compare against eight representative detection methods spanning four paradigms. Uncertainty-based approaches include LN-Entropy Manakul et al. (2023); Xiao & Wang (2021) and Semantic Entropy Farquhar et al. (2024). Consistency-based methods comprise Lexical Similarity Manakul et al. (2023); Wang et al. (2023b) and SelfCheckGPT Manakul et al. (2023). Internal state-based approaches feature EigenScore Chen et al. (2024) and MHAD Zhang et al. (2025a). Supervised probing methods include SAPLMA Azaria & Mitchell (2023) and HaloScope Du et al. (2024). Detailed implementation configurations for all baselines and training details are provided in the Supplementary Material.

Main Results

Quantitative Comparison. Table 2 summarizes performance across all methods and model families. On Gemma, HalluSAE attains an AUC of 0.9286 on the in-distribution test set and 0.8044 on out-of-distribution TriviaQA, establishing new state-of-the-art results in both regimes. Unsupervised baselines display limited efficacy: the strongest (Semantic Entropy) achieves 0.7185 AUC on OOD data, while HalluSAE surpasses this by 11.9% (0.8044 vs. 0.7185), highlighting the insufficiency of surface-level statistical cues for capturing the causal dynamics of hallucination. Consistency-based approaches perform worse still: HalluSAE outperforms SelfCheckGPT (0.6353 OOD AUC) by 16.91 AUC points, corresponding to a 26.6% relative improvement over SelfCheckGPT. Among supervised baselines, HaloScope represents the previous state-of-the-art, utilizing dense hidden states (3,584 dimensions) across multiple layers. HalluSAE outperforms HaloScope by 10.1% on in-distribution data (0.9286 vs. 0.8437) and by 1.9% on OOD data (0.8044 vs. 0.7894).

This improvement is attributable to three core advantages of sparse autoencoder features: (i) semantic disentanglement in the 131k-dimensional sparse space mitigates the polysemanticity of dense activations, where individual neurons encode multiple unrelated concepts; (ii) C-DLA enables precise, mechanistic attribution of hallucination to individual features, rather than diffuse patterns over thousands of variables; and (iii) semantic-level representations demonstrate superior robustness to domain shift compared to raw activation statistics. Notably, these gains are achieved using only 100 sparse features—a 35.8$\times$ reduction relative to HaloScope's single-layer 3,584-dimensional representation—underscoring the efficiency and interpretability of the approach.

Feature ID	H/F Ratio	Purity	Mean Act.
L27-9659	22.92	10/10 (100%)	8.03
L28-87984	81.52	10/10 (100%)	8.54
L23-71479	16.57	9/10 (90%)	7.80
L24-35793	12.41	10/10 (100%)	6.86

Table 3: Identified high-purity hallucination-driving features.

Cross-Model Generalization. Table 2 extends the complete comparison to Llama and Qwen. HalluSAE remains the strongest method on both architectures, reaching 88.67/80.05 ID/OOD AUC on Llama and 90.81/79.32 on Qwen. Table 1 shows that all three models reproduce the Stable–Transition–Escape/Plateau pattern, although the critical interval is model-specific. The transition spans L23–28 in Llama and L16–23 in Qwen, whose endpoint is the peak layer L23; all six Llama and all eight Qwen transition layers satisfy the layer-wise significance filter. Thus, HalluSAE transfers the localization principle rather than a fixed absolute layer range. The Qwen result is particularly informative because its TopK SAE differs from the JumpReLU SAEs used for Gemma and Llama.

Complete AUC, accuracy, and recall values are reported in the Supplementary Material. On the harder numerical subset of TriviaQA, HalluSAE obtains 78.91, 78.43, and 77.86 AUC on Gemma, Llama, and Qwen, respectively.

Qualitative Comparison. We analyze the top-50 C-DLA-ranked features, selecting those with (1) hallucination-exclusive high activations and (2) significant activation gaps between hallucinated and factual samples. This yields four high-purity features (Table 3), whose highest-activation samples reveal distinct error modes: L27-9659 and L28-87984 correspond to numerical substitution, while L23-71479 and L24-35793 reflect entity replacement. Fig. 4 visualizes these consistent feature-to-error correspondences, confirming that our method captures systematic semantic deviations rather than noise. These findings demonstrate three advantages of our approach: (1) robust feature-to-pattern correspondence, indicating hallucinations are driven by specific neural mechanisms; (2) genuine semantic interpretability via SAE features, transparently mapping to functions like numerical shifts and entity confusion; and (3) potential for targeted causal interventions, offering promising directions for SAE-based hallucination mitigation.

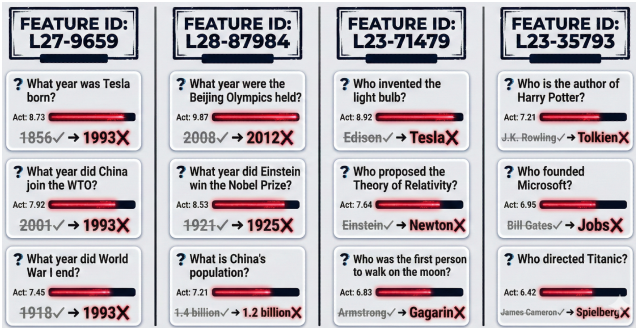
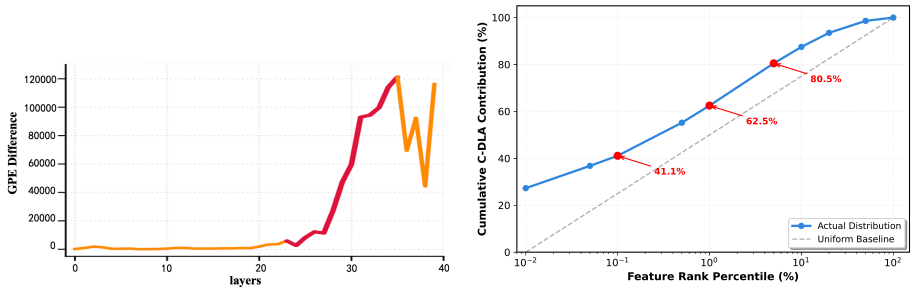


Figure 4: Examples of Samples with High-Activation on Identified Sparse Features. Each feature exhibits stable and interpretable error patterns: L27-9659 shows fixed year substitution (always outputs "1993"), L28-87984 demonstrates numerical shift patterns (+4 years, -0.2 billion), L23-71479 drives domain-specific entity confusion (Tesla/Edison, Newton/Einstein), and L24-35793 triggers cross-domain celebrity substitution (Tolkien/Rowling, Jobs/Gates).

Ablation Study

In this section, we conduct a series of ablation studies to systematically investigate the contributions of HalluSAE's key components. Specifically, we examine the effectiveness of potential-energy-based phase-zone localization, compare C-DLA with alternative sparse-feature attribution strategies, and evaluate the complementary roles of the localization and attribution stages.



(a) Macroscopic GPE dynamics (b) Microscopic C-DLA concentration

Figure 5: Macroscopic and Microscopic Signatures of Hallucination. (a) Layer-wise GPE difference reveals stable (L0–22), transition (L23–35), and plateau (L36–41) phases for Gemma. Error bars are 95% bootstrap confidence intervals ($n = 900$ per group). (b) The cumulative C-DLA curve shows that the top 0.1% (131/131,072) and top 1% (1,310/131,072) of features account for 41.1% and 62.5% of attribution, respectively (Gini 0.912 versus 0.414 for random attribution).

Ablation Study on the Potential Energy Empowered Phase Zone Localization Stage. The complete layer-selection results are reported in Appendix B, Table 4. To assess the necessity of precise phase localization, we compare five layer selection strategies using 100 features for 13-layer selections versus 420 features for all 42 layers. Table 4 shows our transition zone (L23–35) achieves the best performance (ID: 0.93, OOD: 0.80 AUC), matching the full 42-layer baseline with $4.2\times$ fewer

features. Early (L0–12) and late (L29–41) layers underperform (0.81/0.73 and 0.88/0.78 AUC), while random selection performs worst (0.83/0.76 AUC), confirming our energy-based localization captures genuine causal structure. Fig. 5 reveals three GPE phases: stable early (L0–22) with near-zero difference, a sharp transition (L23–35) with a 20.7-fold energy surge (Cohen's $d = 1.64$, $p < 0.001$), and a sustained plateau (L36–41, $d = 0.12$ vs. transition, $p = 0.342$).

These results validate that (1) the transition zone captures the causal shift, (2) early/late layers add no benefit, and (3) random selection fails, confirming all design components are essential. The All-42 Gemma detector is comparable but not identical to the localized detector: its OOD AUC is lower by 0.01, or one AUC percentage point. The same localized advantage appears for Llama and Qwen despite their different transition intervals. Additional permutation controls, centroid robustness, and label-noise results are reported in the Supplementary Material.

Model	Selection	Pur.	ID	OOD	Acc.
Gemma	Wrong-only DLA	73.00	0.8800	0.7900	81.20
Gemma	Correct-only DLA	68.00	0.8700	0.7700	79.80
Gemma	Random	51.00	0.7600	0.6800	70.30
Gemma	**C-DLA**	**89.00**	**0.9300**	**0.8000**	**85.60**
Llama	Wrong-only DLA	70.84	0.8576	0.7691	80.63
Llama	Correct-only DLA	65.39	0.8342	0.7386	78.74
Llama	Random	51.87	0.7264	0.6471	68.32
Llama	**C-DLA**	**85.26**	**0.8867**	**0.8005**	**83.00**
Qwen	Wrong-only DLA	72.41	0.8736	0.7582	81.17
Qwen	Correct-only DLA	67.83	0.8561	0.7437	79.64
Qwen	Random	52.46	0.7429	0.6583	69.41
Qwen	**C-DLA**	**86.74**	**0.9081**	**0.7932**	**82.39**

Table 5: Cross-model feature-attribution ablation using the top 100 features per transition zone. Purity and accuracy are reported as percentages; ID and OOD are measured by AUC.

Ablation Study on the Hallucination-related Sparse Feature Attribution Stage. To validate the C-DLA attribution mechanism, we compare it against two unidirectional rules and random feature selection under identical conditions (Top-100 features per model). As shown in Table 5, C-DLA achieves the highest ID AUC (0.93) and feature purity (89%) on Gemma, substantially outperforming Wrong-only DLA (0.88 AUC, 73% purity) and Correct-only DLA (0.87 AUC, 68% purity). The unidirectional variants miss either suppression of correct answers or enhancement of wrong ones, while C-DLA employs a contrastive vector $\Delta\mathbf{v} = \mathbf{W}_U[:, t_{\text{wrong}}] - \mathbf{W}_U[:, t_{\text{correct}}]$ to capture bidirectional causal pathways.

Fig. 5 further reveals that the top 0.1% of features (131/131,072) account for 41.1% of total attribution, and the top 1% (1,310) cover 62.5%, with a Gini coefficient of 0.912 versus 0.414 for random attribution—confirming hallucination is driven by a sparse subset of high-impact features. This validates our Top-100 selection, filtering out long-tail noise while targeting core mechanistic drivers. The contrastive advantage generalizes across SAE families, with HalluSAE exceeding the random baseline by 0.1244 AUC (12.44 points) on Gemma OOD.

Joint Ablation: Synergy Between Stage I and II. Our joint ablation (Appendix B, Table 5) confirms that neither stage alone is sufficient; the full system reaches 0.80 OOD AUC, demonstrating their complementary contributions.

6 Conclusion

We present HalluSAE, a novel framework that interprets hallucination in LLMs as a phase transition in latent dynamics. By tracking energy trajectories and focusing on sparse, causally significant features, HalluSAE enables interpretable and highly accurate hallucination detection. Our results on Gemma-2-9B, Llama-3.1-8B-Base, and Qwen3-8B-Base set a new standard for both in-distribution and out-of-distribution scenarios and show that the three-stage transition signature generalizes across independently developed model–SAE pairs. HalluSAE relies on the availability and quality of pre-trained SAEs, and its additional SAE projection cost may limit deployment in latency-sensitive or proprietary-model settings.

LLM USAGE STATEMENT

Generative AI tools were used to improve the readability of the manuscript. Generative AI tools were not used to design the research methodology or experiments, implement the proposed method, conduct qualitative or thematic analysis, or interpret experimental results. The remaining required-disclosure tasks were not applicable to this work. The authors verified the cited literature and supporting sources, tested AI-assisted code edits, checked all reported results against the original experiment records, and reviewed all AI-assisted text. The authors developed the method, designed and conducted the experiments, implemented the core approach, performed the analyses, and take full responsibility for the complete submission.

REPRODUCIBILITY STATEMENT

The main paper specifies the datasets, data splits, evaluation metrics, baseline methods, model–SAE configurations, feature-selection procedure, and probe-training protocol. Appendix C provides dataset construction, annotation, quality-control, and split details; Appendix A introduces the SAE and GPE formulations; Appendix D gives the model, SAE, computing, baseline, and detector-training configurations; Appendix E reports robustness analyses, statistical significance tests, and additional ablation studies; and Appendix F provides complete detection metrics and computational-cost comparisons.

REFERENCES

Samir Abdaljalil, Filippo Pallucchini, Andrea Seveso, Hasan Kurban, Fabio Mercorio, and Erchin Serpedin. SAFE: A sparse autoencoder-based framework for robust query enrichment and hallucination mitigation in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9335–9346. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.496. URL https://aclanthology.org/2025.findings-emnlp.496/.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations Workshop*, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying, 2023. URL https://arxiv.org/abs/2304.13734.

Kenza Benkirane, Laura Gongas, Shahar Pelles, Naomi Fuchs, Joshua Darmon, Pontus Stenetorp, David Ifeoluwa Adelani, and Eduardo Sánchez. Machine translation hallucination detection for low and high resource languages using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9647–9665. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.564. URL https://aclanthology.org/2024.findings-emnlp.564/.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, volume 2024, pp. 3056–3076, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0d1986a61e30e5fa408c81216a616e20-Abstract-Conference.html.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, 2024. doi: 10.1093/jla/laae003. URL https://doi.org/10.1093/jla/laae003.

Boyi Deng, Xu Wang, Yaoning Wang, Yu Wan, Yubo Ma, Baosong Yang, Haoran Wei, Jialong Tang, Huan Lin, Ruize Gao, Tianhao Li, Qian Cao, Xuancheng Ren, Xiaodong Deng, An Yang, Fei Huang, Dayiheng Liu, and Jingren Zhou. Qwen-Scope: Turning sparse features into development tools for large language models, 2026. URL https://arxiv.org/abs/2605.11887.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212/.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. HaloScope: Harnessing unlabeled LLM generations for hallucination detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 102948–102972. Curran Associates, Inc., 2024. doi: 10.52202/079017-3270. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ba92705991cfbbcedc26e27e833ebbae-Abstract-Conference.html.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Robert Friel and Atindriyo Sanyal. ChainPoll: A high efficacy method for LLM hallucination detection, 2023. URL https://arxiv.org/abs/2310.18344.

Ahmed Frikha, Muhammad Reza Ar Razi, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. PrivacyScalpel: Enhancing LLM privacy via interpretable feature intervention with sparse autoencoders. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 226–238. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.blackboxnlp-1.13. URL https://aclanthology.org/2025.blackboxnlp-1.13/.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/42ef3308c230942d223c411adf182c88-Abstract-Conference.html.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. Llama Scope: Extracting millions of features from Llama-3.1-8B with sparse autoencoders, 2024. URL https://arxiv.org/abs/2410.20526.

Junjie Hu, Gang Tu, ShengYu Cheng, Jinxin Li, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. HARP: Hallucination detection via reasoning subspace projection. In *International Conference on Learning Representations*, volume 2026, pp. 17783–17802, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1e58b1bf9f218fcd19e4539e982752a5-Abstract-Conference.html.

Zhenglin Hua, Jinghan He, Zijun Yao, Tianxu Han, Haiyun Guo, Yuheng Jia, and Junfeng Fang. Steering LVLMs via sparse autoencoder for hallucination mitigation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10808–10828. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.572. URL https://aclanthology.org/2025.findings-emnlp.572/.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025a. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

Yiming Huang, Junyan Zhang, Zihao Wang, Biquan Bie, Yunzhong Qiu, Xuming Hu, Yi R. Fung, and Xinlei He. RePPL: Recalibrating perplexity by uncertainty in semantic propagation and language generation for explainable QA hallucination detection, 2025b. URL https://arxiv.org/abs/2505.15386.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 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. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.emnlp-main.397. URL https://aclanthology.org/2023.emnlp-main.397/.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023b. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Yinghao Li, Rushi Qiang, Lama Moukheiber, and Chao Zhang. Language model uncertainty quantification with attention chain. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=QTrW2HWNXe.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Qiang Liu, Xinlong Chen, Yue Ding, Bowen Song, Weiqiang Wang, Shu Wu, and Liang Wang. Attention-guided self-reflection for zero-shot hallucination detection in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21005–21021. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1063. URL https://aclanthology.org/2025.emnlp-main.1063/.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Samuel Marks, Can Rager, Eric Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, volume 2025, pp. 23888–23923, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3ba4d47a83e498c2b1a0868cba20f6de-Abstract-Conference.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Valentin Noël. A graph signal processing framework for hallucination detection in large language models, 2025. URL https://arxiv.org/abs/2510.19117.

OpenMOSS-Team and Fudan NLP. Llama3.1-8B-Base LXR-32x sparse autoencoders. Hugging Face model card, 2024. URL https://huggingface.co/OpenMOSS-Team/Llama3_1-8B-Base-LXR-32x.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a712d461e57201efe35d429a6f1731c1-Abstract-Conference.html.

Sangha Park, Seungryong Yoo, Jisoo Mok, and Sungroh Yoon. SAVE: Sparse autoencoder-driven visual information enhancement for mitigating object hallucination. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7935–7944, 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Park_SAVE_Sparse_Autoencoder-Driven_Visual_Information_Enhancement_for_Mitigating_Object_Hallucination_WACV_2026_paper.html.

Qwen Team. Qwen-Scope residual-stream sparse autoencoders for Qwen3-8B-Base. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/SAE-Res-Qwen3-8B-Base-W64K-L0_100.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2. URL https://doi.org/10.1038/s41586-023-06291-2.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. LLM-Check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 34188–34216. Curran Associates, Inc., 2024. doi: 10.52202/079017-1077. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3c1e1fdf305195cd620c118aaa9717ad-Abstract-Conference.html.

Xiaohua Wang, Yuliang Yan, Longtao Huang, Xiaoqing Zheng, and Xuanjing Huang. Hallucination detection for generative large language models by bayesian sequential estimation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15361–15371. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.emnlp-main.949. URL https://aclanthology.org/2023.emnlp-main.949/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2206.07682.

Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2734–2744. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.236. URL https://aclanthology.org/2021.eacl-main.236/.

Guangzhi Xiong, Zhenghao He, Bohan Liu, Sanchit Sinha, and Aidong Zhang. Toward faithful retrieval-augmented generation with sparse autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 112549–112577, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b6f6bfbd260fbf2f5acb0a1d6439ca0e-Abstract-Conference.html.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564, 2023. doi: 10.1162/tacl_a_00563. URL https://aclanthology.org/2023.tacl-1.32/.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *International Conference on Learning Representations*, volume 2024, pp. 32097–32121, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/880ee6e23b6ee548ebcccafe0ef0f513-Abstract-Conference.html.

Luan Zhang, Dandan Song, Zhijing Wu, Yuhang Tian, Changzhi Zhou, Jing Xu, Ziyi Yang, and Shuhao Zhang. Detecting hallucination in large language models through deep internal representation analysis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 8357–8365, 2025a. doi: 10.24963/ijcai.2025/929. URL https://www.ijcai.org/proceedings/2025/929.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418, 2025b. doi: 10.1162/coli.a.16. URL https://aclanthology.org/2025.cl-4.9/.

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R. Fung, Kathleen McKeown, Chengxiang Zhai, Manling Li, and Heng Ji. The law of knowledge overshadowing: Towards understanding, predicting, and preventing LLM hallucination. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23340–23358. Association for Computational Linguistics, 2025c. doi: 10.18653/v1/2025.findings-acl.1199. URL https://aclanthology.org/2025.findings-acl.1199/.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023. URL https://arxiv.org/abs/2310.01405.

A SAE AND GPE BACKGROUND

Sparse Autoencoders (SAEs). SAEs mitigate feature superposition in dense activations by mapping the d_{model}-dimensional residual stream into a substantially higher-dimensional sparse space, where

Model	Strategy	Layer range	# Layers	# Features	ID AUC	OOD AUC	Δ OOD
Gemma	Random-13	Random 13	13	100	0.8300	0.7600	−4.00
Gemma	Early-13	L0–12	13	100	0.8100	0.7300	−7.00
Gemma	Late-13	L29–41	13	100	0.8800	0.7800	−2.00
Gemma	All	L0–41	42	420	0.9100	0.7900	−1.00
Gemma	**Transition**	**L23–35**	**13**	**100**	**0.9300**	**0.8000**	**–**
Llama	All	L0–31	32	–	0.8724	0.7863	−1.42
Llama	**Transition**	**L23–28**	**6**	**100**	**0.8867**	**0.8005**	**–**
Qwen	All	L0–35	36	–	0.8914	0.7811	−1.21
Qwen	**Transition**	**L16–23**	**8**	**100**	**0.9081**	**0.7932**	**–**

Table 4: Layer-selection ablation across model families. Δ OOD denotes the difference in OOD AUC, measured in percentage points, relative to the transition-zone detector for the corresponding model.

$d_{\text{SAE}} \gg d_{\text{model}}$ and each dimension is encouraged to represent a distinct semantic concept Huben et al. (2024); Bricken et al. (2023). We use the open-source Gemma Scope SAE Lieberum et al. (2024), which approximates the residual-stream activation at layer l and token position t as

$$r_l^t \approx \sum_{i=1}^{d_{\text{SAE}}} s_i W_{\text{dec}}[i,:], \tag{8}$$

where

$$s_i = \text{JumpReLU}\left(W_{\text{enc}} r_l^t + b_{\text{enc}}\right)[i] \tag{9}$$

is the activation of the ith sparse feature. We set $d_{\text{SAE}} = 131{,}072$, corresponding to a $36.6\times$ overcomplete representation, and use an average sparsity of $L_0 \approx 30$ to balance representational fidelity and interpretability. JumpReLU preserves activation magnitude while enforcing sparsity, enabling the SAE to retain graded semantic information that may be suppressed by conventional L1-regularized SAEs.

Geometric Potential Energy (GPE). From a dynamical-systems perspective, we quantify deviation from a stable factual state using geometric potential energy (GPE), defined as the squared Euclidean distance from a "truth attractor" in the SAE feature space Marks & Tegmark (2024); Zou et al. (2023):

$$E(l,t) = \left\|\text{SAE}(r_l^t) - \mu_{\text{truth}}\right\|_2^2, \tag{10}$$

where μ_{truth} is the centroid of SAE representations computed from factual samples in the labeled training split. This centroid is estimated once during offline training and cached together with the localized transition layers and selected features. No factual or hallucinated reference pool is required at inference time. GPE captures both the direction and magnitude of representational deviation, emphasizes large departures consistent with phase-transition-like behavior, and measures these departures in a semantically disentangled feature space.

B Complete Ablation Tables

The main text summarizes the experimental design and principal findings. Here, we report the complete ablation results.

Layer-selection ablation. Table 4 compares random, early, late, all-layer, and transition-zone selections. For Gemma, the transition-zone configuration achieves an ID AUC of 0.9300 and an OOD AUC of 0.8000 using 100 features, matching or exceeding all alternatives while using $4.2\times$ fewer features than the All-42 configuration. The same pattern is observed for Llama and Qwen, for which the model-specific transition zones outperform the corresponding all-layer configurations on both ID and OOD evaluation. These results indicate that GPE-based localization identifies the most informative layer interval rather than merely benefiting from a larger layer or feature budget.

To assess the combined contributions of the two stages, we conduct a joint ablation study. As shown in Table 5, using all layers with random feature selection yields an OOD AUC of 0.63. Applying

Configuration	Layers	# Features	AUC	
			ID	OOD
Baseline	All-42	420	0.63	0.63
Stage I only	L23–35	130	0.71	0.64
Stage II only	All-42	420	0.91	0.79
Full model (I+II)	**L23–35**	**100**	**0.93**	**0.80**

Table 5: Joint ablation of Stage I transition-zone localization and Stage II C-DLA feature attribution.

Derivative	Source and use	Added artifacts
HaluEval	HaluEval QA; ID training and testing	Model outputs, binary labels, error types, review decisions, and fixed splits
TriviaQA	TriviaQA; OOD testing only	Model outputs, binary labels, error types, and review decisions

Table 6: Provenance and added content of the curated dataset derivatives. The citations and licenses associated with the original source datasets remain applicable.

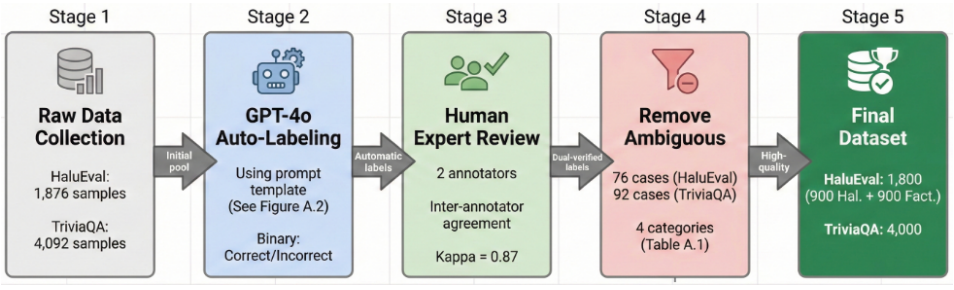


Figure 6: Dataset construction and quality-control pipeline. The five-stage workflow proceeds from raw-data collection to automatic annotation, expert review, ambiguity filtering, and construction of the curated datasets.

Stage I localization while retaining random features increases the OOD AUC to 0.64, whereas applying Stage II C-DLA attribution across all layers increases it to 0.79. Combining transition-zone localization with C-DLA attribution achieves the best OOD AUC of 0.80. Thus, feature attribution provides the largest individual gain, while transition-zone localization yields an additional complementary improvement.

C Data Appendix

Dataset Status and Provenance

Rather than introducing a new source benchmark, we construct curated and re-annotated derivatives of HaluEval Li et al. (2023a) and TriviaQA Joshi et al. (2017). Each record preserves its source provenance and is augmented with model generations, hallucination labels, error categories, quality-control decisions, and fixed data splits. HaluEval is used for in-distribution training and evaluation, whereas TriviaQA is reserved exclusively for out-of-distribution evaluation.

Dataset Construction and Quality Control

Figure 6 summarizes the quality-control pipeline, comprising raw-data collection, GPT-4o annotation, expert review, removal of ambiguous samples, and construction of the final datasets.

Dataset statistics and quality control. Table 7 summarizes the two curated datasets. From initial pools of 1,876 HaluEval records and 4,092 TriviaQA records, we remove 76 and 92 ambiguous cases, respectively, yielding final datasets of 1,800 and 4,000 samples. The removed HaluEval cases

Attribute	HaluEval	TriviaQA
Total samples	1,800	4,000
Hallucinated / factual	900 / 900	2,000 / 2,000
Mean question length	8.5 words	12.3 words
Mean answer length	4.9 tokens	1.5 words
Initial pool size	1,876	4,092
Removed ambiguous cases	76 (4.0%)	92 (2.2%)
Ambiguous HaluEval samples		
Missing units	23 (30.3%)	
Granularity mismatch	19 (25.0%)	
Partially correct	18 (23.7%)	
Annotation conflict	16 (21.0%)	
TriviaQA domain distribution		
Geography / history	28% / 24%	
People / organizations	23% / 25%	
Usage	Training and ID testing	OOD testing only
Train / test split	80% / 20%	–

Table 7: Statistics and quality-control outcomes for the curated HaluEval and TriviaQA derivatives.

```
You are an expert fact-checker
for large language models. Verify
whether the [Model Response]
contains factual hallucinations
based strictly on the [Reference
Knowledge] and [Ground Truth].

Input Data:
[Reference Knowledge]: {knowledge}
[Question]: {question}
[Ground Truth]: {ground_truth}
[Model Response]: {model_response}

Evaluation Criteria:
1. Entity Verification: Check
whether core entities, including
names, places, and dates, match.
2. Numerical Precision: Allow
minor formatting differences,
such as ``20%'' versus ``20
percent,'' but mark substantive
numerical deviations as
incorrect.
3. Contradiction Check: If the
response contradicts the
reference knowledge, mark
INCORRECT.
4. Relevance: If the response is
irrelevant or incomplete, mark
INCORRECT.

Step-by-Step Reasoning:
1. Identify the key claim in
[Ground Truth].
2. Extract the corresponding
claim from [Model Response].
3. Compare the claims and
explicitly identify any
discrepancies.
4. Determine the final verdict.

Output Format (JSON):
{
  "reasoning": "Concise
explanation identifying specific
errors, if any.",
  "label": "CORRECT" or
"INCORRECT"
}
```

Figure 7: GPT-4o annotation prompt. The template provides explicit grounding, numerical-tolerance rules, structured reasoning instructions, and a machine-readable JSON output format.

comprise missing units (30.3%; for example, "330" without the unit "meters"), granularity mismatches (25.0%; for example, "China" instead of "Beijing"), partially correct answers (23.7%), and annotation conflicts (21.0%). HaluEval is partitioned into 1,440 training records (80%) and 360 held-out test records (20%). TriviaQA is used exclusively for OOD evaluation.

Group	Fields
Provenance	`sample_id`, `source_id`, `source_dataset`, and `source_split`
Content	`question`, `ground_truth`, and `model_response`
Generation	`model_name`, `generation_id`, `generation_seed`, and `generation_parameters`
Annotation	`automatic_annotation`, `human_review`, `final_label`, `annotation_conflict`, and `error_type`
Split and QC	`paper_split` and `removal_reason`

Table 8: Schema of the curated HaluEval and TriviaQA derivative records.

GPT-4o Annotation Protocol and Design Principles

For reproducibility, Figure 7 provides the complete GPT-4o annotation prompt. The prompt incorporates four design principles: explicit grounding to reduce judge hallucination, numerical tolerance to accommodate inconsequential formatting differences, structured reasoning to support consistent adjudication, and JSON output to enable automated parsing. Every sample undergoes two-stage annotation: automatic labeling by GPT-4o followed by mandatory expert review. When the automatic and expert judgments disagree, the expert judgment determines the final label.

Record Schema and Fixed Splits

Each released record is uniquely identified by the tuple `(source_id, model_name, generation_id)`, preventing multiple generations of the same source question from being conflated with distinct source examples. Table 8 defines the released schema. The HaluEval split is fixed and stratified by the final binary label: 1,440 records are used for training and model selection, and 360 disjoint records are reserved for held-out ID testing. TriviaQA records are never used to estimate centroids, localize transition zones, rank features, tune hyperparameters, or select decision thresholds; they are used only for final OOD evaluation.

Availability, Licensing, and Limitations

Upon publication, we will release the curated splits, model outputs, annotations, and documentation required to reproduce the reported data pipeline. The release package will contain two curated JSONL files, a split manifest, the annotation prompt, expert-review guidelines, a dataset card, third-party license notices, and SHA-256 checksums. The attribution and licensing terms of the original HaluEval and TriviaQA datasets will remain applicable to redistributed source fields. Newly added annotations and metadata will be released under a permissive license. Our pipeline does not intentionally add personally identifying information.

The labels target claims that conflict with, or are unsupported by, the supplied reference knowledge and ground truth. They should not be interpreted as a comprehensive taxonomy of all possible generation failures. Time-varying claims and samples that could not be adjudicated consistently were excluded. Both GPT-4o judgments and expert adjudication may nevertheless contain annotator-specific errors. The release therefore preserves automatic labels, expert-review decisions, conflicts, and removal reasons rather than providing only the final binary labels.

D Reproducibility Details

Model, SAE, and Computing Configuration

We evaluate three independently developed model–SAE pairs: Gemma-2-9B with Gemma Scope Lieberum et al. (2024), Llama-3.1-8B-Base with Llama Scope/LXR-32x He et al. (2024); OpenMOSS-Team and Fudan NLP (2024), and Qwen3-8B-Base with Qwen-Scope Deng et al. (2026); Qwen Team (2026). Their localized transition zones are L23–35, L23–28, and L16–23, respectively. Gemma and Llama use 131,072-dimensional JumpReLU sparse representations, whereas Qwen uses a 65,536-dimensional TopK SAE. Stage II selects the 100 highest-ranked C-DLA features within each model's transition zone. Factual centroids, transition zones, feature rankings, fea-

Component	Setting
GPU	8× NVIDIA RTX 4090
CPU	AMD EPYC 7T83, 64 cores / 128 threads
RAM / operating system	503 GiB / Ubuntu 22.04.5 LTS
Python / PyTorch	3.10.18 / 2.9.0+cu128
CUDA build / cuDNN	12.8 / 9.10.2
NVIDIA driver	560.28.03
Transformers / SAE Lens	4.52.4 / 6.22.2
scikit-learn / NumPy / SciPy	1.7.2 / 1.26.4 / 1.15.3
Datasets / Accelerate	3.6.0 / 1.7.0

Table 9: Shared computing environment for the Gemma, Llama, and Qwen experiments. The CUDA version of the installed PyTorch build is reported separately from the NVIDIA driver version.

Layers	Mean L_0 range	Representative values
L0–9	25–42	L0 (30), L8 (41), L9 (42)
L10–19	27–35	L16 (35), L17 (35), L18 (34)
L20–29	30–34	L20 (34), L23 (32), L29 (33)
L30–41	26–32	L30–39 (30), L40 (29), L41 (26)
L23–35	**31**	**Approximately the global mean**

Table 10: Layer-wise sparsity configuration of the Gemma SAE. L_0 denotes the mean number of active features.

ture scalers, and probe parameters are fitted using training or validation data only and are subsequently cached for evaluation.

Generation and SAE feature extraction use `bfloat16`. Downstream probe fitting and metric computation use the default numerical precision of the listed scikit-learn and NumPy environment. Greedy decoding is used for the primary generations. Batch sizes range from 4 to 16 depending on sequence length, and peak GPU memory consumption remains below 40 GB per GPU.

Randomness and determinism. Unless otherwise specified in Table 13, all experiments use seed 42. We seed Python, NumPy, PyTorch CPU operations, and all CUDA devices. We also enable deterministic cuDNN execution and disable cuDNN benchmarking. Global deterministic-algorithm enforcement and the cuBLAS workspace configuration are not enabled. These settings reduce stochastic variation but do not guarantee bitwise-identical CUDA execution across hardware or software environments.

Evaluation Metrics and Selection Criteria

We treat hallucination as the positive class and denote true positives, true negatives, false positives, and false negatives by TP, TN, FP, and FN, respectively. Accuracy and recall are defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{12}$$

For detector score s, ROC AUC can be interpreted as the probability that a randomly drawn hallucinated example receives a higher score than a randomly drawn factual example:

$$\text{AUC} = \Pr\left[s(x^+) > s(x^-)\right], \tag{13}$$

with half credit assigned to tied scores. AUC is the primary metric because it evaluates ranking quality independently of a deployment-specific decision threshold. Accuracy summarizes overall classification performance on the balanced test sets, whereas recall measures the proportion of hallucinations detected. Hyperparameters are selected according to mean validation AUC, without access to test data.

Method	Samples	Key configuration
LN-Entropy	1	Final-logit entropy; threshold range 0.1–0.9
Semantic Entropy	10	MiniLM-L6; nucleus sampling $p = 0.9$, $T = 0.8$; clustering threshold 0.75
Lexical Similarity	5	$T = 0.7$; ROUGE-L; threshold range 0.3–0.8
SelfCheckGPT	5	DeBERTa-XL; nucleus sampling $p = 0.95$, $T = 1.0$; BERTScore F1 > 0.5
EigenScore	1	L35; five largest eigenvalues; threshold range 0.5–3.0
MHAD	1	L20/25/30/35; 16 heads; threshold range 0.1–0.5
SAPLMA	1	L35 residual stream; L2 regularization; $C = 1.0$; five-fold CV
HaloScope	1	L30–40; MLP ($21{,}504 \to 512 \to 2$); learning rate 10^{-3}
HalluSAE	**1**	**L23–35; C-DLA; L1 logistic regression; top 100 features**

Table 11: Hyperparameter configurations of HalluSAE and the baseline methods. Supervised methods use the 1,440-record HaluEval training split.

Baseline Implementation and Hyperparameters

All baselines are evaluated under matched experimental conditions, including the same hardware, preprocessing, metrics, random seed, and HaluEval training split for supervised methods. Table 11 summarizes the configurations of the eight baseline methods. Hyperparameters are selected by five-fold cross-validation using validation AUC. Threshold-based methods—LN-Entropy, Semantic Entropy, Lexical Similarity, SelfCheckGPT, EigenScore, and MHAD—select their decision thresholds on the validation data.

Implementation details. Uncertainty-based methods. LN-Entropy computes log-normalized entropy from the final-token logits. Semantic Entropy clusters 10 nucleus-sampled outputs using Sentence-BERT embeddings and agglomerative clustering.

Consistency-based methods. Lexical Similarity computes pairwise ROUGE-L F1 scores across five temperature-sampled outputs. SelfCheckGPT uses five nucleus-sampled outputs with BERTScore-based verification.

Internal-state methods. EigenScore analyzes the eigenspectrum of Layer 35 attention matrices. MHAD computes Jensen–Shannon divergence across 16 attention heads in four selected layers.

Supervised methods. SAPLMA trains a linear probe on the 3,584-dimensional Layer 35 residual stream. HaloScope applies a two-layer MLP to 21,504 concatenated features extracted from six layers. HalluSAE uses 100 selected features, corresponding to a $215\times$ reduction relative to HaloScope's 21,504-dimensional input, and evaluates 13 rather than 42 layers, corresponding to a $4.2\times$ reduction relative to full-layer extraction.

Hallucination Detector Training

The final detector is an L1-regularized logistic regression model trained on the 100 highest-ranked C-DLA features. We use the `liblinear` solver, a maximum of 1,000 iterations, a convergence tolerance of 10^{-4}, and `class_weight=balanced`. Within each cross-validation fold, a `StandardScaler` is fitted only on the training partition and then applied to the corresponding validation partition. After hyperparameter selection, the scaler and detector are refitted using all 1,440 HaluEval training records.

The regularization strength C is selected by stratified five-fold cross-validation. The search space comprises 20 logarithmically spaced values from 10^{-4} to 10^{4}, and the selection criterion is mean validation AUC. The selected value is $C = 0.1$. Each fold contains 1,152 training records and 288 validation records. The 360-record HaluEval test set and all 4,000 TriviaQA records remain untouched until final evaluation.

Component	Configuration
Model / regularizer	Logistic regression / L1
Input / output	Top 100 C-DLA features / binary label
Solver / maximum iterations	`liblinear` / 1,000
Tolerance / class weight	10^{-4} / balanced
Feature scaling	`StandardScaler`, training data only
C search space	$[10^{-4}, 10^{4}]$; 20 log-spaced values
Selection criterion	Stratified five-fold CV; mean AUC
Selected C	0.1
Final training / ID test	1,440 / 360 records
OOD test	4,000 TriviaQA records

Table 12: Configuration of the final hallucination detector and model-selection procedure.

Experiment	Runs	Seed, procedure, and reporting
Final HalluSAE, per model	1	Seed 42; five-fold CV selects C; held-out ID and OOD metrics
Baseline, per method and model	1	Seed 42; CV or validation-threshold selection; held-out ID and OOD metrics
Ablation, per setting and model	1	Seed 42; single final evaluation
Random-feature control	10	Seeds 42–51; mean and standard deviation
Transition-zone bootstrap	1	Seed 42; 1,000 resamples; distributional summaries
Permutation control	1	Seed 42; 100 permutations; empirical false-positive rate

Table 13: Experimental repetitions, random seeds, internal resampling procedures, and aggregation protocols.

Runs, Seeds, and Aggregation

Table 13 distinguishes independent experimental runs from cross-validation folds and statistical resamples. Each baseline and deterministic ablation configuration is executed once using seed 42. The five folds used for hyperparameter selection are not treated as independent experimental repetitions. The random-feature control is the only detector comparison averaged across multiple seeds.

E Robustness and Ablation Studies

Robustness of Transition-Zone Localization

We evaluate the stability of the localized Gemma transition zone, L23–35, using two complementary analyses: parameter sensitivity and bootstrap stability.

Parameter sensitivity. We systematically vary the localization criteria to determine whether the identified zone is sensitive to the choice of hyperparameters. Across 42 configurations, the start multiplier ranges from 1.0 to 2.0, and the endpoint peak tolerance ranges from 5% to 30%. The default configuration uses a start multiplier of 1.3 and an endpoint tolerance of 10%.

Table 14 reports six representative configurations. Thirty of the 42 configurations (71.4%) recover the reference zone exactly, corresponding to an intersection over union (IoU) of 1.0. Across all configurations, the mean IoU is 0.956, and the minimum is 0.846. The estimated start boundary is 23.6 ± 0.9, whereas the endpoint remains fixed at L35 across the evaluated configurations.

Bootstrap stability. As a post hoc stability analysis that is not used to select or train the detector, we perform 1,000 bootstrap iterations. Each iteration samples 1,440 records, corresponding to 80% of the 1,800-record analysis pool, with replacement and reruns the localization algorithm using the default criterion. The resulting mean IoU is 0.925 ± 0.110, and the median IoU is 1.000. A total of 681 iterations (68.1%) achieve IoU ≥ 0.8, and the minimum observed IoU is 0.647.

ID	Start	Tol. (%)	IoU	Zone	Length
1	1.0	5	1.000	L23–35	13
7	1.1	5	1.000	L23–35	13
19 (default)	1.3	10	1.000	L23–35	13
31	1.5	25	1.000	L23–35	13
37	1.8	5	0.846	L25–35	11
42	2.0	30	0.846	L25–35	11
Exact recovery: 30/42; mean IoU: 0.956; minimum IoU: 0.846.					

Table 14: Sensitivity of transition-zone localization to the start multiplier and endpoint tolerance.

Layers	Selections	Stability	Target	Category
L0–22	0	0.000	No	–
L23	998	0.998	Yes	Boundary
L24	999	0.999	Yes	Boundary
L25–35	1000	1.000	Yes	Stable core
L36–39	319	0.319	No	Extended boundary
L40–41	0	0.000	No	–

Table 15: Layer-wise bootstrap stability of the localized transition zone over 1,000 resamples.

Comparison	d	t	df	p	Interpretation
Phase II vs. I	1.64	18.32	1798	< 0.001	Very large
Phase III vs. II	0.12	0.95	1798	0.342	No effect
Phase III vs. I	1.71	19.87	1798	< 0.001	Very large
Phase I: L0–22; Phase II: L23–35; Phase III: L36–41.					

Table 16: Statistical comparisons of GPE differences across the stable, transition, and plateau phases.

The start boundary is 23.0 ± 0.1 layers, with mode L23 and range L23–25. The endpoint is 36.3 ± 1.9 layers, with mode L35 and range L35–39. Table 15 reports layer-wise inclusion frequencies. All 13 layers in the reference transition zone have stability of at least 0.998, and the 11 central layers, L25–35, are selected in every bootstrap iteration.

Statistical significance of phase differences. We quantify the magnitude of adjacent phase differences using Cohen's d:

$$d = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}}, \tag{14}$$

where μ and σ denote the mean and standard deviation of GPE differences within each phase. Table 16 reports the complete results. The transition phase differs strongly from the stable phase ($d = 1.64$, $p < 0.001$), whereas the plateau phase does not differ significantly from the transition phase ($d = 0.12$, $p = 0.342$). These findings support a sharp transition followed by a sustained high-energy plateau rather than a second qualitative shift at the transition–plateau boundary.

Permutation-based false-positive control. We randomly permute the factual and hallucination labels 100 times and rerun the localization procedure using an augmented criterion that requires every candidate layer to pass a layer-wise significance test at $p < 0.01$. None of the 100 permutations produces a detected transition, corresponding to an empirical false-positive rate of 0%. By contrast, the transition layers identified from the unpermuted data pass the significance criterion. This analysis distinguishes systematic energy growth from chance sequences of positive layer-wise differences.

Robustness to centroid estimation and label noise. The factual centroid used by the reported detector is computed exclusively from the labeled training split. As robustness checks, we replace the HaluEval-only factual reference pool with an equally weighted mixture of factual HaluEval and TriviaQA samples and replace the coordinate-wise mean with either a 10% trimmed mean or a coordinate-wise median. Each alternative preserves the qualitative transition from a low-energy to a high-energy regime.

Method	Purity	ID AUC	OOD AUC	ID Acc.	OOD Acc.
Wrong-only DLA	73	0.88	0.79	81.2	74.5
Correct-only DLA	68	0.87	0.77	79.8	72.1
Random features	51	0.76	0.68	70.3	65.2
C-DLA	**89**	**0.93**	**0.80**	**85.6**	**78.3**

Table 17: Detailed feature-attribution comparison for Gemma. Purity and accuracy are percentages. All methods use transition layers L23–35 and the same 100-feature budget.

We additionally inject hallucinated samples into the factual reference pool. The mean GPE across the evaluated high-energy layers decreases smoothly from 59,958 with no contamination to 54,407, 48,797, and 36,959 with 5%, 10%, and 20% contamination, respectively. We interpret this result as evidence that moderate contamination attenuates the energy contrast without eliminating the transition trend; we do not require the estimated boundaries to remain numerically identical.

For the two additional architectures, corruption of the training labels produces gradual rather than catastrophic degradation. Under 10% and 20% label noise, Llama achieves ID AUC values of 86.18 and 82.73, respectively, whereas Qwen achieves 88.31 and 84.92. All Qwen results use L16–23 for localization and probe-feature extraction.

Feature-Attribution Comparisons and Validation

C-DLA versus unidirectional attribution. To evaluate the contribution of contrastive attribution, we compare Contrastive Direct Logit Attribution (C-DLA) with two unidirectional baselines under matched conditions, using the same transition layers and a 100-feature budget. *Wrong-only DLA* measures only the contribution to the incorrect token, $\mathcal{A}_i^{(t_{\text{wrong}})}$, whereas *Correct-only DLA* measures only suppression of the correct token, $-\mathcal{A}_i^{(t_{\text{correct}})}$.

The cross-model attribution table in the main text reports the complete comparison. For Gemma, C-DLA increases feature purity—defined as the proportion of hallucinated samples among the top 10% of samples ranked by activation—from 68–73% for the unidirectional methods to 89%. Paired tests using the five cross-validation folds indicate significant differences between C-DLA and Wrong-only DLA ($t = 8.32$, $p < 0.001$) and between C-DLA and Correct-only DLA ($t = 9.14$, $p < 0.001$). These results suggest that detecting hallucinations benefits from modeling the relative logit advantage of incorrect over correct outputs rather than either direction independently.

Random-feature baseline. To determine whether C-DLA identifies informative features beyond those obtained by chance, we compare it with a random-feature baseline. We randomly select 100 features from L23–35 using 10 seeds while restricting the sampling pool to features with comparable activation strength, defined as a mean activation of at least 5.0.

The 100-feature random baseline achieves an OOD AUC of 0.6800, whereas the final 100-feature C-DLA detector achieves 0.8044. C-DLA therefore improves OOD AUC by 0.1244, or 12.44 AUC points, over an equally sized random feature set. We do not report a z-score because the variance estimated for a different random-feature budget is not applicable to this 100-feature comparison.

Exploratory bidirectional intervention. This analysis is intentionally separated from the 100-feature detection protocol. For selected Gemma examples, we intervene on the five highest-ranked toxic features identified by C-DLA within the transition layers. In the ablation direction, setting these feature activations to zero changes the hallucinated answer "18,000" for the population of the largest town on Kaua'i to the reference answer "10,699". In the steering direction, amplifying features from the same class changes an initially correct response, "Waka Flocka Flame," to the incorrect response, "Gucci Mane." These examples provide qualitative tests of necessity and sufficiency but are not used to train, tune, or evaluate the detector.

Phase	Layers	Pearson r	p-value
Stable	L0–22	0.341	0.112
Transition	**L23–35**	**0.990**	< 0.001
Plateau	L36–41	0.287	0.581
All layers	L0–41	0.876	< 0.001

Table 18: Phase-specific correlations between cumulative C-DLA attribution and the hallucination–factual GPE difference.

JOINT ABLATION OF STAGE I AND STAGE II

We use a 2×2 factorial ablation to assess whether Stage I GPE-based transition-zone localization and Stage II C-DLA feature attribution provide complementary or redundant contributions. Table 5 reports the complete results.

Marginal effects. Relative to the rounded baseline OOD AUC of 0.63, Stage I alone improves OOD AUC by 1 point, whereas Stage II alone improves it by 16 points. Combining both stages yields a 17-point improvement. These directly reproducible differences indicate that C-DLA feature attribution provides the dominant gain, while transition-zone localization contributes an additional improvement. We do not infer a higher-order interaction beyond the precision supported by the reported values.

MICROSCOPIC–MACROSCOPIC FEATURE–ENERGY CORRESPONDENCE

Global correspondence. To evaluate whether microscopic C-DLA feature attribution and macroscopic GPE dynamics capture a shared transition mechanism, we compute the Pearson correlation between layer-wise cumulative C-DLA scores and the GPE difference between hallucinated and factual examples across all 42 layers. The global correlation is strong ($r = 0.876$, $p < 0.001$), and the transition-zone correlation is stronger still.

We also quantify the concentration of feature contributions using the Gini coefficient:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \bar{x}}, \tag{15}$$

where x_i is the C-DLA score of the ith feature. The observed coefficient, $G_{\text{obs}} = 0.912$, is substantially larger than the random-permutation baseline, $G_{\text{random}} = 0.414$, computed from 1,000 permutations. This result indicates that attribution is concentrated in a small subset of features rather than distributed uniformly across the SAE representation.

Phase-specific correlation. To determine whether the association between C-DLA and GPE is distributed uniformly across the network or concentrated in the transition zone, we compute Pearson correlations separately for the stable, transition, and plateau phases. As shown in Table 18, the strongest correlation occurs in the transition zone (L23–35; $r = 0.990$, $p < 0.001$). The correlations are weaker and not statistically significant in the stable phase (L0–22; $r = 0.341$, $p = 0.112$) and plateau phase (L36–41; $r = 0.287$, $p = 0.581$).

We further compare the transition-zone correlation with the correlation over the combined stable and plateau phases using Fisher's Z transformation:

$$Z = \frac{Z_{r_1} - Z_{r_2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}, \qquad Z_r = \frac{1}{2} \ln \left(\frac{1 + r}{1 - r} \right). \tag{16}$$

Comparing the transition zone ($r = 0.990$, $n = 13$) with the combined stable and plateau phases ($r = 0.314$, $n = 29$) yields $Z = 6.74$ and $p < 0.001$. Thus, the transition-zone correlation is statistically distinguishable from that observed in the remainder of the network.

This phase-specific correspondence supports two conclusions. First, C-DLA identifies sparse features whose layer-wise behavior is closely aligned with the macroscopic increase in GPE. Second, this alignment is localized to the same layers in which the representational transition occurs, rather

Model	Split	AUC	Accuracy	Recall
Gemma-2-9B	ID	92.86	85.60	80.88
	OOD	80.44	78.30	75.60
Llama-3.1-8B	ID	88.67	83.00	86.00
	OOD	80.05	72.85	73.25
Qwen3-8B	ID	90.81	82.39	83.74
	OOD	79.32	73.36	74.82

Table 19: Complete HalluSAE detection results across model families. All values are percentages.

Method	ID			OOD		
	AUC	Acc.	Rec.	AUC	Acc.	Rec.
LN-Entropy	54.93	56.50	65.12	70.94	65.75	66.50
Semantic Entropy	62.77	67.98	55.45	71.85	69.21	71.88
Lexical Similarity	64.25	62.50	59.00	65.40	64.25	53.50
SelfCheckGPT	63.78	64.12	56.23	63.53	61.51	57.12
EigenScore	60.53	59.50	77.13	61.51	60.25	66.50
MHAD	64.18	63.32	66.20	61.01	60.75	76.51
SAPLMA	79.14	75.34	75.19	64.33	59.50	55.12
HaloScope	84.37	78.62	75.96	78.94	72.21	72.83
HalluSAE	**92.86**	**85.60**	**80.88**	**80.44**	**78.30**	**75.60**

Table 20: Complete Gemma-2-9B detection results on HaluEval (ID) and TriviaQA (OOD). All values are percentages; AUC is the primary ranking metric.

than reflecting a uniform correlation across the network. The intervention results provide additional qualitative evidence that these features are functionally relevant, although the correlational analysis alone does not establish causality.

F Additional Results and Efficiency

Complete Detection Metrics

Computational Cost and Efficiency

Inference-time cost. Table 21 compares the computational cost of HalluSAE with representative baselines. The $1.0\times$ reference corresponds to one complete standard autoregressive generation without an auxiliary detector, rather than the computation of a single transformer layer. HalluSAE performs SAE projection and probe evaluation during the same generation pass. Its $2.3\times$ cost therefore represents end-to-end latency relative to standard generation, whereas "before completion" refers to the point during generation at which the warning becomes available.

In wall-clock measurements on the benchmark system, standard generation requires 0.65 s per example, whereas HalluSAE requires 1.47 s. SelfCheckGPT with five sampled generations requires 3.78 s ($5.8\times$), and Semantic Entropy with five sampled generations requires 3.99 s ($6.1\times$). HalluSAE can therefore emit a warning from cached transition-layer features during the original generation, before response completion, although the SAE projections increase total generation latency.

Cost decomposition. The measured $2.3\times$ inference-time cost comprises

- standard autoregressive generation: $1.0\times$;
- SAE feature extraction over 13 layers (L23–35): $0.8\times$; and
- prediction using the 100-dimensional linear probe: $0.5\times$.

Thus, the total measured cost is $2.3\times$ that of standard generation. Compared with consistency-based methods such as SelfCheckGPT, which require multiple generations, HalluSAE provides a favorable

Method	Generation passes	Relative cost
Standard generation	1	1.0×
Perplexity / entropy	1	1.1×
HalluSAE (L23–35)	**1**	2.3×
Full-layer probing (L0–41)	1	7.2×
SelfCheckGPT (5 samples)	5	5.8×
SelfCheckGPT (10 samples)	10	10.0×

Table 21: Inference-time computational cost of HalluSAE and representative baselines. Relative cost is normalized to one standard autoregressive generation.

accuracy–efficiency trade-off. Relative to full-layer probing at 7.2× cost, transition-zone localization reduces the measured computational cost by approximately 68% while maintaining comparable or superior detection performance.

One-time training cost. The three-stage training pipeline incurs the following one-time computational costs:

- **Stage I—transition-zone localization:** approximately 3 GPU-hours;
- **Stage II—C-DLA computation and feature selection:** approximately 5 GPU-hours;
- **Stage III—cross-validated probe training:** less than 1 GPU-hour; and
- **Total:** approximately 8 GPU-hours on the eight-GPU system.

These costs are incurred only during training. At inference time, the fitted detector is applied directly to new samples without recomputing the GPE analysis or C-DLA rankings. Inference therefore requires only the original generation pass, SAE feature extraction within the cached transition zone, and evaluation of the trained linear probe.