HALLUCINATIONS EXHIBIT SHALLOWER ENTROPY VALLEYS: TRAINING-FREE HALLUCINATION DETECTION IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models often give answers that sound confident and coherent even when the answers are unsupported or false. This gap between fluent language and reliable content makes it difficult to know when a generated response should be trusted. Existing hallucination detectors usually inspect the completed text, compare it with external evidence, or generate multiple answers to measure disagreement. These approaches can be useful, but they also raise a sharper question: can hallucinations be detected from the model's own internal computation without additional models, retrieval, or repeated decoding? We study this question through large activation events: the small subset of neuron activations whose absolute values are unusually large at a given layer. Rather than treating these events only as a sparsity pattern, we track how often each coordinate emits them across a response and summarize their layer-wise profile with binary entropy. We smooth this profile, retain non-oscillatory local minima, and use the descent prominence of the dominant entropy valley as a training-free hallucination-risk signal without imposing a fixed layer partition. We observe that these events preserve much of the computation needed for reasoning under causal masking, while controls at the same sparsity do not. Empirically, on paired TruthfulQA answers, correct responses exhibit a clearer decrease-and-recovery entropy pattern across layers, and the resulting event-profile score achieves AUROC values of 0.82 for LLaMA-3-8B and 0.79 for LLaMA-3-70B. Valley-stage interventions localize the strongest causal sensitivity to the low-entropy core, while matched-sparsity threshold comparisons support the robustness of the event definition. These results suggest that internal event profiles across layers can serve as an interpretable diagnostic from a single forward pass for identifying responses that warrant further verification.

1 INTRODUCTION

Fluent language is not evidence that a generated answer is correct. Large Language Models (LLMs), which are built on the Transformer architecture and scaled to broad language tasks (Han et al., 2025; Zhang et al., 2025b), can produce statements that are grammatically polished, locally coherent, and nevertheless unsupported by the prompt or by verifiable facts (Zhang et al., 2025b; Janiak et al., 2025). This failure mode, commonly called hallucination, is especially consequential in settings where an unsupported claim can trigger an incorrect decision, action, or allocation of resources.

Most hallucination detectors inspect the output after it has been generated. They compare a response with retrieved evidence, sample several answers to measure disagreement, or use a separate evaluator to judge factual consistency (Han et al., 2025; Janiak et al., 2025; Zhang et al., 2025b). These approaches are useful, but they can be expensive, require external evidence, or depend on another language model. More fundamentally, they assess artifacts of generation rather than the internal computation that produced the response. As a result, they may show that a response is unreliable without revealing where its uncertainty emerged in the computation. Identifying that location matters because it can distinguish a transient ambiguity from a persistent failure in the formation of the response, while a signal available during generation could support verification or abstention before an unsupported claim is finalized. This motivates a question: can the generation process itself provide a signal that separates supported responses from hallucinated ones?

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.

Investigating this possibility requires looking inside the model as it generates. Recent work has shown that hidden representations and their evolution across depth can encode factuality-related information (Han et al., 2025; Zhang et al., 2025a;c). However, a detector based on a single hidden vector is difficult to interpret and may be sensitive to the selected layer. A decoder-only Transformer instead constructs a response through a sequence of layers, making the evolution of its internal states across depth a structured object for locating changes in the model's computation.

Transformer layers repeatedly transform and pass forward the residual representation, and unusually large coordinates can mark components making especially salient contributions at a given stage of this computation. Our causal masking results test this functional role directly. During correct reasoning, these events become organized into a stable pattern over a range of layers as information is progressively integrated, whereas hallucinated responses are less likely to exhibit a clear stable region. To capture this contrast, we measure how consistently each coordinate produces an event across answer tokens and summarize this consistency with binary entropy at each layer. A coherent low-entropy valley in the resulting depth profile thus provides a training-free signal for hallucination detection without requiring a fixed layer partition.

We first use matched-sparsity causal masking to test whether retaining large activation events preserves answer-generation accuracy, and whether masking equally sparse random or small-magnitude coordinates causes greater damage. We then compare layer-wise event profiles to identify the cross-layer pattern associated with reliable generation. Finally, we convert that pattern into a response-level score and test whether it distinguishes hallucinated from supported answers. Together, these experiments establish the functional relevance of the events, characterize their reliability-related dynamics, and evaluate their use for hallucination detection.

The main contributions are:

- **Internal dynamics underlying hallucination detection.** We characterize the changes in internal activations across layers when an LLM generates a correct answer and use these characteristics as the basis for hallucination detection.
- **A causal test of event importance.** We use online activation masking to test whether the events used by the detector are more informative for reasoning than random coordinates or coordinates with small magnitudes at the same sparsity.
- **An empirical evidence chain across model scales and network depths.** Causal masking, entropy-profile comparisons, cross-dataset detection, valley-stage interventions, and threshold analyses connect the event definition to the final score in LLaMA-3-8B and LLaMA-3-70B.

2 Related Work

Hallucination detection from model outputs. Many hallucination detectors inspect a completed response rather than its generating computation. Black-box methods compare sampled answers (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024), requiring repeated decoding. Evidence-based evaluation measures factual support at the claim or response level (Min et al., 2023), but needs external evidence and may introduce a separate evaluator. More recent work has explored lightweight single-sample factuality probes and has shown that evaluation choices can substantially change the apparent performance of a detector (Han et al., 2025; Janiak et al., 2025); conclusions can depend as much on evaluation design as on the detector itself. These limitations motivate detectors using the computation from a single generation pass.

Internal representations for hallucination detection. Another line of work asks whether reliability is already encoded in model internals. Earlier studies found that language models can express calibrated knowledge or truthfulness-related information in their internal states (Kadavath et al., 2022; Azaria & Mitchell, 2023; Burns et al., 2024), making internal-state detection plausible beyond output-only evaluation. Recent approaches use simple factuality probes, prompt-guided internal states, or the information contribution of residual-stream updates to detect hallucinations (Han et al., 2025; Zhang et al., 2025a;c). Probe-based methods, however, tie the detector to a selected representation, while update-based methods focus on a particular statistic of the computation. These approaches therefore leave open how a response-level signal evolves across the full depth of a generated response and whether the internal units used to construct it are causally important for the computation.

Mechanistic analysis and sparse computation. Mechanistic interpretability studies Transformer computation through circuit descriptions and component-level analyses (Elhage et al., 2021; Geva et al., 2021; Dai et al., 2022), as well as causal tracing and activation interventions (Meng et al., 2022; Massidda et al., 2023). Recent work frames such interventions through causal mediation, emphasizing that the mediator determines what a mechanistic claim can establish (Mueller et al., 2026). This literature tests whether an internal component matters for a computation, but does not itself produce a response-level hallucination detector. In parallel, sparsity research studies pruning and targeted low-rank refinement (Frantar & Alistarh, 2023; Shen et al., 2025; Hou et al., 2025), alongside conditional computation through dynamic routing (Fedus et al., 2022; Lv et al., 2025), for efficient inference. These methods target efficiency or component-level function rather than the cross-layer organization of unmodified activations. Our approach connects the two perspectives by summarizing calibrated large activation events with a layer-wise entropy profile and scoring the descent prominence of its dominant smooth local valley; matched-sparsity causal masking then tests whether the same event definition captures computation that matters for model behavior.

3 Problem Formulation

Consider an autoregressive language model with L Transformer layers and hidden dimension d. Given a prompt x, the model generates a response $\boldsymbol{y} = (y_1, \ldots, y_n)$ of n tokens. During this computation, the model produces internal activation signals whose organization across layers and token positions may differ between a reliable response and a hallucinated one. Our goal is to identify a response-level signal from the activations collected along the model's autoregressive response trajectory, without changing the model or requiring external evidence.

At a fixed hook point in layer $\ell \in \{1, \ldots, L\}$, let

$$\boldsymbol{A}_{\ell,t} = (A_{\ell,t,1}, \ldots, A_{\ell,t,d}) \in \mathbb{R}^d \tag{1}$$

denote the activation vector for generated token y_t. The coordinate $A_{\ell,t,i}$ denotes the i-th entry of this vector.

The detector receives the prompt, generated response, and activation collection $\{\boldsymbol{A}_{\ell,t}\}_{\ell,t}$. It should summarize the cross-layer organization of the internal signals into a response-level estimate of reliability, while leaving the generated response unchanged. For evaluation, the ground-truth label is $z = 1$ for a hallucinated response and $z = 0$ for a truthful response; a detector may optionally produce a binary prediction $\widehat{z} \in \{0, 1\}$. The resulting signal can be used to decide which responses warrant verification or abstention.

4 Event Profile Detection

We investigate whether signals intrinsic to an LLM during inference can be used to detect hallucinations without relying on an external model. A useful signal should be comparable across layers, summarize behavior over the whole response, and expose when a set of activation coordinates becomes stable as the computation develops. We therefore focus on large activation events, calibrate their frequency separately at each layer, and trace how their coordinate-wise firing patterns evolve across the generated tokens. We summarize this evolution with a layer-wise entropy profile and use the prominence of a smooth entropy valley as a score to detect hallucinations in LLM-generated responses. An overview of the complete procedure is shown in Figure 1.

4.1 Event Calibration

Raw activation magnitudes vary across layers, so one shared numerical threshold would retain different fractions of coordinates at different depths. Let $k \in (0, 1)$ denote the target event rate. For each layer ℓ, we set $q_\ell(k)$ to the value for which a fraction k of the absolute activation values in the calibration collection $\mathcal{C}$ lie above the threshold. Equivalently, $q_\ell(k)$ is the $(1 - k)$ quantile of those values. For an evaluated response, define the binary event indicator

$$S_{\ell,t,i} = \mathbb{I}(|A_{\ell,t,i}| \geq q_\ell(k)), \quad \ell \in \{1, \ldots, L\}, \quad t \in \{1, \ldots, n\}, \quad i \in \{1, \ldots, d\}. \tag{2}$$

Here, $\mathbb{I}(\cdot)$ equals one when its condition is true and zero otherwise. Thus, $S_{\ell,t,i} = 1$ means that coordinate i of $\boldsymbol{A}_{\ell,t}$ is a large activation event. The absolute value allows both large positive and

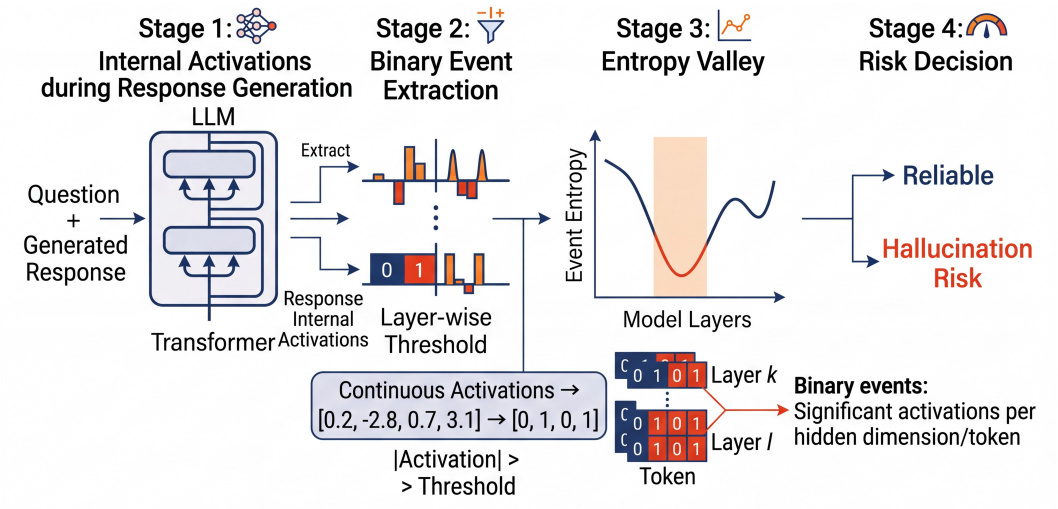

Figure 1: Overview of the proposed training-free hallucination detection method. The method records internal activations during response generation, converts large-magnitude coordinates into layer-calibrated binary events, aggregates their temporal firing patterns into a layer-wise binary-entropy profile, and extracts a smooth entropy valley for the final hallucination-risk decision.

large negative deviations to be selected. Because each $q_\ell(k)$ is calibrated separately, the event rate is comparable across layers without assuming that activations follow a Gaussian distribution.

For later use, collect the d scalar indicators at layer ℓ and token position t into the vector $\boldsymbol{S}_{\ell,t} = (S_{\ell,t,1}, \ldots, S_{\ell,t,d}) \in \{0,1\}^d$. The same event vector supports two distinct branches. In the diagnostic branch, $\boldsymbol{S}_{\ell,t}$ is only counted and the model activations remain unchanged. In the causal branch, inference replaces the original activation by

$$\widehat{\boldsymbol{A}}_{\ell,t} = \boldsymbol{A}_{\ell,t} \odot \boldsymbol{S}_{\ell,t}, \tag{3}$$

where $\widehat{\boldsymbol{A}}_{\ell,t}$ is the masked activation passed to the remaining computation, and $\odot$ denotes coordinate-wise multiplication. The diagnostic branch asks whether the event pattern predicts hallucination; the causal branch asks whether the selected coordinates matter for model behavior.

4.2 From Response Trajectories to Layer-wise Profiles

The event mask is defined at one token position, whereas the label of interest is assigned to a complete response. A single-token event can reflect a transient lexical or syntactic operation and is therefore insufficient to characterize the response as a whole. We instead ask whether each coordinate follows a repeatable event pattern throughout the generated trajectory. For coordinate i at layer ℓ, its response-level firing rate is

$$p_{\ell,i} = \frac{1}{n} \sum_{t=1}^{n} S_{\ell,t,i}, \quad \ell \in \{1, \ldots, L\}, \quad i \in \{1, \ldots, d\}. \tag{4}$$

This quantity is not a second threshold or an additional event definition. It is the fraction of answer positions at which the already-defined event occurs, and is thus the minimal summary needed to distinguish a persistent pattern from an isolated activation. The normalization by n also makes responses of different lengths comparable. We write $\boldsymbol{p}_\ell = (p_{\ell,1}, \ldots, p_{\ell,d})$ for the event profile at layer ℓ.

The firing rate alone does not distinguish a coordinate that is almost never selected from one that is selected at nearly every position; both correspond to a regular pattern. Conversely, a rate near one half indicates that the event alternates substantially across the response. We capture this distinction by treating the event sequence for each coordinate as a Bernoulli process and measuring its uncertainty

with binary entropy:

$$h_\ell = \frac{1}{d}\sum_{i=1}^{d}\left[-p_{\ell,i}\log p_{\ell,i} - (1-p_{\ell,i})\log(1-p_{\ell,i})\right], \tag{5}$$

with the convention $0\log 0 = 0$. Entropy is low when event occurrence is regular (either rarely firing or persistently firing) and high when events occur intermittently. Thus, a decrease in h_ℓ indicates that the event pattern has become more temporally organized at layer ℓ.

4.3 Scoring a Coherent Entropy Valley

A decrease in binary entropy across layers marks a transition toward more regular activation events. When the profile subsequently rises, this decrease forms an *entropy valley*. We focus on the descending segment because it captures the depth-wise stabilization of event patterns. We quantify its depth and smoothness to obtain a response-level score for hallucination detection.

We first denoise the profile using a three-layer symmetric moving average. This is the smallest symmetric window that incorporates both immediate neighbors while preserving layer resolution, and is used to suppress small layer-to-layer fluctuations:

$$\widetilde{h}_\ell = \frac{h_{\ell-1} + h_\ell + h_{\ell+1}}{3}, \quad \ell \in \{2,\dots,L-1\}, \tag{6}$$

with $\widetilde{h}_1 = h_1$ and $\widetilde{h}_L = h_L$. The set of local minima of the smoothed profile is

$$\mathcal{V}_{\text{local}} = \left\{\ell \in \{2,\dots,L-1\} : \widetilde{h}_\ell \le \widetilde{h}_{\ell-1} \text{ and } \widetilde{h}_\ell \le \widetilde{h}_{\ell+1}\right\}. \tag{7}$$

For each candidate minimum ℓ, we measure the descent into the minimum from its left side:

$$\mathcal{J}_\ell^{\text{left}} = \{j_{\ell,1}, j_{\ell,2}, \dots, j_{\ell,m_\ell}\}, \qquad 1 \le j_{\ell,1} < j_{\ell,2} < \cdots < j_{\ell,m_\ell} < \ell,$$

where a candidate shoulder is a local maximum,

$$\widetilde{h}_j \ge \widetilde{h}_{j-1} \quad \text{and} \quad \widetilde{h}_j \ge \widetilde{h}_{j+1}, \qquad j \in \{2,\dots,\ell-1\}. \tag{8}$$

For a candidate shoulder $j_{\ell,r}$, the corresponding descent segment and its prominence are

$$\begin{aligned} \boldsymbol{x}_{\ell,r}^{\text{left}} &= \left(\widetilde{h}_{j_{\ell,r}}, \dots, \widetilde{h}_\ell\right), \\ P_{\ell,r} &= \max\left(\widetilde{h}_{j_{\ell,r}} - \widetilde{h}_\ell,\, 0\right). \end{aligned} \tag{9}$$

The prominence is the amount by which the profile falls from the shoulder to the candidate bottom. It is therefore a direct measure of the strength of the proposed stabilization, rather than a generic measure of activation magnitude.

We additionally require the descent to be smooth. For each segment, let $X_{\ell,r,k}^{\text{left}}$ be its discrete Fourier transform and define

$$\rho_{\ell,r}^{\text{high}} = \frac{E_{\ell,r}^{\text{high}}}{E_{\ell,r}^{\text{total}}}, \qquad E_{\ell,r}^{\text{total}} = \sum_k \left|X_{\ell,r,k}^{\text{left}}\right|^2, \qquad E_{\ell,r}^{\text{high}} = \sum_{k>k_0} \left|X_{\ell,r,k}^{\text{left}}\right|^2. \tag{10}$$

Here k_0 separates low and high Fourier modes. A small $\rho_{\ell,r}^{\text{high}}$ indicates that the descent is not produced by rapid layer-to-layer oscillations. Let $\eta \in [0,1]$ denote the maximum accepted high-frequency energy fraction. The valid shoulder indices and the selected shoulder for minimum ℓ are

$$\mathcal{R}_\ell = \left\{r : P_{\ell,r} > 0,\ \rho_{\ell,r}^{\text{high}} \le \eta\right\}, \qquad r_\ell^* = \arg\max_{r\in\mathcal{R}_\ell} P_{\ell,r}.$$

We retain only local minima with at least one valid descent:

$$\mathcal{L}_{\text{cand}} = \{\ell \in \mathcal{V}_{\text{local}} : \mathcal{R}_\ell \neq \emptyset\}. \tag{11}$$

If several adjacent layers form a tied shoulder plateau, its earliest layer is used as the representative, so one shoulder region cannot contribute several copies of the same descent.

The response-level statistic is the strongest accepted descent:

$$\ell_v = \arg\max_{\ell \in \mathcal{L}_{\text{cand}}} P_{\ell, r^*_\ell}, \qquad V_{\text{BEVP}} = P_{\ell_v, r^*_{\ell_v}}. \tag{12}$$

We call V_{BEVP} the binary-entropy valley-prominence score. If no valid valley exists, we set $V_{\text{BEVP}} = 0$. The reason this statistic can distinguish response types is therefore explicit: it is large only when a response contains a substantial, smooth transition from an unstable event pattern to a stable one over depth.

The firing-rate profile can be accumulated during generation with one counter vector per layer, so the full event tensor need not be stored; after generation, the entropy profile and valley score are computed from the resulting counters.

4.4 Valley Threshold and Decision Rule

The continuous BEVP score ranks responses by the strength of their entropy valley, but a binary detector also needs to decide whether a drop is large enough to be meaningful for that response. A fixed absolute threshold would be inappropriate because the overall entropy scale can vary across responses and models. We therefore normalize the cutoff by the observed range of the smoothed profile. Let

$$\widetilde{h}_{\min} = \min_\ell \widetilde{h}_\ell, \qquad \widetilde{h}_{\max} = \max_\ell \widetilde{h}_\ell,$$

and let $\alpha \in (0, 1)$ specify the minimum fraction of this range required for a valley to count as substantial. Defining the hallucination-risk score as the negative of the truthfulness-oriented BEVP score gives

$$D_{\text{hall}} = -V_{\text{BEVP}}, \qquad \tau_D = -\alpha \left(\widetilde{h}_{\max} - \widetilde{h}_{\min} \right), \qquad \widehat{z} = \mathbb{I}(D_{\text{hall}} > \tau_D). \tag{13}$$

Equivalently, the detector labels a response as hallucinated when $V_{\text{BEVP}} < \alpha(\widetilde{h}_{\max} - \widetilde{h}_{\min})$. This direction is important: a deep accepted valley lowers the hallucination-risk score, while an absent or shallow valley raises it.

For one response, constructing events and accumulating firing rates examines $O(Ldn)$ activation coordinates. The subsequent profile operations require $O(L)$ additional work, apart from the Fourier transforms of candidate descent segments, and streaming requires $O(Ld)$ auxiliary memory.

5 Experiments

5.1 Experimental Setup

Implementation Details. The experiments use the LLaMA-3-8B and LLaMA-3-70B decoder-only models (Grattafiori et al., 2024). Activation events are measured at the residual-update hook point used throughout the method section, so the diagnostic statistic and the intervention masks are built from the same internal representation. For each layer, thresholds are calibrated separately so that the retained event fraction is fixed at $k = 5\%$ by default. Causal masking is evaluated during model inference. For the TruthfulQA profile study, activations are collected only along the answer-token trajectory while conditioning on the same question (Lin et al., 2022); this paired teacher-forcing design compares a dataset-provided correct answer with a dataset-provided incorrect answer for each question. In the detection evaluation (Figure 3), LLaMA-3-8B instead generates answers from questions and task-required context alone, with BEVP computed from the generated answer-token activations. For each model, event thresholds are calibrated once without evaluation labels and reused on held-out responses; no classifier is fitted to the detection labels. We fix $k = 5\%$, $\eta = 0.05$, and the hallucination-risk direction $-V_{\text{BEVP}}$ before evaluation. For the depth-localization intervention, $\alpha = 1/5$ defines the valley core as the maximal contiguous low-entropy interval around ℓ_v; the remaining layers are pre- and post-valley regions.

Baselines. The unmodified dense model serves as the reference for causal accuracy. To isolate the effect of magnitude-based event selection, we compare **top-10%, top-5%, and top-1% masks with random 5% masks and inverse-magnitude 5% masks**. These controls keep the same nominal

sparsity while removing the selection rule that defines the proposed event profile. We also compare percentile, z-score, and median absolute deviation (MAD) thresholding under the same kept fraction. For hallucination detection, the proposed smooth entropy-valley score is compared with minimum layer entropy, all-layer mean entropy, a fixed early–middle contrast used only as a positional baseline, hidden-state norm, and a layer-order-permuted event-profile control. These comparisons test whether discrimination comes from the ordered, smooth local-valley structure rather than an overall activation scale, a fixed layer partition, or a generic entropy summary.

Dataset & Metric. GSM8K evaluates causal masking in mathematical reasoning using answer accuracy and relative accuracy (Cobbe et al., 2021). TruthfulQA provides paired correct and incorrect answers for the layer-wise profile study (Lin et al., 2022); HaluEval supplies a separate hallucination-detection setting (Li et al., 2023). For detection on all three datasets, annotators blind to the BEVP scores judge the generated answers after scoring; human-judged hallucinations form the positive class. AUROC and AUPRC compare the scores with these labels, with question-level bootstrap confidence intervals where reported. Additional causal and stability measures are defined in the table captions.

5.2 Causal Validation of Large Activation Events

Causal result. Before using large activation events as a diagnostic, we test whether they retain information needed to solve GSM8K problems when masked activations are fed back into inference. In Table 1, retaining the largest 5% yields 65.1% accuracy versus 68.4% for dense inference, a drop of 3.3 percentage points; retaining only 1% lowers accuracy to 42.8%. The strong change between these budgets shows that a small set of large events carries much of the task-relevant computation, but also that an overly restrictive event budget discards it.

Matched-sparsity controls. To determine whether this retention depends on magnitude rather than sparsity alone, we compare masks at the same 5% budget. Random and inverse-magnitude masks achieve only 12.4% and 8.2% accuracy, versus 65.1% for top-magnitude events; output KL and hidden-state distortion are also lowest for the latter (0.015 and 0.037). These controls support the event-selection rule as a causal probe of reasoning computation, not by themselves as evidence of hallucination-detection performance.

Table 1: Causal masking on GSM8K with LLaMA-3-8B. Relative is accuracy divided by the dense-baseline accuracy. Dense-Output KL is the mean token-level divergence $D_{\mathrm{KL}}(p_t^{\mathrm{dense}} \| p_t^{\mathrm{masked}})$ under the same teacher-forced prefixes. Relative Hidden-State Distortion is the mean relative ℓ_2 difference between masked and dense hidden states across tokens and layers. Retaining the largest activation events preserves accuracy, output distributions, and internal representations substantially better than controls with the same nominal sparsity.

Configuration	Accuracy (%)	Relative (%)	Output KL ↓	Hidden-State Distortion ↓
Dense baseline	68.4	100.0	0.000	0.000
Top-10% events	67.5	98.6	0.006	0.018
Top-5% events	**65.1**	**95.1**	**0.015**	**0.037**
Top-1% events	42.8	62.5	0.085	0.114
Random 5%	12.4	18.1	0.220	0.245
Inverse 5%	8.2	11.9	0.310	0.302

5.3 Stable Event Profiles and Entropy Valleys

Layer-wise profile. We next examine whether the event profile differs between correct and incorrect answers across depth. Figure 2 shows mean binary event entropy h_ℓ on paired TruthfulQA answers for LLaMA-3-8B (layers 0–31) and LLaMA-3-70B (layers 0–79); unlike the plots' zero-based labels, Section 4 indexes layers from one.

Dominant smooth entropy valley. In Figure 2, correct 8B answers show an early low-entropy basin around layers 5–8 followed by recovery, whereas the 70B basin is broader and occurs later in the network. The shift across model sizes motivates scoring a response's smooth descent locally, as in Section 4.3, rather than selecting one absolute layer range for both models.

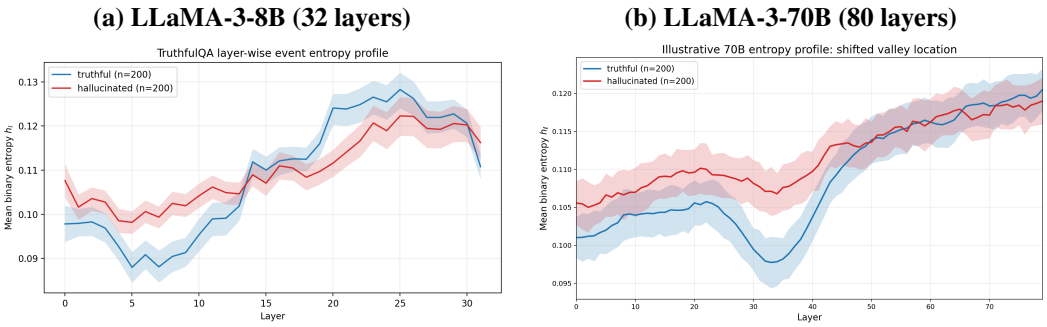

Figure 2: Layer-wise mean binary event entropy on paired TruthfulQA answers at two model scales. (a) LLaMA-3-8B with 32 layers; (b) LLaMA-3-70B with 80 layers. In both panels, blue denotes dataset-provided correct answers and red denotes dataset-provided incorrect answers; shaded bands show bootstrap confidence intervals over 200 answers per class. The 8B profile exhibits a pronounced early low-entropy basin, whereas the 70B profile exhibits a broader basin at later, middle-depth layers.

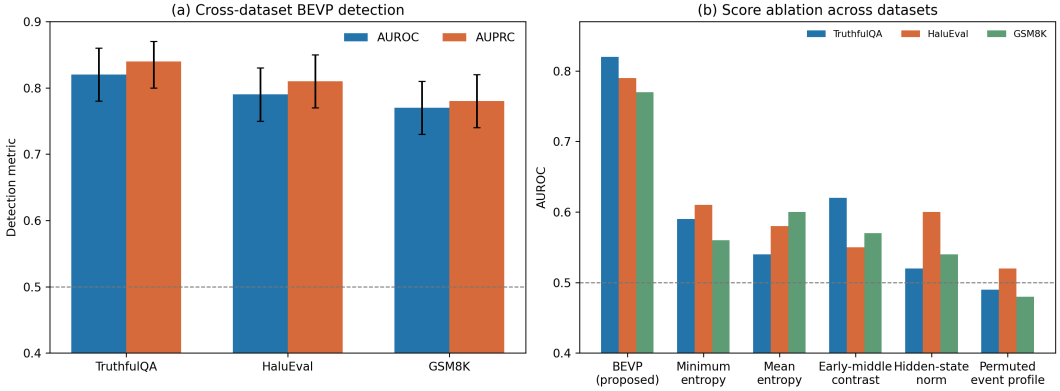

Figure 3: Cross-dataset BEVP detection and score ablation with LLaMA-3-8B. (a) AUROC and AUPRC on TruthfulQA, HaluEval, and GSM8K; error bars denote question-level bootstrap 95% confidence intervals. (b) AUROC of BEVP and alternative score summaries on the same datasets.

5.4 Hallucination Detection with the BEVP Score

Cross-dataset detection. For LLaMA-3-8B, Figure 3(a) reports AUROC/AUPRC of 0.82/0.84 on TruthfulQA, 0.79/0.81 on HaluEval, and 0.77/0.78 on GSM8K. AUROC remains above 0.75 across all three datasets, including the mathematical-reasoning setting, supporting the transfer of the same valley score beyond paired factual answers. Error bars denote question-level bootstrap 95% confidence intervals.

Score ablation. Figure 3(b) shows that BEVP reaches AUROC values of 0.82, 0.79, and 0.77 on TruthfulQA, HaluEval, and GSM8K, respectively. The strongest alternative on each dataset reaches only 0.62 (early–middle contrast), 0.61 (minimum entropy), and 0.60 (mean entropy), gaps of 0.20, 0.18, and 0.17. The permuted event profile scores 0.49, 0.52, and 0.48, indicating that layer order contributes to discrimination; the other entropy and hidden-state summaries likewise do not recover the full valley signal.

5.5 Causal Localization of the Entropy Valley

Valley-stage intervention. We next test whether the low-entropy region is especially sensitive to causal masking. Table 2 applies the same masking intervention to each region on GSM8K. Masking the valley core, which comprises only 12.5% of the layers, causes a 25.03-point accuracy drop, decreases BEVP by 0.0092, and raises mean token NLL from 0.386 to 0.712. By comparison,

Table 2: Causal masking across entropy-valley regions on GSM8K with LLaMA-3-8B. Relative is accuracy divided by the dense-baseline accuracy. Masked Layer Fraction is $r_R = |R|/L$, where R is the intervention region and $L = 32$; the measured partition uses layers 0–4 for pre-valley context, 5–8 for the valley core, and 9–31 for post-valley recovery. ΔV_{BEVP} is the post-intervention minus dense-baseline BEVP score, so a negative value indicates that masking weakens the valley prominence. Mean Token NLL is the teacher-forced average negative log-probability of the answer tokens, $\mathrm{NLL} = -(1/n)\sum_t \log p(y_t \mid x, y_{<t})$, where lower values indicate greater confidence in the answer tokens.

Region	Accuracy (%)	Relative (%)	Masked Layers (%)	ΔV_{BEVP}	Mean Token NLL
Dense baseline	68.40	100.0	0.0	0.000	0.386
Pre-valley context	66.83	97.7	15.6	-0.0015	0.421
Valley core	**43.37**	**63.4**	**12.5**	**-0.0092**	**0.712**
Post-valley recovery	67.65	98.9	71.9	-0.0006	0.408

masking the pre-valley and post-valley regions causes accuracy drops of only 1.57 and 0.75 points, respectively, with substantially smaller BEVP changes. This concentration of behavioral and profile-level effects supports interpreting the smooth entropy valley as a marker of structured computation.

5.6 Robustness of the Event Definition

Threshold comparison. Table 3 tests whether the GSM8K masking result depends on percentile thresholding. At the same 5% kept fraction, percentile, z-score, and MAD rules achieve 65.1%, 64.7%, and 65.2% accuracy, respectively: a 0.5-point range, with across-run standard deviations of 0.7–0.8 points. This supports robustness of the causal masking result to the threshold rule, but does not establish robustness of the hallucination detector.

Table 3: Threshold robustness on GSM8K with LLaMA-3-8B at matched 5% sparsity. Relative is measured against the dense-baseline accuracy of 68.4%. Std. is the standard deviation of accuracy across repeated runs, and CV is the coefficient of variation, $100 \times \mathrm{Std.}/\mathrm{Accuracy}$.

Threshold rule	Accuracy (%)	Relative (%)	Std. (pp)	CV (%)
Percentile	65.1	95.2	0.7	1.08
Z-score	64.7	94.6	0.8	1.24
MAD	65.2	95.3	0.7	1.07

6 Conclusion

This paper reframes hallucination detection as a problem of internal trajectories across layers and introduces a training-free detector based on large activation events and the strongest valid left-side descent of a dominant smooth local entropy valley. The empirical argument follows the proposed evidence chain: causal masking on GSM8K first shows that the selected events are computationally informative; layer-wise profiles then reveal a decrease-and-recovery entropy pattern for correct answers; and BEVP converts that pattern into a response-level hallucination score. On TruthfulQA, LLaMA-3-8B achieves an AUC of 0.82 and LLaMA-3-70B achieves an AUC of 0.79. Valley-stage and threshold analyses further localize the causal sensitivity and assess the stability of the event definition. The resulting framework offers a compact diagnostic that can be computed during ordinary inference and used to trigger downstream verification.

AI use statement

In this work, we did not use generative AI tools for any tasks requiring disclosure. We used generative AI tools to summarize or analyze existing literature, discover research topics or identify gaps, brainstorm ideas, and source or search for information. We have reviewed all AI-assisted work. In particular, we reviewed AI-assisted literature summaries, search results, and suggested ideas for accuracy and relevance, and made the final research decisions ourselves. We take responsibility for

the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Amos Azaria and Tom M. Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21224–21236, 2021.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8493–8502, 2022.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Christopher Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23:1–39, 2022.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pp. 10323–10337, 2023.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhishek Pandey, Adam Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaidyanathan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiatong Han, Neil Band, Muhammed Razzak, Jannik Kossen, Tim G. J. Rudner, and Yarin Gal. Simple factuality probes detect hallucinations in long-form natural language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16209–16226, 2025. doi: 10.18653/v1/2025.findings-emnlp.880.

Bairu Hou, Qibin Chen, Jianyu Wang, Guoli Yin, Chong Wang, Nan Du, Ruoming Pang, Shiyu Chang, and Tao Lei. Instruction-following pruning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 23894–23909, 2025.

Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Jan Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34728–34745, 2025. doi: 10.18653/v1/2025.emnlp-main.1761.

Saurav Kadavath, Tom Conerly, Amanda Askell, Collin Burns, Trenton Nguyen, Sergey Ioffe, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Junyi Li, Xiaoxue Cheng, Wayne 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, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22922–22935, 2022.

Minxuan Lv, Zhenpeng Su, Leiyu Pan, Yizhe Xiong, Zijia Lin, Hui Chen, Wei Zhou, Jungong Han, Guiguang Ding, Wenwu Ou, Di Zhang, Kun Gai, and Songlin Hu. Dsmoe: Matrix-partitioned experts with dynamic routing for computation-efficient dense llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19711–19722, 2025. doi: 10.18653/v1/2025.emnlp-main.997.

Potsawee Manakul, Adian Liusie, and Mark J. F. 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, 2023.

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213, pp. 68–87. PMLR, 2023.

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, 2022.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: Surveying mechanistic interpretability for nlp through the lens of causal mediation analysis. *Computational Linguistics*, 52(1):331–378, 2026. doi: 10.1162/coli.a.572.

Li Shen, Anke Tang, Yong Luo, Tao Sun, Han Hu, and Xiaochun Cao. Targeted low-rank refinement: Enhancing sparse language models with precision. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 54457–54475, 2025.

Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. Prompt-guided internal states for hallucination detection of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21806–21818, 2025a. doi: 10.18653/v1/2025.acl-long.1058.

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.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. Icr probe: Tracking hidden state dynamics for reliable hallucination detection in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17986–18002, 2025c. doi: 10.18653/v1/2025.acl-long.880.