# Uncovering the Knowledge Boundaries and Mechanistic "Knob" of LLMs

**Zihao Huang, Rui Mao, Yong Jiang, and Erik Cambria**
College of Computing and Data Science, Nanyang Technological University, Singapore
Tongyi Lab, Alibaba Group, China
zihao016@e.ntu.edu.sg, {rui.mao, cambria}@ntu.edu.sg
yongjiang.jy@alibaba-inc.com

## Abstract

Large Language Models often generate incorrect factual answers, yet it remains unclear how they internally distinguish facts they know from those they do not. Existing hallucination-detection methods largely rely on behavioral signals or opaque scalar scores, leaving the structure of factual knowledge in hidden state space poorly understood. We address this gap by introducing a coarse-to-fine framework to discover a knowledge boundary: a decision surface in transformer hidden states that separates representations leading to correct factual answers from those leading to errors. Across multiple datasets and model families, we show that factual correctness is predictable from hidden states, and that this signal is concentrated in specific layers and a small number of influential dimensions. Using Qwen3-8B as a case study, we further identify a compact "knob" subspace that retains near-full boundary-prediction performance and supports targeted interventions that bidirectionally modulate the likelihood of the full reference answer under teacher forcing.

## 1 Introduction

Although LLMs are widely deployed, they frequently generate incorrect responses, a phenomenon referred to as hallucination (Ji et al., 2023; Mao et al., 2024; Zhang et al., 2025b). Enhancing their reliability requires understanding how LLMs differentiate prompts they can answer correctly from those they cannot (Petroni et al., 2019), as well as how relevant information is structured in context (Mao et al., 2026). However, how this distinction manifests in their internal representations, such as transformer hidden states, remains poorly understood. Current hallucination detection methods often trade off efficiency, coverage, and interpretability (Huang et al., 2025). Behavioral detectors treat LLMs as black boxes and rely solely on generated outputs.

For example, self-consistency approaches sample multiple responses or reasoning paths and compare their consistency (Manakul et al., 2023; Wang et al., 2023). Uncertainty-based methods estimate how unsure a model is, but they are not directly designed to reveal which internal representations implement factual decisions (Zhu et al., 2026). Semantic entropy detects a subset of hallucinations by sampling multiple answers and clustering them by meaning, but it does not guarantee factuality when the model is systematically and confidently wrong (Farquhar et al., 2024). Hidden state methods expose internal signals, yet many detectors still reduce them to a single scalar score, leaving the decision rule and its underlying geometry difficult to inspect (Zhang et al., 2025c). These limitations leave a persistent gap in our understanding of how knowledge-related decision boundaries are encoded within transformer hidden states and how such internal signals relate to observable model behavior.

This gap motivates us to study knowledge boundaries within LLMs more systematically. We define the knowledge boundary as a decision surface in hidden state space that separates internal states leading to correct factual answers from those leading to incorrect ones. Our goal is not only to detect errors, but to characterize the underlying structure of this boundary in a way that is both compact and interpretable. We hypothesize that the boundary is governed by a small number of influential hidden dimensions, rather than being spread diffusely across the full representation. We refer to such influential dimensions as candidate "knobs": internal coordinates that retain near-full boundary-prediction performance within a compact subspace and support controlled interventions along the distilled boundary direction. Identifying such knobs matters because it turns an abstract notion of "knowing" into a concrete, testable mechanism that links internal representations to observable behavior.



*Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 26679–26694
October 24-29, 2026 ©2026 Association for Computational Linguistics

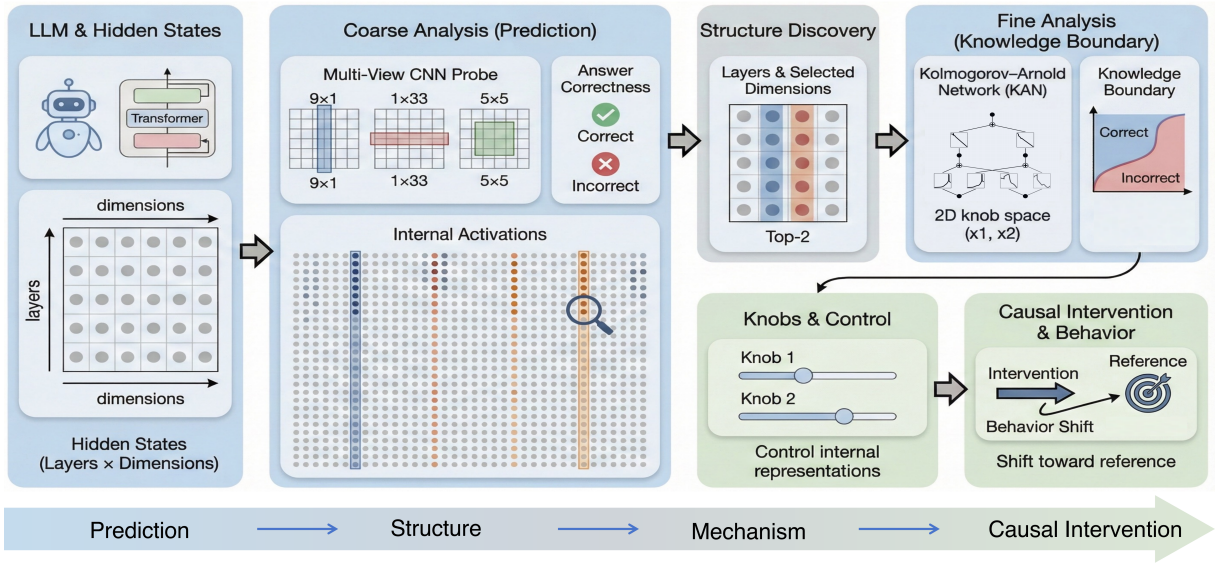



Figure 1: A coarse-to-fine framework for discovering knowledge boundaries and internal knobs in LLMs.

In this work, we aim to address three research questions (RQs). **RQ1.** Can knowledge boundaries be learned from transformer hidden states, across different datasets and model families? **RQ2.** If such a boundary exists, what structural signatures in hidden states distinguish correct from incorrect factual behavior? **RQ3.** In a case study of Qwen3-8B, can the boundary be captured by a compact knob subspace, and can interventions on this subspace directionally modulate the likelihood of the reference answer? Together, these questions move from prediction to characterization, and ultimately to a mechanistic explanation of how internal representations give rise to factual decisions.

To study these questions, we introduce a simple coarse-to-fine framework that gradually uncovers how factual knowledge is organized inside transformer hidden states (see Figure 1). We begin by training a multi-view CNN probe to predict answer correctness from hidden states on the PopQA and TriviaQA benchmarks (Mallen et al., 2023; Joshi et al., 2017), and validate the findings across multiple model families. By inspecting the CNN's internal activations, we identify layer-wise and dimension-level patterns that distinguish correct from incorrect factual behavior. Guided by these patterns, we apply a Kolmogorov–Arnold Network (Liu et al., 2025) (KAN) to a low-dimensional knob space to summarize the boundary as a compact, low-dimensional rule. The dimensions governing this rule are treated as candidate knobs—internal variables that can be adjusted to influence the model's factual decisions.

To test their causal role, we intervene on these knobs and track how the model's preference for a reference answer changes. Our contributions are threefold. 1) We present a systematic and interpretable analysis of how transformer hidden states relate to factual correctness in LLMs. 2) Through a focused case study of Qwen3-8B, we identify a compact knob subspace that retains near-full boundary-prediction performance and supports directional intervention on reference-answer likelihood. 3) We introduce a general coarse-to-fine framework that links prediction, structural analysis, and causal intervention, offering a practical pathway toward mechanistic explainability in LLMs.

## 2 Problem Setup

In this section, we formalize the notion of a knowledge boundary introduced above and specify the hidden state representations and labels used in our analysis. We also define knobs as a compact set of hidden dimensions that can recover near-full predictive performance for this boundary. For each prompt $x$, we extract the transformer hidden states at the last prompt token, just before generation begins. We denote this representation by $H(x) \in \mathbb{R}^{L \times d}$, where $L$ is the number of layers and $d$ is the hidden dimension. We assign a binary label $y(x) \in \{0, 1\}$ using an external evaluator, where $y(x) = 1$ indicates a correct answer and $y(x) = 0$ indicates an incorrect one. We model factual correctness with a scoring function $g : \mathbb{R}^{L \times d} \to \mathbb{R}$, where the sign of $g(H(x))$ predicts whether $x$ is likely to be answered correctly.

If $g(H(x))$ separates $y = 1$ from $y = 0$ on a held-out test set, we say that $g$ captures a learnable decision rule in the hidden state space. As shown in Eq. (1), the boundary induced by $g$ is defined as the zero level set.

$$\mathcal{B} = \left\{ H \in \mathbb{R}^{L \times d} \,:\, g(H) = 0 \right\}. \tag{1}$$

This boundary partitions hidden states into two regions, $\{g(H) > 0\}$ and $\{g(H) < 0\}$, corresponding to states predicted as correct and incorrect.

To test whether the boundary is controlled by only a few dimensions (knobs) of hidden states, we consider an index set $\mathcal{J} \subset [d]$ with $k = |\mathcal{J}| \ll d$. Given $\mathcal{J}$, we form the sub-representation $Z_{\mathcal{J}}(H) = H[:, \mathcal{J}] \in \mathbb{R}^{L \times k}$ by keeping only these dimensions across all layers. Let $\mathrm{Acc}(\mathcal{J})$ denote the best test accuracy achievable by a probe trained using only $Z_{\mathcal{J}}(H(x))$, and let $\mathrm{Acc}_{\mathrm{full}}$ be the test accuracy of a reference probe trained on all $d$ dimensions. We call $\mathcal{J}$ a near-optimal set if its test accuracy drops by at most $\delta$ compared to the full baseline, i.e., $\mathrm{Acc}(\mathcal{J}) \geq \mathrm{Acc}_{\mathrm{full}} - \delta$. We define the knob set $\mathcal{J}_{\mathrm{knob}}$ as the first near-optimal subset within a nested candidate sequence, as specified in Eq. (2). In our implementation, this sequence is instantiated by the importance-ranked Top-$k$ subsets described in Section 3.

$$\begin{aligned} k_{\mathrm{knob}} &= \min \left\{ k : \mathrm{Acc}(\mathcal{J}_k) \geq \mathrm{Acc}_{\mathrm{full}} - \delta \right\}, \\ \mathcal{J}_{\mathrm{knob}} &= \mathcal{J}_{k_{\mathrm{knob}}}. \end{aligned} \tag{2}$$

This definition ensures that knobs are not merely predictive features, but a compact subset that retains essentially all boundary information within the considered search sequence.

## 3 Methodology

We introduce a coarse-to-fine method for analyzing the knowledge boundary in transformer hidden states. The section is structured to address RQ1–RQ3 with increasing depth. We hypothesize that boundary evidence is concentrated in a small set of hidden dimensions (knobs), which motivates low-dimensional distillation and causal probing.

### 3.1 CNN probe for boundary prediction

For RQ1, we instantiate the scoring function in Section 2 with a multi-view CNN probe that predicts answer correctness from hidden states. Given the last-token hidden states $H(x) \in \mathbb{R}^{L \times d}$, we treat $H(x)$ as a single-channel 2D representation indexed by layers and dimensions.

The probe outputs logits $f_\theta(H(x)) \in \mathbb{R}^2$, and we define the boundary score as the logit difference $g_\theta(H) = f_\theta(H)_1 - f_\theta(H)_0$. The network begins with a LayerNorm over the hidden dimension to stabilize scale differences across coordinates. It then applies three convolutional branches with fixed kernels: a **vertical** branch with kernel size $(9, 1)$, a **horizontal** branch with kernel size $(1, 33)$, and a **local** branch with kernel size $(5, 5)$. Appendix C shows that the probe remains stable under reasonable kernel-size variants. These kernels are designed to capture complementary structure along the layer axis, the dimension axis, and their local interactions, respectively. Each branch uses a ReLU nonlinearity followed by global max pooling, and the pooled features are concatenated and fed to a linear classifier. We train the probe by minimizing cross-entropy on labeled pairs $(H(x), y(x))$. We train this CNN not only for prediction, but also for interpretability: its convolutional branches produce layer-dimension activation maps that we inspect in RQ2 to localize boundary evidence in $H$.

### 3.2 Boundary evidence in hidden states

For RQ2, we analyze the trained CNN probe by inspecting its convolutional activation maps over the layer-dimension grid. For each input $H(x)$, we record the pre-pooling activations from a chosen branch, yielding a tensor $A(x) \in \mathbb{R}^{C \times L \times d}$. We summarize channels by averaging over $C$, producing a single activation map $\tilde{A}(x) \in \mathbb{R}^{L \times d}$ that aligns with the geometry of $H(x)$. To compare correct and incorrect states, we compute class-conditional mean maps $\bar{A}^{(1)} = \mathbb{E}[\tilde{A}(x) \mid y(x) = 1]$ and $\bar{A}^{(0)} = \mathbb{E}[\tilde{A}(x) \mid y(x) = 0]$ on a split of interest. We then form a difference map $\Delta A = \bar{A}^{(1)} - \bar{A}^{(0)}$ to visualize where boundary evidence changes across layers and dimensions. To obtain a dimension-level importance score, we aggregate the difference map along the layer axis and define, for each hidden dimension $j \in [d]$, $s_j = \sum_{\ell=1}^{L} |\Delta A_{\ell j}|$. Sorting dimensions by $s_j$ yields a ranked list that guides the construction of compact dimension subsets $\mathcal{J}$. We use this ranking as a search heuristic in the Qwen3-8B case study, where small subsets are evaluated under the near-optimal criterion in Eq. (2).

### 3.3 KAN distillation in the knob subspace

For RQ3, we focus on a case study of Qwen3-8B and distill the knowledge boundary within low-dimensional knob candidates using a KAN. KAN

is a structured function approximator motivated by the Kolmogorov–Arnold representation, which models multivariate functions as compositions of simple univariate functions. This inductive bias aligns with our goal of extracting a compact and interpretable boundary rule in a small knob subspace.

Using the importance scores $s_j$ defined in the previous subsection, we rank dimensions and consider nested dimension subsets $\mathcal{J}_k = \{j_1, \ldots, j_k\}$ formed by the top-$k$ scores. For each $k$, we restrict the hidden states to $\mathcal{J}_k$ across all layers and learn a compact boundary predictor by combining a learned layer attention with a single-output KAN[1].

To aggregate boundary evidence across depth, we introduce trainable layer weights $w \in \mathbb{R}^L$ and define the attention distribution $\alpha = \mathrm{softmax}(w)$. Given $Z_{\mathcal{J}_k}(H)$, we compute a pooled feature vector $\bar{z} \in \mathbb{R}^k$ by attention-weighted summation:

$$\bar{z} = \sum_{\ell=1}^{L} \alpha_\ell \, Z_{\mathcal{J}_k}(H)_{\ell,:}. \tag{3}$$

A single-output KAN $\Phi_\phi : \mathbb{R}^k \to \mathbb{R}$ then maps $\bar{z}$ to a scalar logit score:

$$s = \Phi_\phi(\bar{z}). \tag{4}$$

Combining Eqs. (3) and (4), we define the distilled boundary predictor

$$g_{\phi,k}(H) = \Phi_\phi\left(\sum_{\ell=1}^{L} \alpha_\ell \, H[\ell, \mathcal{J}_k]\right), \tag{5}$$

which serves as a low-dimensional approximation to the boundary score in Section 2.

We train $g_{\phi,k}$ on labeled pairs $(H(x), y(x))$ with a binary classification objective with logits:

$$\mathcal{L} = \mathbb{E}_x\Big[\mathrm{BCEwithLogits}\big(g_{\phi,k}(H(x)), y(x)\big)\Big]. \tag{6}$$

We evaluate $k$ in increasing order starting from $k = 1$ and stop once the near-optimal criterion in Eq. (2) is satisfied, yielding $\mathcal{J}_{\mathrm{knob}}$ by definition. This procedure produces both the first near-optimal subset $\mathcal{J}_{\mathrm{knob}}$ and an explicit scoring rule $g_{\phi,k}$ (Eq. (5)). For interpretability, we further apply KAN's symbolic distillation routine to convert the learned predictor into an explicit closed-form expression.

[1]We use a minimal one-hidden-layer KAN with two nodes, as this capacity is sufficient for our boundary predictor while yielding a more interpretable model.

In the next step, we use the distilled boundary to design targeted interventions on the selected dimensions and test whether perturbing them directionally shifts the model's preference for the reference answer.

### 3.4 Causal interventions on knobs

We use causal interventions to test whether the distilled knobs are not only predictive but also capable of directionally modulating the model's preference for the reference answer. Our interventions are defined in the attention-pooled knob space $\bar{z}$ used by the KAN predictor, and are implemented as layer-wise patches on the selected coordinates rather than arbitrary perturbations of the full hidden state. Let $\Phi : \mathbb{R}^k \to \mathbb{R}$ denote the symbolic boundary score distilled from the trained KAN, so that the distilled decision depends on $\bar{z}$ through $s = \Phi(\bar{z})$. We define the intervention normal as the normalized gradient direction of $\Phi$ at $\bar{z}$:

$$n(\bar{z}) = \frac{\nabla\Phi(\bar{z})}{\|\nabla\Phi(\bar{z})\| + \varepsilon}. \tag{7}$$

As controls, we also consider a tangent direction $\mathrm{orth}(n(\bar{z}))$, defined as any unit vector orthogonal to $n(\bar{z})$ in knob space, and a random unit direction.

For a step size parameter $\tau$, we define a normalized perturbation in the pooled space as

$$\Delta\bar{z}(\tau, d) = \frac{\tau}{\|\nabla\Phi(\bar{z})\| + \varepsilon}\, d, \quad d \in \{n(\bar{z}), \mathrm{orth}(n(\bar{z})), r\}. \tag{8}$$

We then implement $\Delta\bar{z}$ in the original hidden state space by adding a small patch to the last prompt token, restricted to the selected dimensions $\mathcal{J}_k$ across all layers. Concretely, we construct a layer-wise perturbation $\Delta H[\ell, \mathcal{J}_k] \in \mathbb{R}^k$ whose attention-weighted sum matches the desired pooled shift:

$$\sum_{\ell=1}^{L} \alpha_\ell \, \Delta H[\ell, \mathcal{J}_k] \;=\; \Delta\bar{z}. \tag{9}$$

A simple closed-form solution is obtained by distributing $\Delta\bar{z}$ across layers proportionally to $\alpha_\ell$:

$$\Delta H[\ell, \mathcal{J}_k] = \frac{\alpha_\ell}{\sum_{m=1}^{L} \alpha_m^2} \Delta\bar{z}. \tag{10}$$

Eq. (10) satisfies Eq. (9) by the derivation in Appendix A. To connect internal perturbations to answer likelihood, let $y^\star = (y_1^\star, \ldots, y_T^\star)$ denote a dataset-provided reference answer.

Under teacher forcing, we measure the intervention-induced change in its full-sequence log-likelihood:

$$\Delta \log p(\tau, d) = \sum_{t=1}^{T} \Big[ \log p_\theta\Big(y_t^\star \mid x, y_{<t}^\star; \mathrm{do}\big(H \leftarrow H + \Delta H(\tau, d)\big)\Big) - \log p_\theta\big(y_t^\star \mid x, y_{<t}^\star\big)\Big]. \tag{11}$$

Each reference token is conditioned on the ground-truth prefix $y_{<t}^\star$. The first-reference-token diagnostic corresponds to the special case $T = 1$. We analyze the resulting response patterns across directions and step sizes in Section 5. If $\Phi$ reflects the model's internal boundary in knob space, we expect normal-direction interventions to induce the largest (and roughly monotonic in $\tau$) shifts in $\Delta \log p$, while tangent and random directions yield substantially weaker and less consistent effects.

## 4 Experimental Setup

**Datasets, labeling, and splits.** We primarily use PopQA and additionally evaluate on TriviaQA for RQ1 and RQ2, which play complementary roles. PopQA's long-tailed distribution yields many natural errors, providing enough incorrect cases to learn a clean boundary and evaluate knob-based interventions without synthetic negatives. TriviaQA, from a different question source, serves as a cross-dataset validation setting to test whether the signals found on PopQA are dataset-specific. For each question, we prompt the model to answer with the entity name only and use greedy decoding. We label correctness by matching the generated answer against the reference set (gold entity plus aliases) via normalized exact match with whole-word matching. For each dataset-model pair, we use an 80/20 stratified split and downsample the majority class for balanced 1:1 sets (Table 1)[2]. This abundance of incorrect responses is useful for our RQ3 case study on knob distillation and causal interventions.

[2]The PopQA splits differ across models because its long-tail knowledge makes correct answers scarce. We collect all correctly answered questions for each model and build a balanced 1:1 split from that set, so models with different success patterns yield different sample sizes. In contrast, TriviaQA is much easier, and all models generate many correct answers. Since the correct pool is large and similar across models, we uniformly sample 1,800 correct and 1,800 incorrect examples for each, producing identical train/test sizes.

| Dataset | Model | $N$ | $\text{Train}_{\text{bal}}$ | $\text{Test}_{\text{bal}}$ |
|---|---|---|---|---|
| PopQA | Qwen3-8B | 4932 | 3946 | 986 |
| PopQA | Llama-3.1-8B | 7114 | 5694 | 1420 |
| PopQA | Mistral-7B | 6226 | 4978 | 1248 |
| TriviaQA | Qwen3-8B | 3600 | 2880 | 720 |
| TriviaQA | Llama-3.1-8B | 3600 | 2880 | 720 |
| TriviaQA | Mistral-7B | 3600 | 2880 | 720 |

Table 1: Sample counts for each dataset–model pair after constructing balanced (bal.) 1:1 splits. $N$ is the total number of examples retained after balancing.

**Models and hidden states.** For RQ1 and RQ2, we study Qwen3-8B (Qwen Team, 2025), Llama-3.1-8B (Grattafiori et al., 2024), and Mistral-7B (Jiang et al., 2023); for RQ3, we focus on Qwen3-8B on PopQA as a focused mechanistic case study, since this setting provides sufficient incorrect cases and stable support for layer-wise activation patching. We additionally replicate the full RQ3 pipeline on Llama-3.1-8B/PopQA in Appendix G, independently rediscovering the knob subspace and intervention directions in its native hidden space. We further report Qwen3-1.7B and Qwen3-32B results in Appendix E to test whether the observed boundary signal is specific to the 8B scale. We extract the last-token hidden state tensor $H(x)$ (Section 2) and treat the embedding output as layer 0, so $H(x) \in \mathbb{R}^{L \times d}$ with $L$=37 for Qwen3-8B (36 blocks + embedding), $L$=33 for Llama/Mistral (32 blocks + embedding), and $d$=4096 for these three models.

Additional robustness analyses across 25 data-split/initialization combinations and an anchor-position ablation are reported in Appendix H; these results show stable boundary prediction and support our use of the last prompt token.

## 5 Results

We report results for RQ1-RQ3, moving from boundary prediction to structural characterization and directional intervention. RQ1 establishes that factual correctness is decodable from last-token hidden states across datasets and model families. RQ2 shows that this signal is structured rather than diffuse, concentrating in specific depth regions and a small subset of hidden dimensions. RQ3 presents a Qwen3-8B case study where a compact knob subspace supports boundary distillation and directional modulation of reference-answer likelihood.

| | Model | Qwen3 | Llama-3.1 | Mistral | Avg. |
|---|---|---|---|---|---|
| PopQA | LR | 0.8012 | 0.7709 | 0.7752 | 0.7824 |
| | MLP | 0.8114 | 0.8012 | 0.8005 | 0.8044 |
| | CNN | 0.8083 | 0.8027 | 0.8013 | 0.8041 |
| TriviaQA | LR | 0.7333 | 0.6681 | 0.6931 | 0.6982 |
| | MLP | 0.7722 | 0.7014 | 0.7375 | 0.7370 |
| | CNN | 0.7639 | 0.6917 | 0.7264 | 0.7273 |
| Avg. | LR | 0.7673 | 0.7195 | 0.7342 | 0.7403 |
| | MLP | 0.7918 | 0.7513 | 0.7690 | 0.7707 |
| | CNN | 0.7861 | 0.7472 | 0.7639 | 0.7657 |

Table 2: Balanced accuracy for boundary prediction from last-token hidden states. Chance is 0.50.

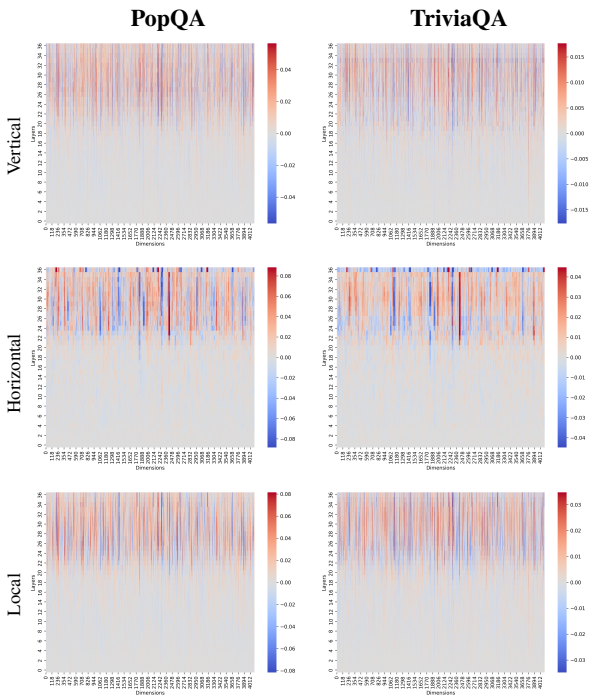
### 5.1 RQ1: The knowledge boundary is learnable from hidden states

Table 2 reports balanced accuracy on class-balanced test sets across two datasets and three model families. All settings achieve strong performance, well above chance (0.50), showing that a hidden-state decision boundary separating correct from incorrect responses is indeed learnable. We include LR and MLP as baselines and use our multi-view CNN as the analysis probe. CNN achieves comparable accuracy while producing spatially aligned activation maps for RQ2. We next localize where boundary evidence concentrates in the layer-dimension grid.

### 5.2 RQ2: Boundary evidence is structured rather than diffuse

RQ2 asks whether the boundary evidence revealed by the CNN probe is structured rather than diffuse. We visualize this evidence by computing signed activation differences $\Delta A = \bar{A}^{(1)} - \bar{A}^{(0)}$ between correct ($y = 1$) and incorrect ($y = 0$) examples, and summarize the maps with simple statistics.[3] For readability, Figures 2 and 3 show the Qwen3-8B results in the main text, while the corresponding Llama-3.1-8B and Mistral-7B results are provided in Appendix D. Together, these results reveal three consistent structural signatures.

**Vertical structure across kernels.** A striking regularity in Figure 2 is that boundary evidence forms vertically aligned stripes. For many dimensions, the activation difference varies more coherently across layers than across neighboring dimensions. This pattern appears across the three




| | PopQA | | | | TriviaQA | | | |
|---|---|---|---|---|---|---|---|---|
| Kernel | V_corr | H_corr | SNR | LateRatio | V_corr | H_corr | SNR | LateRatio |
| Vertical | 0.92 | 0.18 | 21.72 | 0.85 | 0.92 | 0.14 | 22.29 | 0.83 |
| Horizontal | 0.78 | 0.65 | 18.02 | 0.88 | 0.77 | 0.71 | 20.43 | 0.86 |
| Local | 0.93 | 0.71 | 20.94 | 0.86 | 0.93 | 0.59 | 20.79 | 0.85 |

Figure 2: Localization of boundary evidence on Qwen3-8B via CNN activation maps and summary statistics. Rows correspond to the vertical, horizontal, and local CNN branches, while columns correspond to PopQA and TriviaQA. Colors show signed activation differences between correct and incorrect examples. Across both datasets, the maps exhibit strong vertical organization and greater boundary evidence in later layers. The table summarizes vertical continuity, horizontal continuity, dimension-level concentration, and late-layer concentration. Additional model results are provided in Appendix D.

CNN branches: the vertical branch, the horizontal branch, and the local branch all show strong vertical alignment, with $V_{\text{corr}}$ consistently larger than $H_{\text{corr}}$. Overall, correctness evidence appears to be organized primarily along transformer depth, rather than as purely dimension-wise bands.

**Sparsity and long-tailed dimension energy.** The difference maps are dominated by near-zero regions, indicating that most dimensions contribute little to the correct–incorrect separation. This sparsity is also reflected by high SNR values. To further probe dimension-level concentration, Figure 3 plots ranked energy profiles over dimensions and detects knee points. The profiles exhibit a clear long-tailed distribution, with knee points located near the head, indicating that a small subset of dimensions dominates the boundary evidence.

[3] Figure 2 reports four summary metrics. $V_{\text{corr}}$ and $H_{\text{corr}}$ are Pearson correlations of $|\Delta A|$ across adjacent layers and dimensions, respectively. SNR measures dimension-level concentration. LateRatio is the mass of $|\Delta A|$ in the later half of layers.

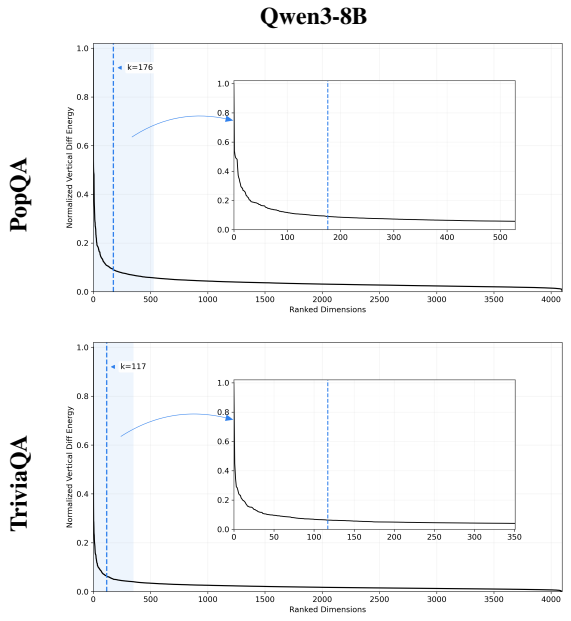



Figure 3: Ranked dimension-energy profiles with knee points on Qwen3-8B. Both datasets show long-tailed profiles, indicating that boundary evidence is concentrated in a small subset of hidden dimensions. Additional model results are provided in Appendix D.

The same qualitative pattern is observed for Llama-3.1-8B and Mistral-7B in Appendix D. This dimension-dominated structure motivates a dimension-based search for compact knob candidates in the Qwen3-8B case study (RQ3).

**Late-layer concentration.** Boundary evidence also concentrates in later layers. In Figure 2, early layers are largely neutral, while later layers exhibit stronger signed differences. Consistently, LateRatio indicates that the later half of layers contains most of the total energy. This depth localization motivates the layer-attention pooling used in RQ3 (Eq. (3)), which learns to upweight informative layers when aggregating knob evidence. Overall, boundary evidence is structured rather than diffuse. We next leverage these signatures to form compact dimension subsets and distill a low-dimensional boundary with directional intervention (RQ3).

### 5.3 RQ3: Knob distillation and directional intervention in a Qwen3-8B case study

RQ3 examines whether the knowledge boundary can be captured by a compact set of controllable hidden dimensions (*knobs*) in a focused case study. We instantiate this study on Qwen3-8B with PopQA. Following the knob definition in Eq. (2), we search for a compact dimension subset

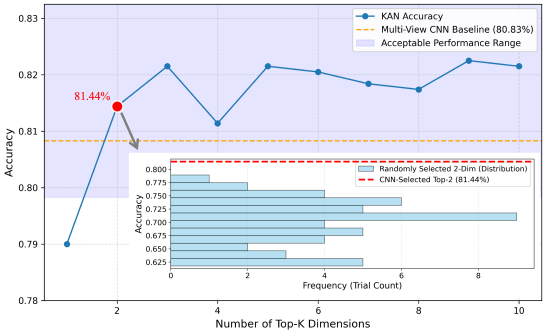



Figure 4: Top-$k$ sweep for compact knob selection on Qwen3-8B/PopQA ($\delta$=1%), with a Random-2 control (inset).

$\mathcal{J}$ whose predictive performance remains within a near-optimal tolerance of the full probe, using $\delta = 1\%$ to define the acceptable range. Concretely, we evaluate nested top-$k$ subsets $\mathcal{J}_k$ (ranked by the RQ2 importance scores) for $k = 1, \ldots, 10$ and identify the first $k$ that satisfies the near-optimal criterion. We validate the specificity of the selected knob set by comparing it against randomly selected subsets of the same size. After identifying a compact knob set, we distill an explicit low-dimensional boundary in the corresponding knob space and visualize its decision geometry. Finally, we test the intervention effect of these knobs by intervening along directions implied by the distilled boundary and measuring directional changes in the model's preference for the reference answer.

**Compactness: the knob set is small and non-trivial.** We first evaluate the *compactness* of the knob set under the near-optimal criterion in Eq. (2). Using the RQ2 importance ranking, we evaluate nested top-$k$ subsets $\mathcal{J}_k$ for $k = 1, \ldots, 10$ on Qwen3-8B/PopQA. Figure 4 shows that $k = 2$ is the first point that enters the acceptable band defined by $\delta = 1\%$, yielding $\mathcal{J}_{\text{knob}} = \mathcal{J}_2$ as the first importance-ranked subset satisfying the near-optimal criterion. To verify that this is not a generic low-dimensional effect, we compare against randomly selected 2D subsets. The Random-2 method with 50 tries is substantially weaker (mean: 0.6996; std.: 0.0426; min: 0.6176; max: 0.7890). None of the random subsets matches the Top-2 accuracy (fraction $\geq$ Top-2: 0.0), indicating that the discovered knob pair is highly specific rather than arbitrary.

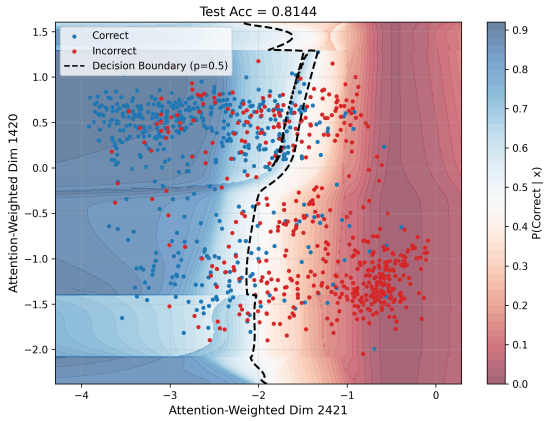



Figure 5: KAN decision surface in the 2D knob space. The x-axis and y-axis show the attention-pooled coordinates of the two selected knob dimensions (Dim 2421 and Dim 1420), respectively. Points are colored by the ground-truth label (CORRECT vs. INCORRECT), and the dashed contour marks the decision boundary at $p(\text{CORRECT} \mid x) = 0.5$.

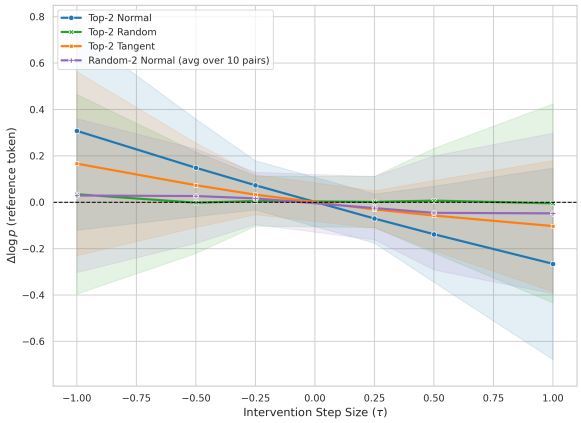



Figure 6: Directional activation patching in the Top-2 knob space, measured by changes in the first-reference-token log-probability: normal/tangent/random, with Random-2 Normal averaged over 10 random 2D pairs. Shaded regions show variability.

**Geometry: a compact boundary emerges in the 2D knob space.** Given $|\mathcal{J}_{\text{knob}}| = 2$, we visualize the induced boundary in the corresponding knob space. Figure 5 plots the KAN decision surface on the attention-pooled coordinates of the two selected dimensions, where color indicates the predicted probability of CORRECT and the 0.5 contour corresponds to the boundary $g_{\phi,2}(H) = 0$. Blue (CORRECT) and red (INCORRECT) examples form separated regions with the boundary passing through their transition, providing geometric evidence that the two-dimensional knob subspace captures a meaningful separation of factual states.

**Directional intervention: knob perturbations modulate reference-answer likelihood.** Predictive sufficiency alone is not sufficient for a knob: a valid knob should also support stable and directional intervention effects. We therefore test whether the discovered Top-2 knob set $\mathcal{J}_{\text{knob}} = \{2421, 1420\}$ of Qwen3-8B supports directional changes in the model's preference for the reference answer. Concretely, we take the KAN from the previous subsection and distill it into a closed-form symbolic boundary score $\Phi$[4]. In our case study, the selected log-based symbolic form achieves strong agreement with the learned KAN logits ($R^2 = 0.8807$) and competitive balanced accuracy (0.8043) on balanced data. Let $x_1$ and $x_2$ denote the attention-pooled coordinates of the two selected dimensions. At a high level, the distilled boundary takes the additive form

$$\Phi(x_1, x_2) = c + \sum_{i=1}^{2} \sum_{m=1}^{M_i} w_{i,m} \log\left(u_{i,m}^{\top}[1, x_i]\right), \tag{12}$$

where $M_i$ denotes the number of log terms over $x_i$ (in our instantiation, $M_1 = M_2 = 2$); $u_{i,m} = [\alpha_{i,m}, \beta_{i,m}]$ collects the affine coefficients inside the log. The instantiated numeric coefficients are omitted for brevity. The full closed-form expression produced by symbolic distillation is provided in Appendix F.

Following the directional patching procedure in Section 3.4, Figure 6 shows that normal-direction interventions induce the largest and most consistent shifts in $\Delta \log p(\tau, d)$ (see Eq. (11)), with an approximately monotonic trend over $\tau$. In contrast, tangent and random controls stay much closer to zero and are substantially less consistent, consistent with the normal direction being maximally aligned with the local boundary gradient.

As a specificity control, we apply the same analysis to 10 random 2D pairs and average their normal-direction responses (Random-2 Normal in Fig. 6). This baseline remains near zero, showing no comparable dominance and indicating that the effect is not generic to arbitrary subspaces.

[4]We sweep a small library of candidate univariate nonlinearities. Since the variants yield comparable fidelity and classification accuracy, we report the log-based form as a compact representative expression; additional variants are shown in Appendix F.

| Setting | Mean | Median | Improved (%) |
|---|---|---|---|
| *Normal-direction strength sweep* | | | |
| $\tau = -0.10$ | $+0.0378$ | $+0.0156$ | 63.9 |
| $\tau = -0.25$ | $+0.0580$ | $+0.0469$ | 70.4 |
| $\tau = -0.50$ | $+0.1609$ | $+0.0938$ | 74.3 |
| $\tau = -1.00$ | $+0.3583$ | $+0.1855$ | 73.5 |
| *Direction comparison at* $\tau = -0.10$ | | | |
| Normal | $+0.0378$ | – | – |
| Tangent | $-0.0096$ | – | – |
| Random | $-0.0023$ | – | – |

Table 3: Teacher-forced full-reference sequence intervention results on Qwen3-8B/PopQA. Mean and median denote changes in full-reference sequence log-likelihood. "Improved" is the percentage of examples whose full-reference sequence log-likelihood increases after intervention.

To verify that the intervention effect extends beyond the first reference token, we additionally evaluate the teacher-forced log-likelihood of the complete reference-answer sequence. Table 3 summarizes the results. Negative normal-direction interventions consistently increase the full-reference likelihood, with larger mean changes at stronger intervention strengths. The improvements at all four strengths are statistically significant under one-sided Wilcoxon signed-rank tests against zero ($p < 3 \times 10^{-26}$). At $\tau = -0.1$, the normal intervention is significantly positive against zero ($p = 2.57 \times 10^{-26}$) and significantly stronger than both the tangent ($p = 2.12 \times 10^{-20}$) and random ($p = 1.81 \times 10^{-26}$) controls under sample-level paired Wilcoxon signed-rank tests. For the random control, the three random-direction results are first averaged within each example before the paired comparison.

Intervention in the opposite direction consistently decreases the full-reference likelihood: as $\tau$ increases from $+0.1$ to $+1.0$, the mean change decreases from $-0.0437$ to $-1.2616$. These results extend the intervention effect from the first reference token to the complete reference-answer sequence under teacher forcing.

Overall, the Top-2 dimensions form a compact, intervention-effective knob subspace whose boundary-guided perturbations directionally modulate reference-answer likelihood.

## 6 Related Work

Hallucination detection commonly uses only model outputs, treating an LLM as a black box (Manakul et al., 2023; Zhang et al., 2023). Representative methods rely on behavioral consistency (Li et al., 2024; Manakul et al., 2023) and on uncertainty estimates from sampled generations (Farquhar et al., 2024). While practical, these black-box detectors provide limited insight into which internal representations implement factual decisions. Moreover, self-evaluation signals can be unreliable when the model is systematically biased or confidently wrong, motivating hidden-state based analysis (Geng et al., 2024). Beyond outputs, recent work detects hallucinations from internal signals, e.g., hidden states and attention patterns, using lightweight probes or simple derived features (Chen et al., 2024; Chuang et al., 2024; Zhang et al., 2025a,c). However, most internal-signal detectors remain predictive, producing opaque scores rather than explicit decision structures, motivating boundary-centric analyses that make the decision surface inspectable.

Closely related work studies knowledge-boundary perception and truthfulness through internal representations. Ni et al. (2025) exploit LLM internal states before and after generation to estimate model confidence and improve knowledge-boundary perception. Burns et al. (2023) study latent knowledge in language models without supervision, while Li et al. (2023) intervene on internal representations at inference time to elicit more truthful answers. Huang et al. (2026) further investigate hidden-state boundaries under conversational pressure, characterizing how internal belief-related representations evolve across interaction turns. In contrast, our work explicitly models factual correctness as a decision boundary in hidden-state space, localizes a compact set of influential dimensions, distills the resulting boundary geometry, and evaluates boundary-guided interventions on reference-answer likelihood.

## 7 Conclusion

We identify an interpretable boundary in transformer hidden states separating correct from incorrect factual answers. Across three LLMs, this boundary is decodable from last-token representations, organized across layers, and concentrated in a few influential dimensions. With Qwen3-8B, we compress the boundary into a low-dimensional knob space, distill it into symbolic scores, and show that boundary-guided interventions directionally modulate the teacher-forced likelihood of the full reference answer.

## Limitations

Our study focuses on short-answer factual QA, where correctness can be evaluated cleanly. The intervention analysis measures the teacher-forced likelihood of the reference answer rather than free-generation accuracy; therefore, it does not establish that an intervention flips an incorrect generated answer to a correct one. Extending the framework to open-ended generation, multi-step reasoning, and long-context dialogue remains future work.

We primarily analyze hidden states at the last prompt token immediately before generation. Although an anchor-position ablation supports this choice, it may miss mechanisms that emerge during multi-token decoding, such as self-correction, drift, or error accumulation. Future work can track boundary geometry and intervention directions across decoding steps.

Our detailed knob identification and symbolic distillation center on Qwen3-8B/PopQA, with an independent Llama-3.1-8B replication showing the same functional pattern in a model-specific subspace. Thus, our intended generalization is at the level of the analysis procedure and qualitative properties, rather than shared dimension indices or symbolic coefficients. Broader validation across prompts, datasets, and model families remains necessary.

## Acknowledgments

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL).

## References

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

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

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

Zihao Huang, Rui Mao, Pengjun Xie, Anupam Chattopadhyay, and Erik Cambria. 2026. How conversational pressure alters belief in large language models: A hidden state boundary perspective. In *Third Conference on Language Modeling*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. 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)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11858–11875, Miami, Florida, USA. Association for Computational Linguistics.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y. Hou, and Max Tegmark. 2025. KAN: Kolmogorov–arnold networks. In *The Thirteenth International Conference on Learning Representations*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. 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*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. GPTEval: A survey on assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7844–7866, Torino, Italia. ELRA and ICCL.

Rui Mao, Zihao Huang, Xulang Zhang, and Erik Cambria. 2026. Neurosymbolic AI: fundamental learning for context engineering in computational linguistics. *npj Artificial Intelligence*.

Shiyu Ni, Keping Bi, Jiafeng Guo, Lulu Yu, Baolong Bi, and Xueqi Cheng. 2025. Towards fully exploiting LLM internal states to enhance knowledge boundary perception. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24315–24329, Vienna, Austria. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023. SAC$^3$: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458, Singapore. Association for Computational Linguistics.

Luan Zhang, Dandan Song, Zhijing Wu, Yuhang Tian, Changzhi Zhou, Jing Xu, Ziyi Yang, and Shuhao Zhang. 2025a. Detecting hallucination in large language models through deep internal representation analysis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 8357–8365. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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. 2025b. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. 2025c. 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)*, pages 17986–18002, Vienna, Austria. Association for Computational Linguistics.

Xianxun Zhu, Erik Cambria, and Hui Chen. 2026. Uncertainty-aware multimodal affective data fusion for personalized mental health dialogue. *Pattern Recognition*, 180:114574.

## A Derivation of the Layer-wise Patch

We derive the closed-form patch in Eq. (10) as the minimum-norm solution to the attention-weighted constraint in Eq. (9). Let $\Delta H_\ell \in \mathbb{R}^k$ denote the layer-wise perturbation restricted to the selected coordinates, i.e., $\Delta H_\ell = \Delta H[\ell, \mathcal{J}_k]$. We solve

$$\min_{\{\Delta H_\ell\}_{\ell=1}^L} \sum_{\ell=1}^{L} \|\Delta H_\ell\|_2^2 \quad \text{s.t.} \quad \sum_{\ell=1}^{L} \alpha_\ell \Delta H_\ell = \Delta \bar{z}. \tag{13}$$

The Lagrangian is

$$\mathcal{L} = \sum_{\ell=1}^{L} \|\Delta H_\ell\|_2^2 + \lambda^\top \left( \sum_{\ell=1}^{L} \alpha_\ell \Delta H_\ell - \Delta \bar{z} \right), \tag{14}$$

where $\lambda \in \mathbb{R}^k$ is a Lagrange multiplier. Setting $\partial \mathcal{L} / \partial \Delta H_\ell = 0$ gives

$$2\Delta H_\ell + \alpha_\ell \lambda = 0 \quad \Rightarrow \quad \Delta H_\ell = -\frac{\alpha_\ell}{2} \lambda. \tag{15}$$

Substituting into the constraint yields

$$\begin{aligned} -\frac{1}{2} \left( \sum_{\ell=1}^{L} \alpha_\ell^2 \right) \lambda &= \Delta \bar{z} \\ \Rightarrow \quad \lambda &= -\frac{2}{\sum_{m=1}^{L} \alpha_m^2} \Delta \bar{z}. \end{aligned} \tag{16}$$

Therefore,

$$\Delta H_\ell = \frac{\alpha_\ell}{\sum_{m=1}^{L} \alpha_m^2} \Delta \bar{z}, \tag{17}$$

which is Eq. (10). In implementation, we add a small $\epsilon$ to the denominator for numerical stability.

## B RQ2 Summary Metrics

Given a signed activation-difference map $\Delta A \in \mathbb{R}^{L \times d}$, we compute all summary metrics on its magnitude $M = |\Delta A|$. Let $M_{\ell,j}$ denote the entry at layer $\ell \in \{1, \ldots, L\}$ and dimension $j \in \{1, \ldots, d\}$.

**Vertical and horizontal smoothness ($V_{\text{corr}}$, $H_{\text{corr}}$).** We quantify whether evidence varies smoothly along depth (vertical) or along the hidden dimension axis (horizontal) using Pearson correlations between adjacent slices. Define the adjacent-layer pairs $M^{\uparrow} = M_{1:L-1,:}$ and $M^{\downarrow} = M_{2:L,:}$, and the adjacent-dimension pairs $M^{\leftarrow} = M_{:,1:d-1}$ and $M^{\rightarrow} = M_{:,2:d}$. We flatten each pair into vectors and compute

$$V_{\text{corr}} = \text{corr}\Big(\text{vec}(M^{\uparrow}),\ \text{vec}(M^{\downarrow})\Big), \tag{18}$$

$$H_{\text{corr}} = \text{corr}(\text{vec}(M^{\leftarrow}),\ \text{vec}(M^{\rightarrow})), \tag{19}$$

where $\text{corr}(\cdot, \cdot)$ denotes Pearson correlation and $\text{vec}(\cdot)$ flattens a matrix into a vector. Intuitively, larger $V_{\text{corr}}$ indicates stronger layer-aligned (vertical) continuity, while larger $H_{\text{corr}}$ indicates stronger dimension-aligned (horizontal) continuity.

**Dimension-level concentration (SNR).** We summarize how concentrated the evidence is over hidden dimensions by first computing a per-dimension energy vector

$$e_j = \frac{1}{L} \sum_{\ell=1}^{L} M_{\ell,j}, \qquad j \in \{1, \ldots, d\}. \tag{20}$$

Let $P$ be a percentile threshold (we use $P = 95$). Let $\tau_P$ be the $P$-th percentile of $\{e_j\}_{j=1}^d$. We define the *signal* set $S = \{j : e_j > \tau_P\}$ (top $(100 - P)\%$ dimensions) and the *noise* set $N = \{j : e_j \leq \tau_P\}$. We report an SNR-style ratio in decibels:

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \left( \frac{\frac{1}{|S|} \sum_{j \in S} e_j^2}{\text{Var}\big(\{e_j\}_{j \in N}\big) + \varepsilon} + \varepsilon \right), \tag{21}$$

where $\varepsilon$ is a small constant for numerical stability.

**Late-layer concentration (LateRatio).** To quantify whether evidence concentrates in later layers, we compute the fraction of total magnitude mass contained in the later $\rho$ fraction of layers (we use $\rho = \frac{1}{2}$). Let $\ell_0 = \lfloor (1 - \rho) L \rfloor + 1$ be the starting index of the later segment. We define

$$\text{LateRatio} = \frac{\sum_{\ell=\ell_0}^{L} \sum_{j=1}^{d} M_{\ell,j}}{\sum_{\ell=1}^{L} \sum_{j=1}^{d} M_{\ell,j} + \varepsilon}. \tag{22}$$

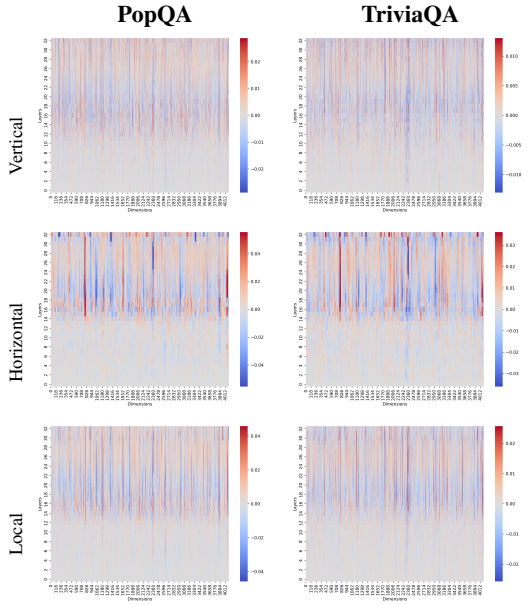
## C Kernel-Size Ablation for the CNN Probe

To test whether the RQ2 structural patterns are artifacts of a specific CNN kernel choice, we evaluate several variants of the vertical, horizontal, and local convolutional kernels on Qwen3-8B/PopQA. Table 4 shows that the probe accuracy remains stable across reasonable kernel-size choices. This suggests that the observed boundary localization patterns are not driven by a single hand-picked kernel configuration.

## D Additional RQ2 Visualizations

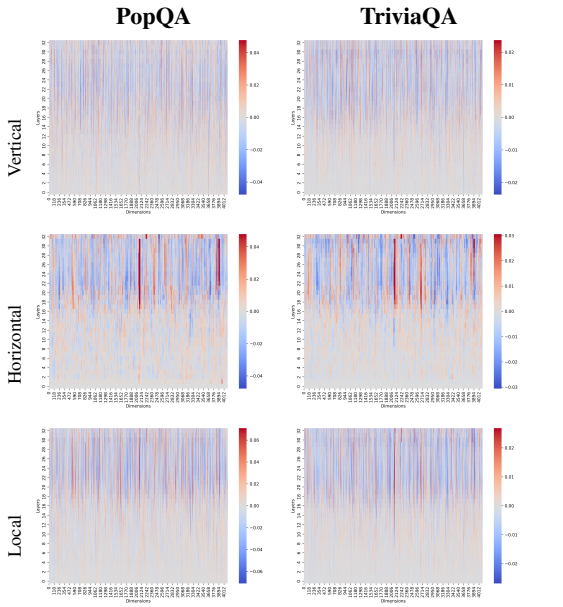
| Setting | Vertical | Horizontal | Local | Accuracy (%) |
|---|---|---|---|---|
| Base | $(9, 1)$ | $(1, 33)$ | $(5, 5)$ | 80.83 |
| Variant 1 | $(9, 1)$ | $(1, 17)$ | $(5, 5)$ | 80.22 |
| Variant 2 | $(9, 1)$ | $(1, 65)$ | $(5, 5)$ | 81.74 |
| Variant 3 | $(11, 1)$ | $(1, 33)$ | $(5, 5)$ | 80.73 |
| Variant 4 | $(7, 1)$ | $(1, 33)$ | $(5, 5)$ | 81.64 |
| Variant 5 | $(9, 1)$ | $(1, 33)$ | $(7, 7)$ | 80.63 |
| Variant 6 | $(9, 1)$ | $(1, 33)$ | $(3, 3)$ | 81.14 |

Table 4: Kernel-size ablation for the CNN probe on Qwen3-8B/PopQA. Accuracy remains stable across reasonable variants of the vertical, horizontal, and local kernels, indicating that the RQ2 patterns are not artifacts of one specific kernel choice.




| | PopQA | | | | TriviaQA | | | |
|---|---|---|---|---|---|---|---|---|
| Kernel | V_corr | H_corr | SNR | LateRatio | V_corr | H_corr | SNR | LateRatio |
| Vertical | 0.83 | 0.09 | 23.02 | 0.66 | 0.84 | 0.12 | 22.12 | 0.71 |
| Horizontal | 0.76 | 0.58 | 19.41 | 0.76 | 0.73 | 0.47 | 18.24 | 0.76 |
| Local | 0.85 | 0.52 | 22.49 | 0.71 | 0.85 | 0.48 | 20.22 | 0.73 |

Figure 7: Additional CNN activation maps and summary statistics for Llama-3.1-8B. The qualitative patterns are consistent with Qwen3-8B in Figure 2, including vertical organization, sparse dimension-level evidence, and late-layer concentration.




| | PopQA | | | | TriviaQA | | | |
|---|---|---|---|---|---|---|---|---|
| Kernel | V_corr | H_corr | SNR | LateRatio | V_corr | H_corr | SNR | LateRatio |
| Vertical | 0.91 | 0.06 | 22.15 | 0.68 | 0.92 | 0.08 | 21.86 | 0.72 |
| Horizontal | 0.84 | 0.59 | 20.76 | 0.74 | 0.80 | 0.52 | 19.33 | 0.76 |
| Local | 0.94 | 0.39 | 19.68 | 0.74 | 0.92 | 0.50 | 20.86 | 0.75 |

Figure 8: Additional CNN activation maps and summary statistics for Mistral-7B. The qualitative patterns are consistent with Qwen3-8B in Figure 2, including vertical organization, sparse dimension-level evidence, and late-layer concentration.

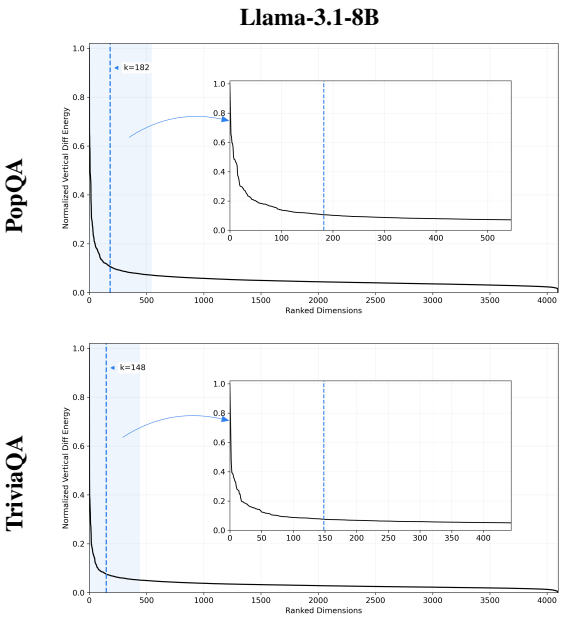



Figure 9: Additional ranked energy profiles with knee points for Llama-3.1-8B. Both PopQA and TriviaQA show long-tailed dimension-energy distributions, consistent with the Qwen3-8B profiles in Figure 3.

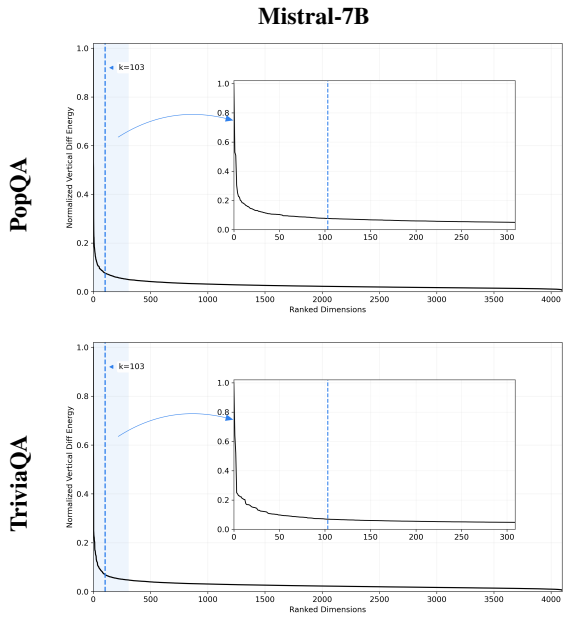



Figure 10: Additional ranked energy profiles with knee points for Mistral-7B. Both PopQA and TriviaQA show long-tailed dimension-energy distributions, consistent with the Qwen3-8B profiles in Figure 3.

## E Cross-Scale Validation on Qwen3

To test whether the observed boundary signal is specific to a single model scale, we additionally evaluate Qwen3-1.7B and Qwen3-32B on the same balanced PopQA setting. Table 5 reports two results: CNN prediction using full hidden states, and KAN prediction using only the Top-2 dimensions selected by the same structural procedure. Both additional scales retain strong boundary detectability, and the compact Top-2 KAN predictor remains substantially above chance. These results suggest that the qualitative boundary signal is not limited to Qwen3-8B.

| Model | CNN full hidden states | KAN Top-2 |
|---|---|---|
| Qwen3-1.7B | 92.02 | 83.27 |
| Qwen3-32B | 84.90 | 78.44 |

Table 5: Cross-scale validation on Qwen3 models. Accuracy is reported on balanced PopQA subsets. Both full hidden-state boundary detection and compact Top-2 KAN prediction remain effective at additional model scales.

## F Symbolic Distillation Details

**KAN structure and where nonlinearity is introduced.** Figure 11 visualizes the distilled KAN structure in the Top-2 knob space. The first layer contains four edges that map the raw knobs $(x_1, x_2)$ into intermediate units, while the second layer aggregates them into a single boundary score. This structure suggests that expressive *univariate* nonlinearities are primarily needed on the first-layer edges, whereas the second-layer edges are well approximated by linear mixing.

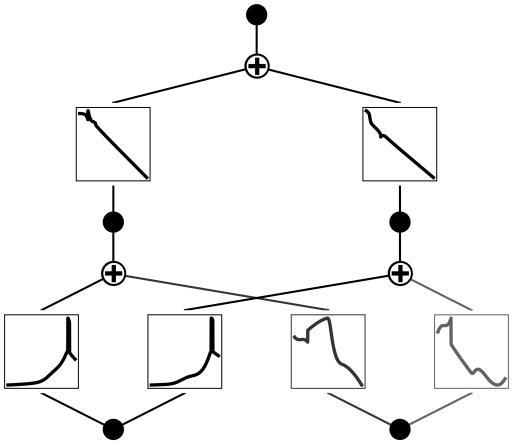
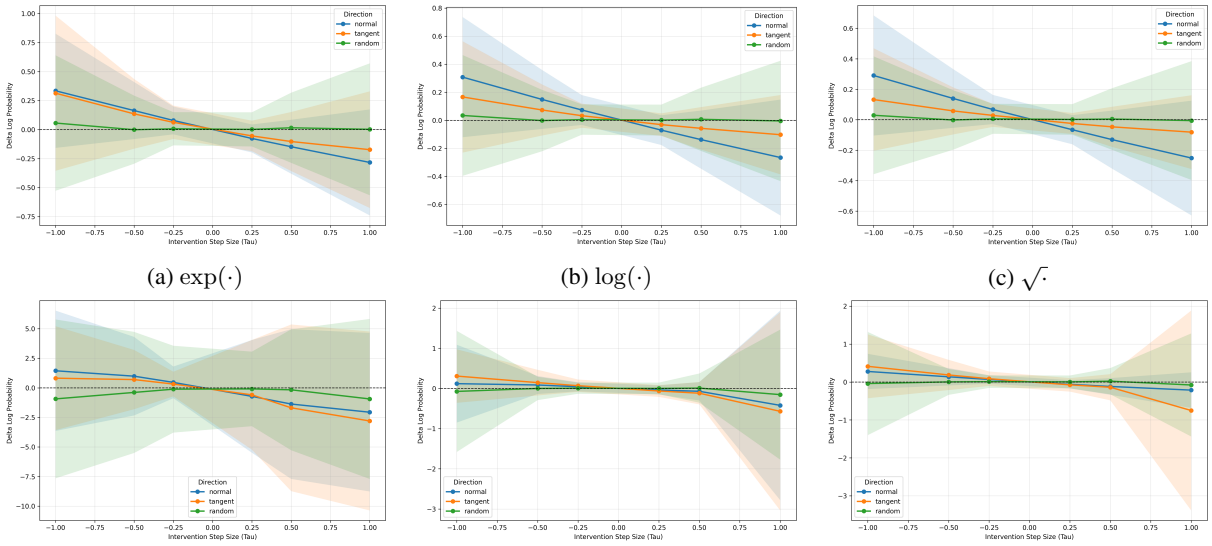


Figure 11: KAN structure distilled in the Top-2 knob space. The first-layer edges capture univariate nonlinear transformations over $x_1$ and $x_2$, and the second layer linearly aggregates these intermediate signals into a boundary score.

| Primitive library | $R^2$ | Symbolic Acc. |
|---|---|---|
| $\exp(\cdot)$ | 0.884687 | 0.806288 |
| $\log(\cdot)$ | 0.880658 | 0.804260 |
| $\sqrt{\cdot}$ | 0.868750 | 0.796146 |
| $\tanh(\cdot)$ | 0.879672 | 0.783976 |
| $x^2$ | 0.885620 | 0.806288 |
| $x^3$ | 0.883424 | 0.799189 |

Table 6: Symbolic distillation across a small library of candidate univariate primitives. Fidelity and accuracy are broadly comparable across libraries, so the log-based form is reported as a compact representative expression. Direction-controlled intervention patterns for all variants are shown in Figure 12.

**Library sweep.** Motivated by the above structure, we sweep a small library of candidate univariate primitives for symbolic distillation: $\{x^2, x^3, \sqrt{x}, \exp(x), \tanh(x), \log(\cdot)\}$. For each library, we distill a closed-form expression $\Phi(x_1, x_2)$ and evaluate (i) its agreement with the KAN logits using $R^2$, and (ii) the balanced test accuracy when using the symbolic score as a classifier. Table 6 shows that multiple libraries yield comparable fidelity and classification performance.




(a) $\exp(\cdot)$ (b) $\log(\cdot)$ (c) $\sqrt{\cdot}$

(d) $\tanh(\cdot)$ (e) $x^2$ (f) $x^3$

Figure 12: Direction-controlled activation intervention results across symbolic variants. Each panel shows the likelihood shift under normal-direction interventions guided by the corresponding symbolic boundary, alongside tangent-direction and random-direction controls. Across variants, normal-direction interventions generally induce stronger and more systematic shifts than tangent and random controls. The log-based variant is reported in the main text as a compact representative expression.

**Symbolic variants and representative closed-form expression.** Since Table 6 shows comparable fidelity and classification accuracy across primitive libraries, we report the log-based expression as a compact representative closed-form boundary. Figure 12 further visualizes direction-controlled intervention patterns across all tested variants. The qualitative separation between normal-direction interventions and tangent/random controls is broadly consistent across variants, indicating that the main conclusion does not rely on a single symbolic primitive. We use the log-based form in the main text because it provides a compact expression with clear directional separation.

**Full closed-form expression for the $\log(\cdot)$ variant.** For completeness, we provide the instantiated closed-form expression distilled from the KAN using the $\log(\cdot)$ library. Let $(x_1, x_2)$ denote the Top-2 knob coordinates. The resulting symbolic boundary score is:

$$\begin{aligned}\Phi(x_1, x_2) = \; & 2.171\,\log\big(0.584 - 3.715\,x_1\big) \\ & + 1.307\,\log\big(2.420 - 9.300\,x_1\big) \\ & - 0.468\,\log\big(8.965 - 5.577\,x_2\big) \\ & + 0.326\,\log\big(9.350 + 3.990\,x_2\big) \\ & - 7.849\,.\end{aligned} \tag{23}$$

## G Cross-Model RQ3 Replication on Llama-3.1-8B

To test whether the RQ3 findings are specific to Qwen3-8B, we independently repeat the full analysis pipeline on Llama-3.1-8B/PopQA in the model's native hidden space, without transferring or reusing any dimensions from Qwen3. The pipeline includes CNN localization, dimension ranking, Top-$k$ search, layer-attention + KAN, and directional intervention.

**Compact knob subspace.** The pipeline independently identifies dimensions 788 and 2978 as the Top-2 subset. Table 7 summarizes the resulting predictive performance. The full-hidden CNN achieves an accuracy of 0.8027, giving a 1% near-optimal threshold of 0.7927. Top-1 does not satisfy this criterion, while Top-2 reaches 0.8123 and is the first subset in the nested Top-$k$ search to do so.

| Representation | Accuracy |
|---|---|
| Full-hidden CNN | 0.8027 |
| Top-1 KAN | 0.5952 |
| Top-2 KAN | **0.8123** |
| Best Top-$k$ result | 0.8185 |

Table 7: RQ3 replication on Llama-3.1-8B/PopQA. The Top-2 subset is the first importance-ranked subset satisfying the 1% near-optimal criterion.

**Directional intervention.** We next evaluate the same boundary-guided intervention using the teacher-forced full-reference sequence log-likelihood. At $\tau = -0.1$, the normal direction increases the full-reference log-likelihood by $+0.1149$, compared with $-0.0739$ for the tangent control and $+0.0033$ for the random control. The normal intervention is significantly positive against zero ($p = 4.33 \times 10^{-35}$) and significantly stronger than both the tangent ($p = 3.52 \times 10^{-48}$) and random ($p = 1.09 \times 10^{-33}$) controls under sample-level paired Wilcoxon signed-rank tests. For the random control, random-direction results are first averaged within each example before the paired comparison.

| Direction | Mean $\Delta$ log-likelihood |
|---|---|
| Normal | **+0.1149** |
| Tangent | -0.0739 |
| Random | +0.0033 |

Table 8: Teacher-forced full-reference sequence intervention results on Llama-3.1-8B/PopQA at $\tau = -0.1$.

Across intervention strengths, the normal direction also shows a clear bidirectional response: the mean full-reference log-likelihood changes by $+0.1149$, $+0.2221$, $+0.2469$, and $+0.3748$ for $\tau = -0.10, -0.25, -0.50, -1.00$, respectively, and by $-0.1391$, $-0.3556$, $-0.6794$, and $-1.3296$ for $\tau = +0.10, +0.25, +0.50, +1.00$.

These results show that the same pipeline can independently recover a compact, intervention-effective knob subspace in another model family. The selected dimensions differ from those identified in Qwen3-8B, so our result supports cross-model functional replication rather than shared dimension indices or a universal coordinate-level mechanism.

## H Robustness and Anchor-Position Ablations

**Robustness across splits and initializations.** To evaluate whether the boundary-prediction and dimension-localization results depend on a particular data split or model initialization, we repeat the Qwen3-8B/PopQA experiment using five data-split seeds and five initialization seeds, resulting in 25 runs. Across these runs, the CNN boundary probe achieves a mean accuracy of $0.8126$ with a standard deviation of $0.0139$; the half-width of the 95% confidence interval is $0.0055$.

We additionally recompute the dimension-importance ranking independently in every run. Dim 2421 appears in the Top-2 in all 25 runs, while Dim 1420 and Dim 1838 appear in the Top-2 in 12 and 6 runs, respectively. Across runs, the mean Top-10 overlap is $0.700$, and the mean Spearman rank correlation is $0.579$. These results indicate that the same core dimension and a similar sparse high-importance structure recur across different data splits and initializations.

| Robustness statistic | Result |
|---|---|
| Mean accuracy | $0.8126$ |
| Std. accuracy | $0.0139$ |
| 95% CI half-width | $0.0055$ |
| Dim 2421 in Top-2 | $25/25$ |
| Dim 1420 in Top-2 | $12/25$ |
| Dim 1838 in Top-2 | $6/25$ |
| Mean Top-10 overlap | $0.700$ |
| Mean Spearman correlation | $0.579$ |

Table 9: Robustness of boundary prediction and dimension ranking across five data-split seeds and five initialization seeds on Qwen3-8B/PopQA.

**Anchor-position ablation.** Our main analysis uses the hidden state of the last prompt token as the pre-generation representation. To test whether this choice is important, we compare it with the second-to-last prompt token and mean pooling over all prompt tokens under the same CNN setting on Qwen3-8B/PopQA. As shown in Table 10, both alternatives yield lower boundary-prediction accuracy than the last prompt token.

| Anchor representation | Accuracy |
|---|---|
| Last prompt token | $\mathbf{0.8083}$ |
| Second-to-last token | $0.7582$ |
| Prompt mean pooling | $0.7633$ |

Table 10: Anchor-position ablation on Qwen3-8B/PopQA using the CNN boundary probe.