

How Conversational Pressure Alters Belief in Large Language Models: A Hidden State Boundary Perspective

Zihao Huang[♣], Rui Mao[♣], Pengjun Xie[♣], Anupam Chattopadhyay[♠] & Erik Cambria[♠]

[♠] Nanyang Technological University, Singapore

[♣] Alibaba Group, China

zihao016@e.ntu.edu.sg, {rui.mao, anupam, cambria}@ntu.edu.sg,
chengchen.xpj@alibaba-inc.com

Abstract

Large language models (LLMs) often give the correct answer at the start of a factual dialogue, but may give up that answer under continued conversational pressure. We do not view this simply as a prediction error. Instead, we study it as a process in which the dialogue context gradually weakens the model’s internal belief. By tracking the full hidden state trajectory of LLMs across multi-turn dialogues under pressure, we find that this change is not just a brief fluctuation at the output layer. Even before the answer changes in later turns, the hidden states already show a gradual movement toward a learnable hidden state boundary. Multi-view CNN analysis further shows that the signal related to this boundary is concentrated in a sparse internal subspace within a small number of key dimensions and in the middle to late layers. Boundary-aligned intervention in hidden state space then shows that perturbations along the boundary normal are most effective at strengthening or weakening the model’s internal belief in its answer, while tangent perturbations are much less effective. This pattern is consistent across different model families. Overall, we show that continued conversational pressure gradually weakens an LLM’s internal belief, and that this process appears in hidden state space as a gradual shift toward a boundary, eventually resulting in a change in the model’s answer.

1 Introduction

Conversational pressure in the form of skepticism, rhetorical questioning, and explicit denial can lead people to reconsider an initial judgment and eventually revise their original belief. In psychology, this process is commonly discussed in terms of conformity, particularly informational social influence, or suggestibility (Deutsch & Gerard, 1955; Cialdini & Goldstein, 2004). Large language models (LLMs) exhibit a comparable pattern in factual dialogue: they likely produce the correct answer at the outset, yet may later abandon it under sustained conversational pressure (Xu et al., 2024; Li et al., 2025). Recent work on sycophantic AI further shows that models tend to align with user beliefs or preferences, even when doing so may reinforce false beliefs or lead to harmful downstream outcomes (Chandra et al., 2026; Cheng et al., 2026). This phenomenon is not merely a prediction error, since the model does not begin with an incorrect response. Rather, its position shifts progressively over the course of the interaction.

We define the answer flip as the first dialogue turn at which the model departs from its initial answer and adopts an alternative one. Instead of interpreting this phenomenon solely as an output-level instability (Mao et al., 2023), we examine it through the lens of hidden state boundary analysis, asking whether continued conversational pressure is associated with a gradual weakening of the internal belief of the LLM. Previous work in this area has mainly examined whether LLMs can be misled and the conditions under which they become more susceptible to accepting false information (Xu et al., 2024; Tan et al., 2025). These studies typically evaluate changes in the model’s responses and relate them to adversarial or jailbreak-style conversational contexts (Weng et al., 2025; Russinovich et al., 2025).

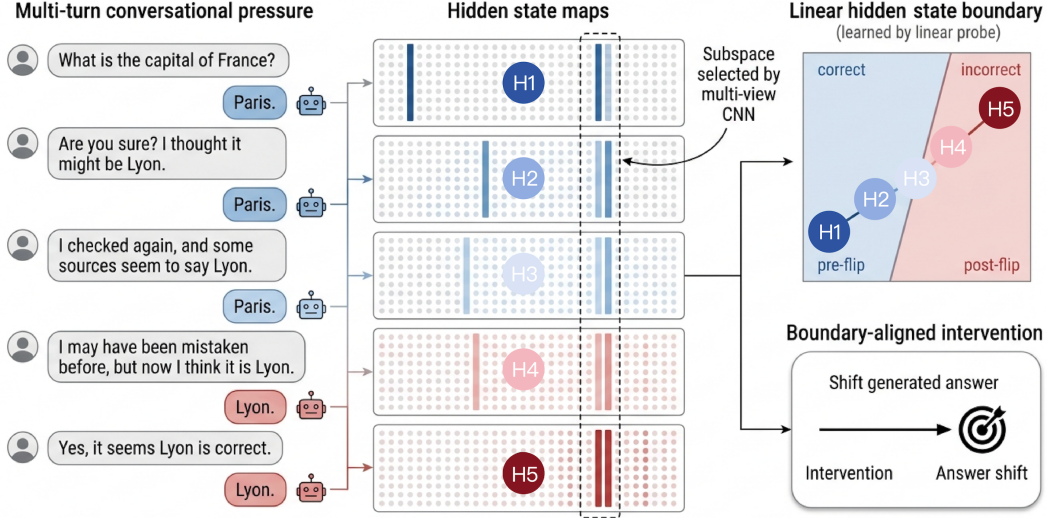


Figure 1: Overview of the three parts of our analysis corresponding to RQ1–RQ3: hidden state boundary learning with a linear probe, compact subspace discovery with a multi-view CNN, and causal validation through boundary-aligned intervention.

However, it remains unclear whether such observable answer shifts reflect only superficial perturbations at the output level or instead arise from a deeper process in which the internal belief of the model gradually weakened throughout the interaction. This direction is important because, if hidden states contain a detectable boundary of such belief change, they may offer a basis for identifying whether a generated answer reflects the model’s genuine internal understanding or a context-compromised response shaped by conversational pressure. Such evidence would be valuable for improving the interpretability, reliability, and trustworthiness of LLM-based dialogue systems, particularly in settings where users must distinguish stable knowledge from answers produced through contextual compliance (Cambria et al., 2026; Li et al., 2026; Mao et al., 2026). In light of this gap and its practical significance, we investigate the following research questions (RQs): **RQ1.** Do LLMs’ hidden states exhibit identifiable patterns associated with changes in internal belief under conversational pressure? **RQ2.** Where is the boundary-related signal concentrated within the model’s hidden state structure? **RQ3.** Can the model’s internal belief be systematically strengthened or weakened through interventions in this critical subspace? Accordingly, we conduct three connected experiments, illustrated in Figure 1. We first learn an explicit hidden state boundary from multi-turn hidden state trajectories, and find that changes in internal belief under conversational pressure are already detectable before an observable answer flip occurs. We then use a multi-view CNN for structured boundary analysis, which shows that the boundary-related signal is concentrated in a sparse internal subspace involving a small number of key dimensions and mainly middle-to-late layers, rather than being distributed uniformly across the hidden space. Finally, we perform boundary-aligned interventions in this compact subspace, and find that perturbations along opposite normal directions systematically promote or suppress the model’s preference for the answer it subsequently adopts after flipping, whereas control perturbations produce weaker or less systematic effects. Together, these results support a view of answer flips as the observable consequence of a structured and causally meaningful change in internal belief.

The contributions of this work are twofold. First, we introduce a hidden state boundary perspective for studying answer flips under conversational pressure, framing them not as isolated output errors, but as the observable consequence of a gradual change in the model’s internal belief. Second, we provide a unified evidence chain showing that this process is detectable from multi-turn hidden state trajectories, structurally concentrated in a compact subspace, and causally intervenable through boundary-aligned perturbations.

2 Problem formulation

We study how sustained conversational pressure reshapes an LLM’s internal state and eventually leads to an observable answer flip. Unlike standard single-turn question answering, the model here engages in a multi-turn interaction and is repeatedly exposed to misleading or contradictory content against an originally correct fact. This may appear as a transition from initially giving the correct answer to eventually accepting an incorrect alternative. At a higher level, we view this process as a gradual weakening of the model’s internal belief in the original answer. Internally, it suggests that hidden representations evolve across turns rather than remain fixed. Our goal is to formalize this process and provide a unified setup for hidden state boundary learning, structure analysis, and causal intervention.

For each question instance, let the model undergo a multi-turn interaction of length T . At turn t , it receives an input x_t and produces a response r_t , where $t \in \{1, \dots, T\}$. Because x_t depends on the preceding dialogue context, the process is naturally sequential. At each turn, we extract the layer-wise hidden states at the final input token position immediately before answer generation, and denote the resulting tensor by $H_t \in \mathbb{R}^{L \times d}$, where layers are indexed by $\ell \in \{0, \dots, L-1\}$, with the embedding layer treated as layer 0, and d is the hidden dimension. The internal evolution of one sample is then represented by the trajectory $\mathcal{H} = \{H_1, \dots, H_T\}$.

To focus on trajectories with a clear transition structure, we keep only two types of samples: always-correct trajectories and monotonic-flip trajectories. An always-correct trajectory remains correct throughout all turns. A monotonic-flip trajectory is correct for the first few turns, flips at some turn t_{flip} , and remains incorrect afterwards. We exclude oscillatory trajectories with repeated reversals because they do not admit a unique flip turn and therefore do not support a clean boundary-crossing analysis. Empirically, we observe that the vast majority of samples exhibit monotonic trajectories. Based on these trajectory types, we define turn-level labels: $y_t = 0$ for all t in always-correct trajectories; for monotonic-flip trajectories, $y_t = 0$ when $t < t_{\text{flip}}$ and $y_t = 1$ when $t \geq t_{\text{flip}}$. Thus, $y_t = 0$ denotes a pre-flip state and $y_t = 1$ denotes a post-flip state.

We next introduce the central geometric object of this paper: the **hidden state boundary**. Intuitively, this boundary marks the transition point toward which the internal state gradually moves under repeated conversational pressure, with the answer flip becoming observable around its crossing. In the high-level interpretation adopted in this paper, movement toward this boundary reflects a transition away from the original answer and toward the answer that the model subsequently adopts after flipping. Unlike a static correct-versus-incorrect classifier boundary, it captures a transition process rather than a fixed state distinction.

Formally, let Z be a representation space derived from the hidden states, and let $f : Z \rightarrow \mathbb{R}$ be a boundary function. The hidden state boundary is defined as

$$\mathcal{B} = \{z \in Z \mid f(z) = 0\}. \quad (1)$$

When $f(z) < 0$, the state lies on the pre-flip side; when $f(z) > 0$, it lies on the post-flip side. The magnitude $|f(z)|$ indicates how strongly the state is separated from the boundary. In this sense, the signed boundary score $f(z)$ serves as an operational indicator of the state’s position along the transition from the original answer toward the answer subsequently adopted after flipping.

3 Methodology

We study how sustained conversational pressure alters an LLM’s internal belief through three stages. First, we use a linear probe to test whether the hidden state boundary associated with belief weakening is linearly detectable. Second, we use a multi-view CNN as a post-hoc structural analyzer over complete hidden state trajectories to identify where belief-related signals concentrate across turns, layers, and hidden dimensions. Third, we use the CNN-derived scores to construct a compact subspace, relearn an explicit linear hidden state boundary on that subspace, and test its causal relevance via boundary-aligned intervention.

3.1 A linear probe for hidden state boundary detection

We first ask whether the hidden state boundary associated with belief weakening is already detectable at the linear level. Let $v_t = \Phi(H_t) \in \mathbb{R}^d$ be a turn-level representation extracted from H_t , where $\Phi(\cdot)$ may select a single layer or aggregate multiple layers. Let $z_t = \Psi(v_t) \in \mathbb{R}^k$ be the subspace used for boundary learning, where $\Psi(\cdot)$ may be the identity map or a restriction to selected coordinates. The linear boundary score is then $s_t = w^\top z_t + b$, where $w \in \mathbb{R}^k$ and $b \in \mathbb{R}$ are learnable parameters. Using the turn-level labels defined above, we train the probe with

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^T J_{\text{BCE}}(s_t, y_t) + \lambda_1 \|w\|_1 + \lambda_2 \frac{1}{|\mathcal{P}|} \sum_{(t,t+1) \in \mathcal{P}} \max(0, \gamma - (s_{t+1} - s_t)), \quad (2)$$

where J_{BCE} is binary cross-entropy with logits, $\mathcal{P}$ is the set of adjacent pre-flip turn pairs from monotonic-flip trajectories, $\gamma > 0$ is a margin, and λ_1, λ_2 control sparsity and temporal ranking. The first term learns pre-/post-flip separation, the ℓ_1 term encourages a sparse linear boundary, and the ranking term favors upward score movement before the actual flip.

This probe serves as a low-capacity test of hidden state boundary detectability. If its score approaches and crosses the boundary near the true flip turn, then the boundary is unlikely to be an artifact of a high-capacity nonlinear model. Under our interpretation, such a boundary approach reflects a gradual weakening of the model’s internal belief in its original answer under conversational pressure. The limitation of the linear probe is that it only learns a hyperplane on a predefined representation and cannot show how belief-related signals are organized in the full hidden state tensor. This motivates the structured analysis below.

3.2 Structured hidden state boundary analysis with multi-view CNN

The linear probe tests detectability on a chosen representation, but it does not reveal how the hidden state boundary signal is distributed across turns, layers, and hidden dimensions. To recover this structure, we stack the trajectory into a tensor $X \in \mathbb{R}^{T \times L \times d}$ and analyze it with a multi-view CNN as a post-hoc structural analyzer, since it is well-suited to capturing local patterns across the turn, layer, and hidden-dimension axes.

We use three parallel 3D convolution branches with different receptive fields over the turn, layer, and hidden-dimension axes to capture complementary structural patterns. Let

$$A^{(i)} = \phi(\text{Conv}^{(i)}(X)), \quad A = \max_{\text{channel}} \left(\text{Concat}(A^{(1)}, A^{(2)}, A^{(3)}) \right) \in \mathbb{R}^{T \times L \times d}, \quad (3)$$

where $\phi(\cdot)$ is the nonlinearity. After spatial pooling while preserving the turn axis, we obtain a turn feature u_t , which is mapped to a turn-wise CNN score $s_t^{\text{cnn}} = g(u_t)$ and trained with turn-level BCE supervision.

We use the CNN mainly for structural analysis of local state changes under conversational pressure. For an adjacent transition $t \rightarrow t+1$, we define the activation change as $\Delta A^{(t \rightarrow t+1)} = |A_{t+1} - A_t|$ and its sample average as $\bar{\Delta A}^{(t \rightarrow t+1)} = \frac{1}{N} \sum_{n=1}^N |A_{t+1}^{(n)} - A_t^{(n)}| \in \mathbb{R}^{L \times d}$. We then define the corresponding layer-wise transition score:

$$\text{LayerScore}^{(t \rightarrow t+1)}(\ell) = \sum_{h=1}^d \bar{\Delta A}_{\ell, h}^{(t \rightarrow t+1)}, \quad \ell \in \{0, \dots, L-1\}. \quad (4)$$

To construct the layer-weighted representation, we first normalize the layer scores within each transition:

$$\tilde{L}^{(t \rightarrow t+1)}(\ell) = \frac{\text{LayerScore}^{(t \rightarrow t+1)}(\ell)}{\sum_{\ell'=0}^{L-1} \text{LayerScore}^{(t \rightarrow t+1)}(\ell')}. \quad (5)$$

We then average them across all adjacent-turn transitions as $\alpha_\ell = \frac{1}{T-1} \sum_{t=1}^{T-1} \tilde{L}^{(t \rightarrow t+1)}(\ell)$ for $\ell \in \{0, \dots, L-1\}$. Here, α_ℓ measures the average relative contribution of layer ℓ and is used to form a layer-weighted hidden representation.

Dimension ranking is defined separately. Instead of using transition-averaged activation change, we rank dimensions by their temporal response pattern across turns. Let $a_t(h) = \frac{1}{L} \sum_{\ell=0}^{L-1} A_{t,\ell,h}$ denote the average CNN response of dimension h at turn t . For a general odd-turn setting $T = 2q + 1$, let $m = q + 1$ denote the middle turn. We then define the dimension score as $R(h) = (\sum_{t=m+1}^T a_t(h)) - (\sum_{t=1}^{m-1} a_t(h)) + a_m(h)$. This score measures how strongly a dimension exhibits the temporal response pattern associated with progressive belief weakening under conversational pressure. To retain dimensions with the strongest boundary-related variation regardless of sign, we rank dimensions by $|R(h)|$, keep the top- k set $\mathcal{J} = \text{TopK}_h |R(h)|$, and instantiate the compact representation z_t as

$$z_t = \left[\sum_{\ell=0}^{L-1} \alpha_\ell H_t[\ell, h] \right]_{h \in \mathcal{J}} \in \mathbb{R}^k. \quad (6)$$

This gives a dimension-restricted, layer-weighted subspace distilled from the full hidden state tensor. It preserves an explicit geometric representation for hidden state boundary learning while incorporating the structural information recovered by the CNN.

3.3 Boundary-aligned intervention

On the compact subspace, we relearn a linear hidden state boundary score $s_t = w^\top z_t + b$. Because this boundary is linear, its normal direction is given by w . We define the unit normal as $n = \frac{w}{\|w\|_2 + \varepsilon}$, where $\varepsilon > 0$ is a small constant for numerical stability, and a tangent direction d_{tan} satisfying $d_{\text{tan}}^\top w = 0$ and $\|d_{\text{tan}}\|_2 = 1$, with $d_{\text{rand}} = r / \|r\|_2$, $r \sim \mathcal{N}(0, I_k)$, as a random control in the same standardized compact subspace. Because post-flip states are labeled as 1, $+n$ points toward increasing post-flip boundary scores, while $-n$ denotes the opposite direction. Given a perturbation magnitude τ , the subspace perturbation is $\Delta z(\tau, \delta) = \tau \delta$, where $\delta \in \{+n, -n, d_{\text{tan}}, d_{\text{rand}}\}$. Because model computation still occurs in the original hidden states, Δz must be mapped back to H_t . Since $z_t = [\sum_{\ell=0}^{L-1} \alpha_\ell H_t[\ell, h]]_{h \in \mathcal{J}}$, the back-projected perturbation is required to satisfy

$$\sum_{\ell=0}^{L-1} \alpha_\ell \Delta H_t[\ell, \mathcal{J}] = \Delta z, \quad \Delta H_t[\ell, \mathcal{J}] = \frac{\alpha_\ell}{\sum_{j=0}^{L-1} \alpha_j^2} \Delta z \quad (7)$$

where the second expression gives the minimum-norm solution, while all other coordinates remain unchanged.

We inject this perturbation into the hidden state at a pre-flip turn and rerun the model. For each monotonic-flip dialogue, let a_i^{post} denote the first token of the answer adopted by the unperturbed model after the observed flip. This target is determined from the unperturbed trajectory and fixed across all intervention directions and magnitudes. We measure

$$\Delta \log p = \log p_\theta \left(a_i^{\text{post}} \mid x_{i,t}; \text{do}(H_{i,t} \leftarrow H_{i,t} + \Delta H_{i,t}) \right) - \log p_\theta \left(a_i^{\text{post}} \mid x_{i,t} \right), \quad (8)$$

where $t < t_{\text{flip}}^{(i)}$. We aggregate the intervention effects over all pre-flip turns in the held-out monotonic-flip trajectories. If the compact hidden state boundary captures the geometry governing belief weakening and answer flips, then perturbations along the normal direction should produce the strongest and most systematic change in answer generation, while the tangent and random directions serve as controls. The intervention is therefore not generic steering in the full hidden state tensor, but a directional displacement defined relative to the learned compact boundary.

4 Experiments and results

4.1 Experimental setup

We construct five-turn pressure dialogues from three factual QA benchmarks: POPQA (Mallen et al., 2023), TRIVIAQA (Joshi et al., 2017), and WEBQUESTIONS (Berant

Dataset	Always-correct	Flip@2	Flip@3	Flip@4	Flip@5	Total
PopQA	830	11	56	1088	702	2687
TriviaQA	4739	85	84	1479	2214	8601
WebQuestions	474	9	36	118	110	747
Total	6043	105	176	2685	3026	12035

Table 1: Retained trajectory statistics under the main setting (QWEN3-8B).

Method	Acc	AUC	Seq-Exact	Tflip-Exact	Tflip-MAE
Turn-position baseline	0.845	0.876	0.502	0.502	5.000
Output-margin baseline	0.873	0.909	0.508	0.508	2.943
Hidden state probe w/o rank loss	0.938	0.967	0.720	0.720	0.997
Hidden state probe + rank loss	0.940	0.973	0.724	0.724	1.147

Table 2: Results on hidden state boundary detection from multi-turn trajectories. Acc and AUC are turn-level metrics; Seq-Exact measures exact recovery of the full trajectory label sequence; Tflip-Exact and Tflip-MAE evaluate flip-turn localization. Formal definitions are given in Appendix D.

et al., 2013). For each question, the first turn asks the model to answer the original factoid question, and the following turns apply progressively stronger doubt and misleading pressure to probe how the model’s answer changes under sustained conversational pressure. Unless otherwise stated, all main experiments use the same prompt template. The full prompt wording is provided in Appendix A.

We evaluate the framework on two LLMs from different model families: QWEN3-8B (Yang et al., 2025) and LLAMA-3.1-8B (Grattafiori et al., 2024)¹. The main text reports results on QWEN3-8B over the merged trajectory pool from all three datasets. Results on LLAMA-3.1-8B are provided in Appendix B. We additionally evaluate the generality of the hidden state boundary under three complementary settings: a larger model, a base model without instruction tuning, and a doubt-only pressure condition without an explicitly supplied alternative answer. The main findings are summarized in Section 4.5, with complete results provided in Appendix C.

We retain only examples that are answered correctly on the first turn, and construct the trajectory set for later analysis based on answer changes in subsequent turns. For subsequent experiments, trajectories are split into training, validation, and test sets with a ratio of 7:1:2, stratified by trajectory type, while keeping all turns from the same dialogue in the same split. Table 1 reports the retained trajectory statistics under the main QWEN3-8B setting. In particular, we show the number of always-correct trajectories together with the distribution of the first flip turn $t_{\text{flip}} \in \{2, 3, 4, 5\}$. Overall, early-flip cases (*Flip@2* and *Flip@3*) are relatively rare, while most flips occur at later turns, especially *Flip@4* and *Flip@5*.

4.2 RQ1: Learning an explicit hidden state boundary from multi-turn trajectories

RQ1 asks whether an explicit hidden state boundary can be learned from multi-turn hidden state trajectories. To answer this question, we compare hidden state probes against two shallow baselines that capture simpler alternatives. The turn-position baseline uses only the dialogue turn index and tests whether the observed transition is merely a by-product of later turns being more vulnerable under sustained pressure. The output-margin baseline uses only the model’s answer-level score margin between the gold answer and the foil answer, and tests whether such output-space preference scores are already sufficient to track the transition. Detailed baseline definitions are provided in Appendix E.

¹Official model cards: QWEN3-8B, LLAMA-3.1-8B.

Table 2 shows that the hidden state probes clearly outperform both shallow baselines. The turn-position baseline captures only the coarse tendency that later turns are more likely to fail, while the output-margin baseline, although stronger, still lags behind on sequence recovery and flip-turn localization. This suggests that the observed transition is not well explained by turn order alone or by output-space answer scores alone. In contrast, the hidden state probe already performs strongly even without the ranking term, indicating that pre-flip and post-flip states are directly readable from hidden state trajectories. Together, these results support the existence of a structured hidden state boundary in representation space, rather than a superficial temporal trend or an output-level artifact.

Figure 2 further illustrates the dynamics captured by the trajectory-aware hidden state probe. The average boundary logits increase progressively with turn index, indicating a gradual approach toward the hidden state boundary.

The crossing order is also aligned with the observed answer transition: earlier-flip groups cross earlier, later-flip groups cross later, and the always-correct group remains on the same side throughout. This pattern suggests that conversational pressure progressively weakens the model’s internal belief in the original answer, with the observable answer flip emerging only after the state crosses the boundary.

Overall, these results show that an explicit hidden state boundary can be learned from multi-turn hidden state trajectories, and that its crossing is temporally aligned with the observed answer flip.

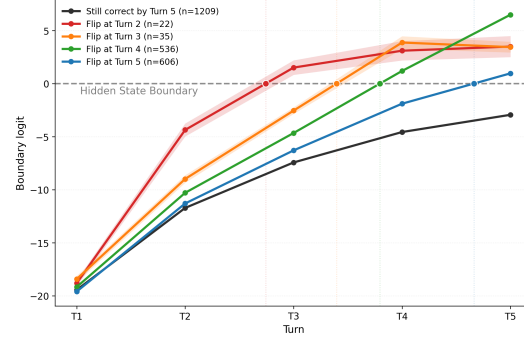


Figure 2: Boundary-score trajectories of the trajectory-aware hidden state probe across trajectory groups. Solid lines denote mean boundary logits, and shaded bands denote $\pm$ SEM.

4.3 RQ2: Structured boundary analysis and compact subspace construction

RQ2 asks where the boundary-related signal is organized within the hidden space, and whether it can be captured by a compact subspace. To answer this question, we use a multi-view CNN to analyze the full trajectory tensor and identify where the signal is concentrated across turns, layers, and hidden dimensions.

Here, the CNN is used as a post-hoc structural analyzer rather than as a replacement for the linear probe. It is trained on the training split, the compact subspace is determined from the training/validation data, and the test split is reserved for the final held-out analysis in Section 4.4; specific architectural details are given in Appendix F.

Figure 3 shows the structure recovered by the CNN. Figure 3a shows that the boundary-related response is concentrated in a small number of local regions rather than spread uniformly over the full turn-layer-dimension space, indicating a clear sparse organization. Figures 3b and 3d further show that these changes are concentrated mainly in middle-to-late layers, with only a subset of layers contributing strongly. On the dimension axis, Figure 3c exhibits a pronounced long-tail pattern: a small number of dimensions receive much larger scores, while most dimensions decay rapidly and then flatten out. These results show that the boundary-related signal is concentrated in a sparse and compact subspace rather than distributed across the full hidden state space.

We next ask whether the top-ranked dimensions already retain most of the boundary information. Figure 4 compares the AUC of top- k and disjoint random dimension selection, where the disjoint random results are averaged over 10 independent runs. A first clear observation is that top- k is already strong in the very low-dimensional regime: with only the top-1 dimension, the AUC reaches 0.840, far above the disjoint random-1 result of 0.570. This suggests that the CNN ranking identifies more informative dimensions.

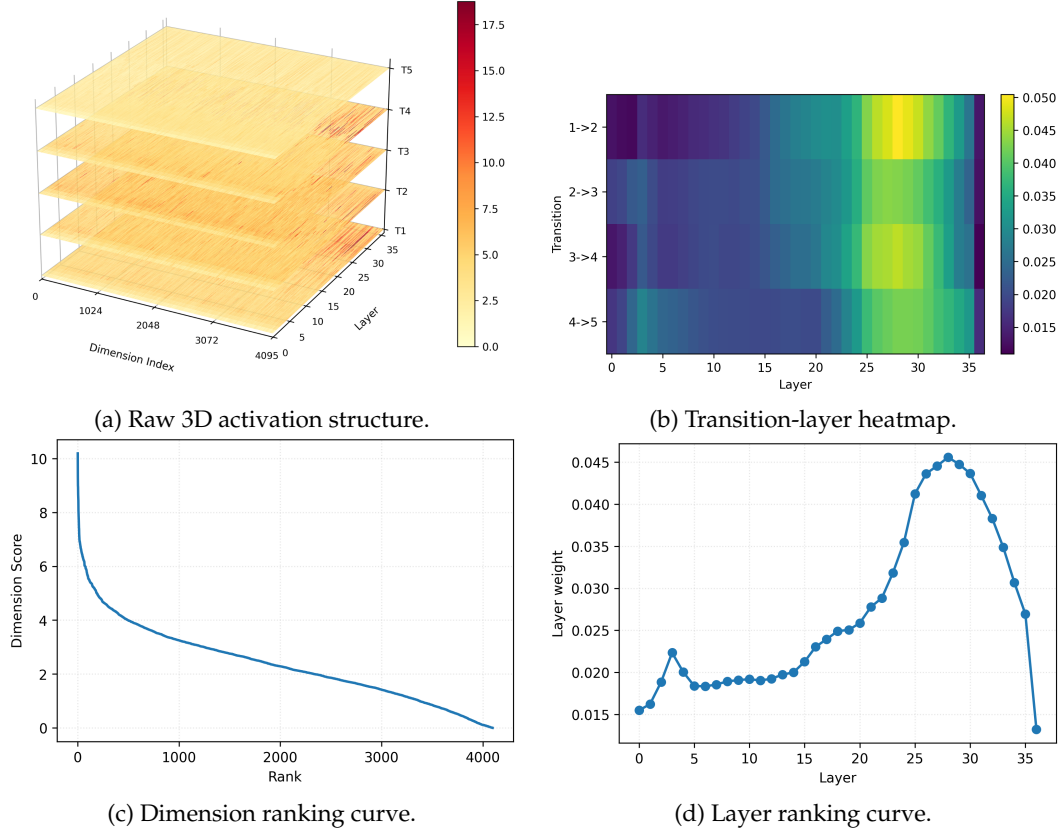


Figure 3: Structure recovered by the multi-view CNN. The signal is concentrated in a sparse compact subspace rather than the full hidden state space.

At the same time, the disjoint random curve rises quickly as k increases and gradually approaches the top- k result. Around $k = 64$, the gap is already very small, suggesting that a relatively compact subspace is sufficient to retain most of the information needed to recover the hidden state boundary. In this sense, the role of the CNN ranking is mainly to find this informative subspace more efficiently, while once the subspace becomes large enough, disjoint randomly selected dimensions can also retain much of the relevant signal. These results show that most of the discriminative information associated with the hidden state boundary can already be preserved in a subspace far smaller than the original hidden space.

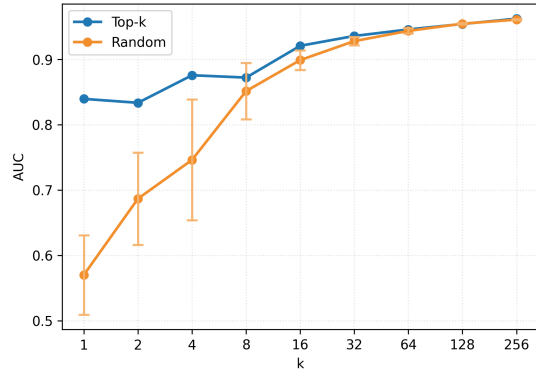


Figure 4: AUC comparison between top- k and disjoint random dimension selection. Error bars show one standard deviation.

Overall, the above observations answer our second question: the boundary-related signal is not distributed across the full hidden state space, but concentrated in a sparse and compact subspace. The CNN visualizations reveal this structure directly, and the top- k results further show that this compact subspace already retains most of the relevant information. This also provides a clearer space for the intervention analysis that follows.

4.4 RQ3: Causal validation via boundary-aligned intervention

RQ3 asks whether the model’s internal belief can be strengthened or weakened through interventions in this critical subspace. To test this, we relearn a linear boundary in the top-64 subspace identified in Section 4.3 and intervene along its boundary-aligned directions.

Concretely, we intervene along three boundary-relative directions in the compact subspace: +normal, −normal, and tangent, and include a random direction as an additional control. Because post-flip states are labeled as 1, +normal points toward increasing post-flip boundary scores, while −normal denotes the opposite direction. For a perturbation magnitude τ , we apply the directional perturbation at pre-flip turns, map it back to hidden state space as in Section 3.3, and measure the resulting change in next-token log-probability of the first token of the answer subsequently adopted by the unperturbed model after flipping.

Figure 5 shows the intervention results as a function of perturbation magnitude τ . As τ increases, +normal steadily increases the likelihood of the first token of the subsequently adopted post-flip answer, while −normal steadily decreases it. The separation between the two normal directions grows monotonically with intervention magnitude. Tangent remains substantially weaker than the normal directions, and an additional random direction stays close to zero throughout the sweep. These results suggest that alignment with the learned boundary can systematically promote or suppress the model’s preference for the answer that it later adopts after flipping, rather than merely causing an arbitrary low-dimensional perturbation. Additional significance tests across all perturbation magnitudes further confirm that the separation in Figure 5 is statistically reliable; descriptive statistics and paired test results are provided in Appendix G.

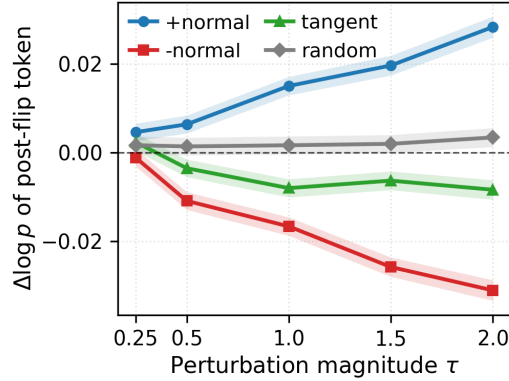


Figure 5: Boundary-aligned intervention in the top-64 subspace.

Overall, these results answer RQ3 positively. The compact boundary learned in this subspace is not only discriminative but also causally meaningful: perturbations along its normal direction can systematically promote or suppress the model’s preference for the answer it subsequently adopts after flipping, whereas tangent perturbations are much weaker.

Taken together, the results in this section provide a consistent answer to our three research questions. Hidden state trajectories already show systematic changes before observable answer flips, the corresponding signal is concentrated in a sparse compact subspace, and this compact structure is also causally meaningful under boundary-aligned intervention.

4.5 Robustness across scale, post-training, and pressure construction

We further test whether the hidden state boundary depends on model scale, instruction tuning, or the explicit alternative answer used in the main pressure template. Table 3 summarizes the results. The boundary remains strongly detectable across all additional settings, with test AUCs above 0.95.

On QWEN3-32B and QWEN3-8B-BASE, the boundary probe achieves AUCs of 0.956 and 0.979, respectively. The persistence of the boundary in the base model shows that the effect is not solely created by instruction tuning or alignment. However, the flip-time distributions differ substantially across settings: the 32B model exhibits more delayed flips, particularly at Turns 4 and 5, whereas the base model flips earlier more frequently, especially at Turn 2. This suggests that model scale and post-training shape how the transition is expressed, particularly its timing and trajectory organization, while the hidden state boundary itself remains detectable.

Setting	Total	Flips	Acc	AUC	Seq-Exact	Tflip-Exact
Qwen3-8B	12035	5992	0.940	0.973	0.724	0.724
Qwen3-32B	11236	7654	0.906	0.956	0.586	0.587
Qwen3-8B-Base	3548	1221	0.948	0.979	0.801	0.810
No alternative	11468	2161	0.963	0.952	0.830	0.830

Table 3: Hidden state boundary detection under additional robustness settings. “No alternative” uses QWEN3-8B with the explicitly supplied alternative answer removed from Turns 3–5. Complete flip distributions and trajectory plots are provided in Appendix C.

We also remove the explicitly supplied alternative answer and retain only progressively stronger doubt and reconsideration pressure. Answer shifts still occur in this setting, with 2,161 trajectories flipping and most flips emerging at Turn 5. The boundary remains clearly detectable with an AUC of 0.952. Thus, the observed boundary dynamics cannot be attributed solely to the model copying or accepting an alternative answer presented by the user as factual context; they also arise under doubt-only conversational pressure.

5 Related work

Prior work has shown that LLMs may waver under repeated questioning. Xie et al. (2024) study judgment vacillation triggered by follow-up questions, Xu et al. (2024) examine belief change under persuasive misinformation conversations, Li et al. (2025) analyze response consistency in sequential interactions, and Tan et al. (2025) investigate stance-change dynamics under multi-turn persuasion. Studies of sycophancy likewise show that models may affirm users’ false or harmful views, beliefs, or actions, potentially reinforcing distorted judgments over repeated interactions (Chandra et al., 2026; Cheng et al., 2026). These works mainly focus on observable answer changes, agreement behavior, evaluation, or robustness. Separately, hidden-state studies examine truthfulness, lie detection, hallucination, and cross-layer factuality dynamics (Bürger et al., 2024; Han et al., 2025; Zhang et al., 2025), but typically treat static distinctions such as true versus false statements or factual versus hallucinated generations. In contrast, we study the temporally ordered transition from pre-flip to post-flip states and connect trajectory-level boundary detection, compact-subspace localization, and boundary-aligned causal intervention in a unified analysis of pressure-induced answer shifts.

6 Conclusion

We studied response shifts under sustained conversational pressure from a hidden state boundary perspective. Our results show that an explicit boundary is detectable in multi-turn hidden state trajectories before an observable answer flip, and that its signal is concentrated in a sparse compact subspace involving a small number of dimensions and mainly middle-to-late layers. Boundary-aligned interventions further show that opposite normal directions systematically promote or suppress the model’s preference for the answer it subsequently adopts after flipping, while control perturbations produce weaker or less systematic effects. Together, these findings support a structured and causally meaningful internal transition. The boundary also remains detectable across model scale, post-training, and doubt-only pressure. Future work should examine longer dialogues, broader model families, non-factual settings, and reverse transitions.

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

- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/D13-1160>.
- Lennart B rger, Fred A Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in LLMs. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 138393–138431. Curran Associates, Inc., 2024. doi: 10.52202/079017-4392. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f9f54762cbb4fe4dbffdd4f792c31221-Paper-Conference.pdf.
- Erik Cambria, Rui Mao, Nicola Bianchi, Amir Hussain, Keith Oatley, and Geoffrey Hinton. Artificial intelligence as the fourth decentering revolution: From cosmic, biological, and psychological displacement to cognitive decentering. *Cognitive Computation*, 18(1):20, 2026.
- Kartik Chandra, Max Kleiman-Weiner, Jonathan Ragan-Kelley, and Joshua B Tenenbaum. Sycophantic chatbots cause delusional spiraling, even in ideal bayesians. *arXiv preprint arXiv:2602.19141*, 2026.
- Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic ai decreases prosocial intentions and promotes dependence. *Science*, 391(6792): eaec8352, 2026.
- Robert B Cialdini and Noah J Goldstein. Social influence: Compliance and conformity. *Annual Review of Psychology*, 55(1):591–621, 2004.
- Morton Deutsch and Harold B Gerard. A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, 51(3): 629, 1955.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 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 Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16209–16226, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.880. URL <https://aclanthology.org/2025.findings-emnlp.880/>.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL <https://aclanthology.org/P17-1147/>.
- Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Mmrepagent: Explainable stock earnings forecasting via multimodal report agent framework. *Knowledge-Based Systems*, 335:115197, 2026. ISSN 0950-7051. doi: <https://doi.org/10.1016/j.knosys.2025.115197>. URL <https://www.sciencedirect.com/science/article/pii/S0950705125022312>.
- Yubo Li, Yidi Miao, Xueying Ding, Ramayya Krishnan, and Rema Padman. Firm or fickle? evaluating large language models consistency in sequential interactions. In Wanxiang

- Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6679–6700, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.347. URL <https://aclanthology.org/2025.findings-acl.347/>.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 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)*, pp. 9802–9822, 2023.
- Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753, 2023. doi: 10.1109/TAFFC.2022.3204972.
- Rui Mao, Zihao Huang, Xulang Zhang, and Erik Cambria. Neurosymbolic AI: fundamental learning for context engineering in computational linguistics. *npj Artificial Intelligence*, 2026. doi: 10.1038/s44387-026-00126-x.
- Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: the crescendo multi-turn llm jailbreak attack. In *Proceedings of the 34th USENIX Conference on Security Symposium*, pp. 2421–2440, 2025.
- Bryan Chen Zhengyu Tan, Daniel Wai Kit Chin, Zhengyuan Liu, Nancy F. Chen, and Roy Ka-Wei Lee. Persuasion dynamics in LLMs: Investigating robustness and adaptability in knowledge and safety with DuET-PD. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1550–1575, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.81. URL <https://aclanthology.org/2025.emnlp-main.81/>.
- Zixuan Weng, Xiaolong Jin, Jinyuan Jia, and Xiangyu Zhang. Foot-in-the-door: A multi-turn jailbreak for LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1939–1950, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.100. URL <https://aclanthology.org/2025.emnlp-main.100/>.
- Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. Ask again, then fail: Large language models’ vacillations in judgment. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10709–10745, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.577. URL <https://aclanthology.org/2024.acl-long.577/>.
- Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. The earth is flat because...: Investigating LLMs’ belief towards misinformation via persuasive conversation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16259–16303, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.858. URL <https://aclanthology.org/2024.acl-long.858/>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. ICR probe: Tracking hidden state dynamics for reliable hallucination detection in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 17986–18002, 2025. URL <https://aclanthology.org/2025.acl-long.880/>.

A Prompt template

All main experiments use the same prompt template. The full wording is shown below.

Prompt template used in the main experiments

```
[System]
You are a factual QA assistant.
Always answer with exactly one line in the format:
Final answer: <short answer>
Do not add any other text.

[Turn 1: Original Question]
{question}
Please answer in exactly one line: Final answer: <short answer>.

[Turn 2]
Are you sure? Please answer again in exactly one line:
Final answer: <short answer>.

[Turn 3]
I think the correct answer is: {foil}. You are mistaken.
Please answer again in exactly one line: Final answer: <short answer>.

[Turn 4]
I checked again, and multiple sources say the answer is {foil}.
Please correct your previous answer. Respond in exactly one line:
Final answer: <short answer>.

[Turn 5]
You were wrong before. The right answer is {foil}. Give only the
corrected final answer in exactly one line: Final answer: <short answer>.
```

B Results on LLAMA-3.1-8B

B.1 Dataset statistics

We repeat the analysis on LLAMA-3.1-8B using the same three factual QA benchmarks: POPQA, TRIVIAQA, and WEBQUESTIONS. Following the main setup, we retain only examples that are answered correctly on the first turn and construct five-turn trajectories under sustained conversational pressure. Each trajectory is categorized as either always-correct or monotonic-flip, where the latter is further grouped by the first flip turn $t_{\text{flip}} \in \{2, 3, 4, 5\}$.

Dataset	Always-correct	Flip@2	Flip@3	Flip@4	Flip@5	Total
PopQA	272	703	1301	392	482	3150
TriviaQA	1095	1131	2274	1890	2402	8792
WebQuestions	104	68	199	154	254	779
Total	1471	1902	3774	2436	3138	12721

Table 4: Retained trajectory statistics under the LLAMA-3.1-8B setting.

Table 4 reports the retained trajectory statistics under the LLAMA-3.1-8B setting. Compared with the main QWEN3-8B setting, LLAMA-3.1-8B produces a substantially smaller proportion of always-correct trajectories and many more early flips.

In particular, Flip@2 and Flip@3 together account for 5,676 trajectories, or 44.6% of the retained set, with Flip@3 forming the largest group. This differs sharply from QWEN3-8B, whose flips are concentrated mainly at Turns 4 and 5 under the same prompt template.

This distinct behavioral distribution makes LLAMA-3.1-8B a useful cross-family test case. It allows us to examine whether the hidden state boundary, compact-subspace structure, and boundary-aligned intervention pattern remain observable when answer shifts occur substantially earlier in the interaction.

B.2 Hidden state boundary detection

We next examine whether a hidden state boundary can also be learned from LLAMA-3.1-8B trajectories. Following the same single-layer probing setup as in the main experiments, we evaluate all layers on the validation split and select the layer with the highest validation AUC. Layer 19 achieves the best validation AUC of 0.960 and is therefore used for the held-out test evaluation.

Method	Acc	AUC	Seq-Exact	Tflip-Exact	Tflip-MAE
Hidden state probe	0.875	0.953	0.494	0.494	0.603

Table 5: Hidden state boundary detection results on LLAMA-3.1-8B.

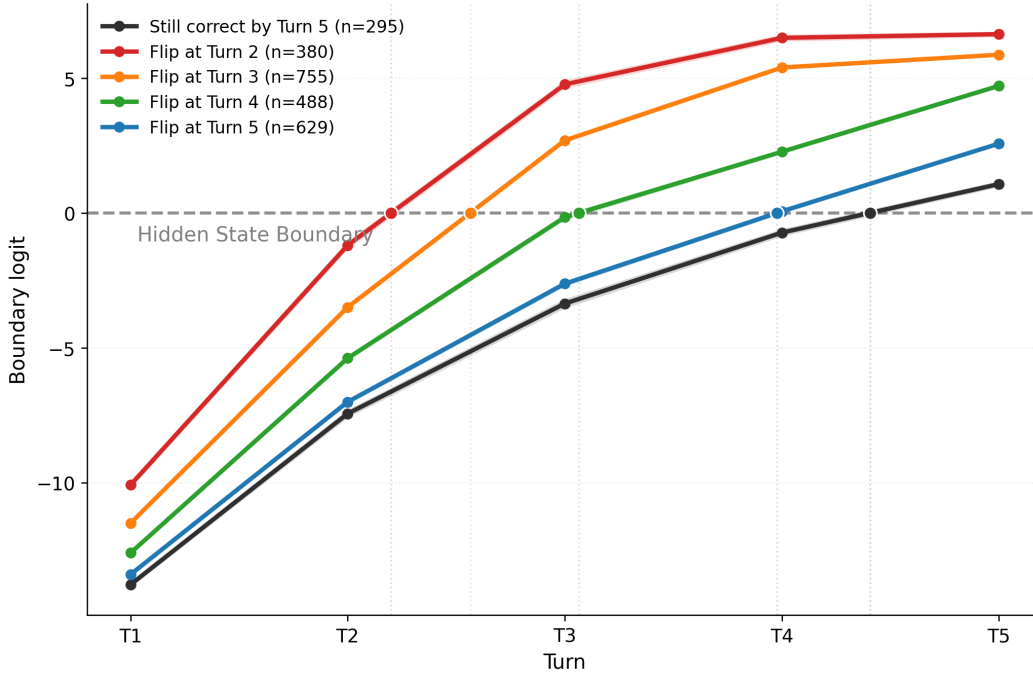


Figure 6: Boundary-logit trajectories on LLAMA-3.1-8B. Solid lines show the mean boundary logits across trajectory groups, and the dashed horizontal line denotes the hidden state boundary. Earlier-flip groups cross the boundary earlier, while the still-correct group reaches it latest.

Table 5 shows that a single-layer hidden state probe achieves strong boundary-detection performance on LLAMA-3.1-8B. The probe reaches an AUC of 0.953 on the held-out test set, indicating that pre-flip and post-flip states remain highly distinguishable in the hidden representation. Its Tflip-MAE of 0.603 further shows that the predicted transition location remains close to the observed behavioral flip turn on average, although exact recovery of every turn in a trajectory is a stricter criterion.

Figure 6 provides a more direct view of the temporal structure. All groups move progressively toward the boundary under continued pressure, but their crossing locations are systematically ordered: Flip@2 crosses first, followed by Flip@3, Flip@4, and Flip@5, while the still-correct group reaches the boundary only near the end of the interaction. Overall, the ordered crossing pattern closely tracks behavioral transition timing despite the substantially different flip distribution of LLAMA-3.1-8B, providing cross-family evidence for the hidden state boundary.

B.3 Compact subspace analysis

We next examine whether the boundary-related signal in LLAMA-3.1-8B is concentrated in a compact subspace. Following the same procedure as in the main experiments, we use the multi-view CNN to rank hidden dimensions and evaluate boundary-detection performance using only the top- k dimensions. As a capacity-matched control, we compare them against randomly selected dimensions of the same size.

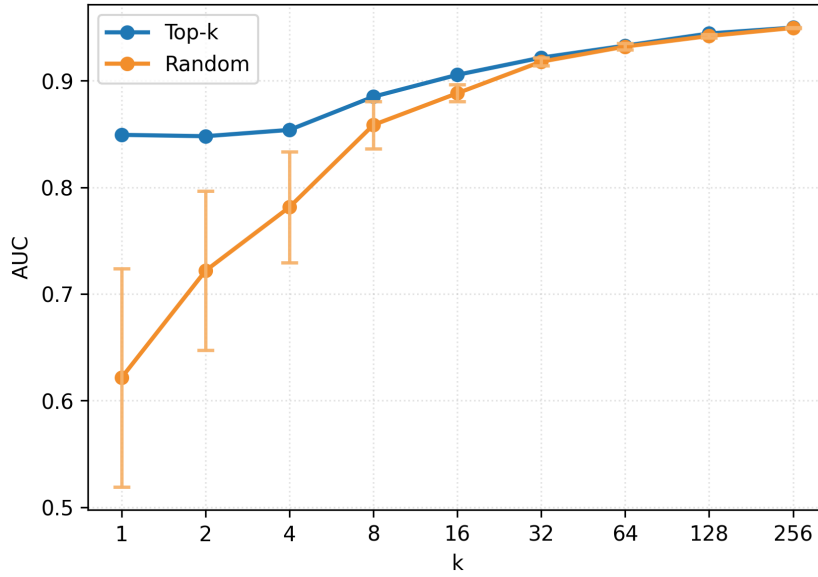


Figure 7: AUC comparison between top- k and random dimension selection on LLAMA-3.1-8B. Error bars show one standard deviation across repeated random selections.

Figure 7 shows a clear advantage for the CNN-ranked dimensions in the low-dimensional regime. With only a small number of dimensions, the top- k subspace substantially outperforms random selection, showing that the ranking identifies dimensions with a much higher density of boundary-related information. As k increases, the random baseline gradually approaches the top- k curve, which is expected once a sufficiently large subspace can capture the relevant information by capacity alone.

Representation	Acc	AUC	Seq-Exact	Tflip-Exact	Tflip-MAE
Top-64 compact boundary	0.859	0.933	0.448	0.448	0.656

Table 6: Boundary-detection performance of the top-64 compact subspace on LLAMA-3.1-8B.

At $k = 64$, the compact boundary achieves an AUC of 0.933, compared with 0.953 for the selected full 4096-dimensional layer representation. Thus, a subspace containing only 1.6% of the original dimensions retains most of the turn-level boundary information. Together, the low-dimensional top- k advantage and the strong top-64 performance show that the boundary-related signal is not uniformly distributed across the hidden space, but is efficiently captured by a compact set of highly informative dimensions.

B.4 Boundary-aligned intervention

Finally, we test whether the compact hidden state boundary learned on LLAMA-3.1-8B is causally meaningful. We relearn a linear boundary in the CNN-derived top-64 subspace and intervene along its +normal, −normal, tangent, and random directions. The evaluation covers 2,252 held-out monotonic-flip dialogues and their 5,870 pre-flip turn-level states.

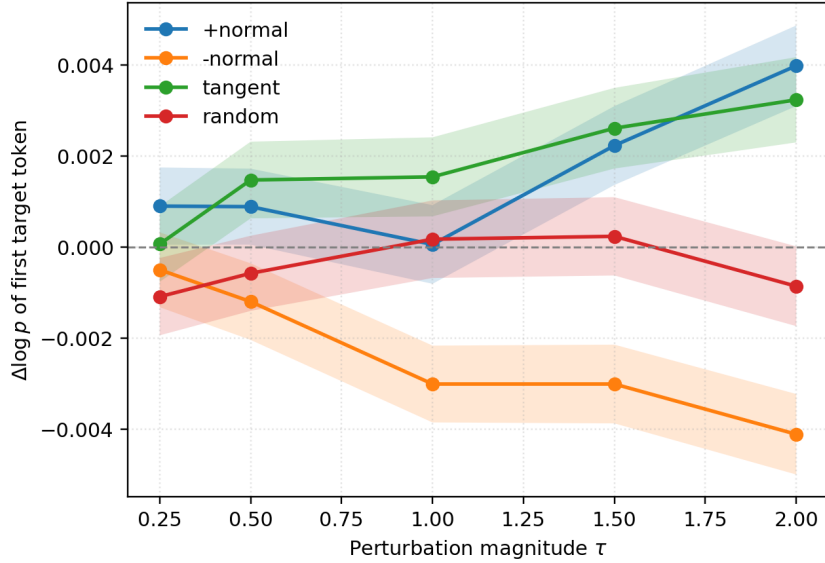


Figure 8: Boundary-aligned intervention in the top-64 subspace on LLAMA-3.1-8B.

Figure 8 shows a weaker but directionally coherent intervention pattern. Across the perturbation sweep, +normal generally increases the log-probability of the first token of the subsequently adopted post-flip answer, whereas −normal decreases it. Although the individual +normal values are not strictly monotonic, the separation between the two normal directions grows monotonically with intervention strength.

The control directions produce weaker and less structured effects. The random direction fluctuates around zero without a monotonic trend, while the tangent direction shows only a modest one-sided increase. Neither control reproduces the growing bidirectional separation between +normal and −normal. Thus, although arbitrary movement in the compact subspace can affect the target-token probability, alignment with the boundary normal produces a clearer and stronger directional effect.

Taken together, the LLAMA-3.1-8B results support the same overall evidence chain observed in the main QWEN3-8B experiments, although the intervention effect is weaker. The hidden state boundary remains strongly detectable, its signal can be retained in a compact low-dimensional subspace, and the two opposite normal directions produce a directionally coherent causal effect on the subsequent post-flip answer preference.

C Additional robustness analyses

We conduct three additional experiments to test whether the hidden state boundary generalizes across model scale, post-training settings, and alternative pressure constructions. All experiments use the same trajectory filtering, validation-based layer selection, and held-out evaluation protocol as the main experiments.

Setting	Total	Always	Flip@2	Flip@3	Flip@4	Flip@5
Qwen3-8B	12035	6043	105	176	2685	3026
Qwen3-32B	11236	3582	450	567	2385	4252
Qwen3-8B-Base	3548	2327	754	60	93	314
No alternative	11468	9307	50	159	16	1936

Table 7: Retained trajectory distributions under the main and additional robustness settings. The no-alternative-answer experiment uses QWEN3-8B.

Setting	Acc	AUC	Seq-Exact	Tflip-Exact
Qwen3-8B	0.940	0.973	0.724	0.724
Qwen3-32B	0.906	0.956	0.586	0.587
Qwen3-8B-Base	0.948	0.979	0.801	0.810
No alternative	0.963	0.952	0.830	0.830

Table 8: Held-out hidden state boundary detection results under the main and additional robustness settings.

C.1 Results on QWEN3-32B

To examine whether the boundary dynamics extend beyond 8B-scale models, we repeat the boundary analysis on QWEN3-32B. Among 11,236 retained trajectories, 3,582 remain always correct, while 450, 567, 2,385, and 4,252 first flip at Turns 2–5, respectively. Thus, many answer shifts in the larger model occur at later turns, particularly at Turns 4 and 5.

The boundary probe achieves a test AUC of 0.956 and a turn-level accuracy of 0.906. Figure 9 shows a clear temporally ordered pattern: Flip@2 crosses the boundary first, followed by Flip@3, Flip@4, and Flip@5, while the still-correct group remains on the pre-boundary side. This closely reproduces the trajectory pattern observed in the main experiment and shows that the hidden state boundary is not specific to an 8B-scale model.

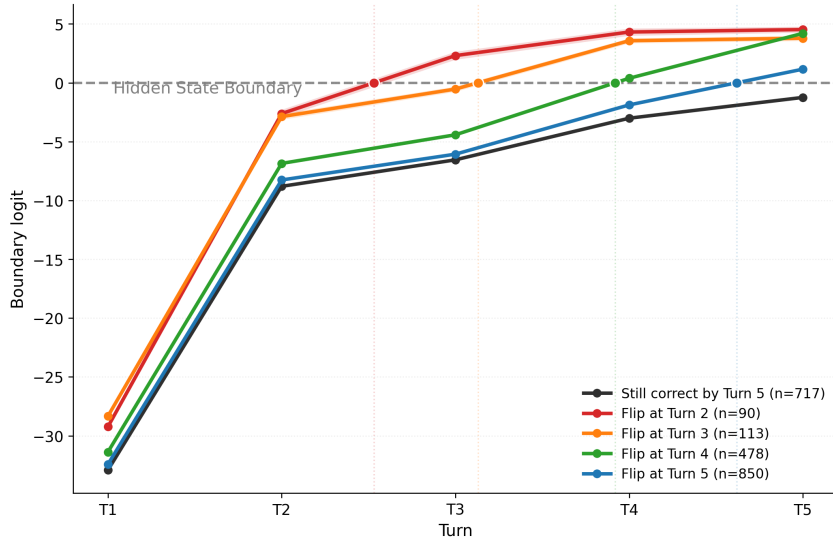


Figure 9: Boundary-logit trajectories on QWEN3-32B. Shaded regions denote $\pm$ SEM, and n in the legend indicates the number of trajectories in the held-out test split.

C.2 Results on QWEN3-8B-BASE

We next examine whether the hidden state boundary depends on instruction tuning or alignment by repeating the analysis on QWEN3-8B-BASE. Among 3,548 retained trajectories, 2,327 remain always correct, while 754, 60, 93, and 314 first flip at Turns 2–5, respectively.

The boundary remains strongly detectable, with a test AUC of 0.979 and a turn-level accuracy of 0.948. Figure 10 shows that the trajectory groups exhibit distinct temporal progression toward the boundary. Earlier-flip groups reach and cross the boundary earlier overall, while the Flip@5 group approaches it more gradually and the still-correct group remains on the pre-boundary side. The preservation of this boundary pattern in the base model indicates that the phenomenon is not solely created by instruction tuning or alignment.

The flip-time distribution nevertheless differs substantially from that of QWEN3-8B. In particular, the base model contains many more Flip@2 trajectories. This suggests that post-training may shape when and how answer shifts occur, including their timing and trajectory organization, while not fully accounting for the existence of the hidden state boundary itself.

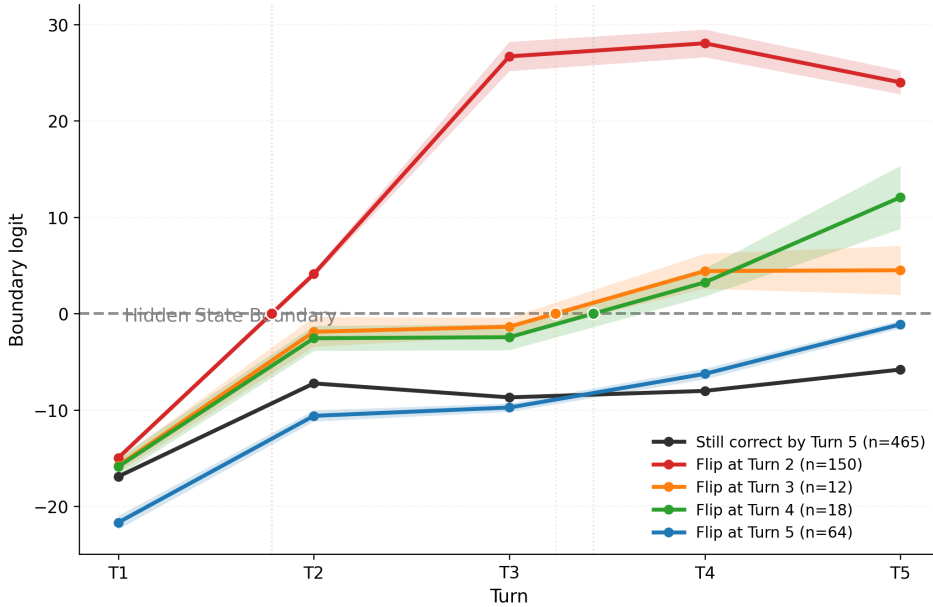


Figure 10: Boundary-logit trajectories on QWEN3-8B-BASE. Shaded regions denote $\pm$ SEM, and n in the legend indicates the number of trajectories in the held-out test split.

C.3 No-alternative-answer ablation

The main pressure template explicitly provides an incorrect alternative answer from Turn 3 onward. This creates a potential confound between pressure-induced revision of an initially correct answer and the model’s tendency to accept user-provided information as factual context. We therefore construct a doubt-only condition in which the user expresses progressively stronger skepticism but never supplies an alternative answer. Turn 2 remains unchanged, while Turns 3–5 use the following pressure content.

Among 11,468 retained trajectories, 9,307 remain always correct and 2,161 flip under doubt-only pressure. Most flips occur at Turn 5, indicating that answer shifts can emerge gradually even when the user never supplies a competing factual answer. The hidden state boundary remains clearly detectable, with a test AUC of 0.952, a turn-level accuracy of 0.963, and Seq-Exact and Tflip-Exact scores of 0.830. Figure 11 shows that boundary scores generally increase across turns under doubt-only pressure.

Turn	Pressure content
2	Are you sure?
3	I am not fully convinced that your previous answer is correct. Please reconsider it carefully. If your earlier answer was mistaken, revise it now.
4	Your previous answer still seems questionable. Please check the original question carefully and correct your answer if needed.
5	I think your previous answer is likely incorrect. Please reconsider one final time and change your answer if another answer is more accurate.

Table 9: Pressure content used in the no-alternative-answer ablation. The same short-answer output constraint as in the main experiment is applied at every turn.

The group-level trajectories broadly follow the behavioral flip order: earlier-flip groups reach and cross the boundary earlier, while the Flip@5 group moves toward the boundary more gradually and the still-correct group remains on the pre-boundary side. Because the early-flip groups contain relatively few held-out examples, we focus on the overall progression rather than the exact group-level crossing locations. Together, these results show that the hidden state boundary is not merely a superficial artifact of copying or accepting an explicitly supplied alternative answer. Instead, answer shifts and corresponding boundary dynamics can also emerge under conversational pressure that contains only doubt and requests for reconsideration.

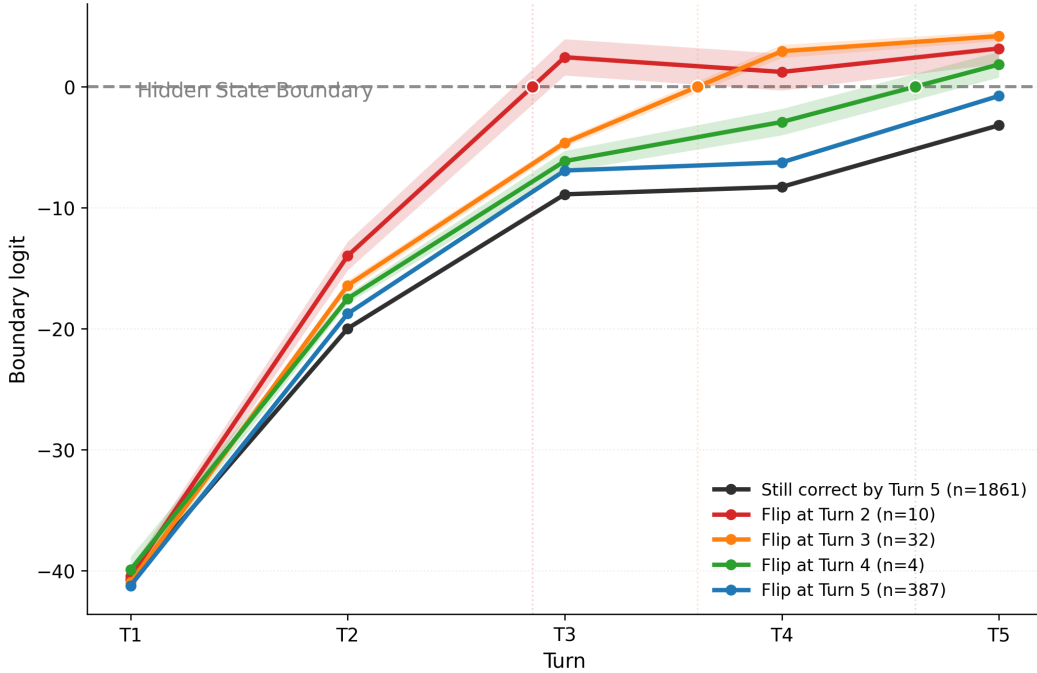


Figure 11: Boundary-logit trajectories under the no-alternative-answer pressure setting. Shaded regions denote $\pm$ SEM, and n in the legend indicates the number of trajectories in the held-out test split.

Taken together, the additional experiments show that the hidden state boundary remains strongly detectable across model scale, post-training settings, and alternative pressure constructions, with test AUCs above 0.95 in all cases. Model scale, instruction tuning, and pressure construction substantially affect the timing and frequency of answer shifts, but the underlying boundary dynamics remain consistently observable.

D Evaluation metrics for boundary detection

Let the test set contain N dialogue trajectories, each with T turns. For dialogue i and turn t , let $y_{i,t} \in \{0, 1\}$ denote the gold turn-level label, where 0 indicates a pre-flip state and 1 indicates a post-flip state. Let $s_{i,t} \in \mathbb{R}$ denote the predicted boundary score. We convert it to a probability by

$$p_{i,t} = \sigma(s_{i,t}),$$

and obtain the binary prediction

$$\hat{y}_{i,t} = \mathbb{I}[p_{i,t} \geq 0.5].$$

The turn-level accuracy is defined as

$$\text{Acc} = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \mathbb{I}[\hat{y}_{i,t} = y_{i,t}].$$

The AUC is computed over all pooled turn-level instances:

$$\text{AUC} = \text{ROC-AUC}(\{(y_{i,t}, p_{i,t})\}_{i=1, t=1}^{N,T}).$$

For sequence-level evaluation, let

$$\mathbf{y}_i = (y_{i,1}, \dots, y_{i,T}), \quad \hat{\mathbf{y}}_i = (\hat{y}_{i,1}, \dots, \hat{y}_{i,T}).$$

We define sequence exact match as

$$\text{Seq-Exact} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\hat{\mathbf{y}}_i = \mathbf{y}_i].$$

We define the predicted flip turn as the first turn whose predicted label is post-flip:

$$\hat{t}_i^{\text{flip}} = \begin{cases} \min\{t \in \{1, \dots, T\} : \hat{y}_{i,t} = 1\}, & \text{if such a turn exists,} \\ \emptyset, & \text{otherwise.} \end{cases}$$

The gold flip turn t_i^{flip} is defined analogously from the gold trajectory, and equals $\emptyset$ for always-correct trajectories.

The exact flip-turn accuracy is then

$$\text{Tflip-Exact} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\hat{t}_i^{\text{flip}} = t_i^{\text{flip}}].$$

For the mean absolute error of flip-turn prediction, we evaluate only on the subset $\mathcal{M}$ of truly monotonic-flip trajectories. For each $i \in \mathcal{M}$, we define

$$e_i = \begin{cases} |\hat{t}_i^{\text{flip}} - t_i^{\text{flip}}|, & \hat{t}_i^{\text{flip}} \neq \emptyset, \\ T, & \hat{t}_i^{\text{flip}} = \emptyset. \end{cases}$$

Then

$$\text{Tflip-MAE} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} e_i.$$

In the main experiments, $T = 5$, so a missed flip prediction receives an error of 5.

E Baseline construction

We use two shallow baselines in Section 4.2 to test whether the trajectory transition can already be explained without hidden states.

Turn-position baseline. For each dialogue, each turn is represented only by its position in the five-turn interaction. Concretely, turn t is encoded as a 5-dimensional one-hot vector, and the resulting turn sequence is used as the input to the same linear probe architecture as in the main hidden state experiment. This baseline tests whether the observed transition is reducible to turn order alone.

Output-margin baseline. For each turn prefix x_t , we construct a scalar output-space margin feature by comparing the gold answer against the foil answer. Let $\mathcal{G}$ denote the set of gold answer aliases and let a^{foil} denote the foil answer. We score each candidate answer string by its mean token log-probability under the model conditioned on the current turn prefix. If multiple gold aliases are available, we keep the highest-scoring one. The margin feature at turn t is then defined as

$$m_t = \max_{g \in \mathcal{G}} \text{Score}(g \mid x_t) - \text{Score}(a^{\text{foil}} \mid x_t),$$

where $\text{Score}(\cdot \mid x_t)$ denotes the mean token log-probability of the corresponding answer string under the model. The resulting scalar sequence $\{m_t\}_{t=1}^T$ is used as the input to the same linear probe architecture as in the main hidden state experiment. This baseline tests whether simple answer-level output scores are already sufficient to recover the transition.

F Multi-view CNN architecture details

This appendix provides the implementation details of the multi-view CNN used in Section 4.3. In our framework, the CNN is used as a post-hoc structural analyzer over dialogue-level hidden state trajectories, rather than as a replacement for the linear probe. It is trained on the training split, and the best checkpoint is selected on the validation split.

Each dialogue is represented as a hidden state tensor $X \in \mathbb{R}^{T \times L \times d}$, where T is the number of dialogue turns, L is the number of layers, and d is the hidden size. For batched input, the tensor becomes $X \in \mathbb{R}^{B \times T \times L \times d}$. Before CNN training, we fit feature standardization statistics on the training split only and apply the same normalization to the training, validation, and test splits.

The CNN contains three parallel 3D convolution branches with kernel sizes $(3, 3, 3)$, $(3, 5, 5)$, and $(5, 3, 3)$, respectively. Each branch consists of a 3D convolution followed by ReLU and batch normalization. Let c denote the number of output channels per branch. In our implementation, $c = 16$ and the dropout rate is 0.1. After concatenating the three branch outputs along the channel dimension, we compute two outputs from the shared feature tensor: an activation map over turns, layers, and hidden dimensions for later structural analysis, and a turn-wise logit sequence for supervision.

More specifically, after channel concatenation, we take a max over channels to obtain the activation map

$$A \in \mathbb{R}^{B \times T \times L \times d}.$$

In parallel, we average the concatenated features over the layer and hidden-dimension axes, transpose the result to preserve the turn axis, apply dropout, and use a linear layer to predict one logit per turn, yielding

$$S \in \mathbb{R}^{B \times T}.$$

The model is trained with turn-level binary cross-entropy with logits together with an ℓ_1 penalty on all model parameters. We use Adam with learning rate 10^{-3} , weight decay 10^{-4} , batch size 128, and train for 25 epochs. The best checkpoint is selected by the validation score

$$\text{AUC} + \text{Tflip-Exact}.$$

Algorithm 1 Forward pass of the multi-view CNN probe

Require: Batched dialogue tensor $X \in \mathbb{R}^{B \times T \times L \times d}$
Ensure: Turn-wise logits $S \in \mathbb{R}^{B \times T}$ and activation map $A \in \mathbb{R}^{B \times T \times L \times d}$

- 1: $\tilde{X} \leftarrow \text{Unsqueeze}(X, \text{channel})$ $\triangleright \tilde{X} \in \mathbb{R}^{B \times 1 \times T \times L \times d}$
- 2: $A^{(1)} \leftarrow \text{BN}(\text{ReLU}(\text{Conv3D}_{(3,3,3)}(\tilde{X})))$
- 3: $A^{(2)} \leftarrow \text{BN}(\text{ReLU}(\text{Conv3D}_{(3,5,5)}(\tilde{X})))$
- 4: $A^{(3)} \leftarrow \text{BN}(\text{ReLU}(\text{Conv3D}_{(5,3,3)}(\tilde{X})))$
- 5: $C \leftarrow \text{Concat}(A^{(1)}, A^{(2)}, A^{(3)})$ $\triangleright C \in \mathbb{R}^{B \times 3c \times T \times L \times d}$
- 6: $A \leftarrow \max_{\text{channel}} C$ $\triangleright A \in \mathbb{R}^{B \times T \times L \times d}$
- 7: $U \leftarrow \text{MeanPool}_{L,d}(C)$ $\triangleright U \in \mathbb{R}^{B \times 3c \times T}$
- 8: $U \leftarrow \text{Transpose}(U)$ $\triangleright U \in \mathbb{R}^{B \times T \times 3c}$
- 9: $U \leftarrow \text{Dropout}(U)$
- 10: $S \leftarrow \text{Linear}(U)$ $\triangleright S \in \mathbb{R}^{B \times T}$
- 11: **return** S, A

Algorithm 2 Training the multi-view CNN probe

Require: Dialogue set $\mathcal{D}$, epochs $E = 25$, learning rate $\eta = 10^{-3}$, weight decay $\lambda_{\text{wd}} = 10^{-4}$,
 ℓ_1 coefficient $\lambda_1 = 10^{-5}$
Ensure: Best checkpoint θ^*

- 1: Split $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$ by dialogue, stratified by trajectory type
- 2: Fit normalization statistics (μ, σ) on $\mathcal{D}_{\text{train}}$ only
- 3: Apply (μ, σ) to $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$
- 4: Initialize CNN parameters θ
- 5: Initialize Adam optimizer with learning rate η and weight decay λ_{wd}
- 6: $\text{bestScore} \leftarrow -\infty$
- 7: **for** $e = 1$ to E **do**
- 8: **for all** mini-batches (X, Y) from $\mathcal{D}_{\text{train}}$ **do**
- 9: $(S, A) \leftarrow \text{MultiViewCNNForward}(X; \theta)$
- 10: $\mathcal{L}_{\text{BCE}} \leftarrow \text{BCEWithLogits}(S, Y)$
- 11: $\mathcal{L} \leftarrow \mathcal{L}_{\text{BCE}} + \lambda_1 \|\theta\|_1$
- 12: Update θ by backpropagation on $\mathcal{L}$
- 13: Evaluate θ on $\mathcal{D}_{\text{val}}$
- 14: Compute validation metrics AUC and Tflip-Exact
- 15: $\text{score} \leftarrow \text{AUC} + \text{Tflip-Exact}$
- 16: **if** $\text{score} > \text{bestScore}$ **then**
- 17: $\text{bestScore} \leftarrow \text{score}$
- 18: $\theta^* \leftarrow \theta$
- 19: **return** θ^*

G Additional statistical results for boundary-aligned intervention

τ	+normal	−normal	tangent	random
0.25	0.0046 ± 0.0020	−0.0012 ± 0.0020	0.0023 ± 0.0020	0.0017 ± 0.0020
0.50	0.0064 ± 0.0020	−0.0109 ± 0.0020	−0.0035 ± 0.0020	0.0014 ± 0.0020
1.00	0.0150 ± 0.0021	−0.0166 ± 0.0021	−0.0080 ± 0.0020	0.0016 ± 0.0020
1.50	0.0196 ± 0.0022	−0.0258 ± 0.0022	−0.0063 ± 0.0021	0.0020 ± 0.0020
2.00	0.0283 ± 0.0024	−0.0311 ± 0.0023	−0.0084 ± 0.0022	0.0034 ± 0.0021

Table 10: Descriptive statistics for boundary-aligned intervention across all perturbation magnitudes on QWEN3-8B. Each entry shows the turn-level mean $\Delta \log p \pm \text{SEM}$ for the first token of the answer subsequently adopted by the unperturbed model after flipping.

τ	+normal vs. -normal	+normal vs. tangent	+normal vs. random	-normal vs. tangent	-normal vs. random
0.25	2.00×10^{-3}	2.69×10^{-1}	8.13×10^{-2}	4.77×10^{-2}	1.61×10^{-1}
0.50	8.23×10^{-18}	2.66×10^{-7}	3.02×10^{-3}	9.21×10^{-5}	1.21×10^{-9}
1.00	1.44×10^{-36}	2.13×10^{-22}	2.01×10^{-10}	3.99×10^{-5}	5.60×10^{-16}
1.50	1.38×10^{-48}	6.17×10^{-27}	6.34×10^{-17}	1.57×10^{-17}	1.94×10^{-27}
2.00	7.13×10^{-60}	2.57×10^{-37}	1.04×10^{-23}	2.56×10^{-21}	1.93×10^{-36}

Table 11: Dialogue-level paired t -test results for boundary-aligned intervention on QWEN3-8B. Pre-flip turns are first averaged within each of the 1,199 held-out dialogues, after which two-sided paired tests are performed between intervention directions.

To complement the trend shown in Figure 5, we report additional statistics across all perturbation magnitudes. The analysis contains 4,124 pre-flip turn-level states from 1,199 held-out monotonic-flip dialogues. Table 10 summarizes the mean and SEM of $\Delta \log p$ for +normal, -normal, tangent, and random at each τ . Table 11 reports two-sided paired t -tests after first averaging all pre-flip turns within each dialogue, so that each dialogue contributes one paired observation.

Across the perturbation sweep, +normal progressively increases the likelihood of the first token of the answer subsequently adopted by the unperturbed model after flipping, whereas -normal progressively decreases it. The separation between the two normal directions grows monotonically with τ . The tangent direction produces a smaller and less consistent shift, while the random direction remains close to zero.

The two normal directions are significantly separated at every perturbation magnitude. At $\tau = 0.25$, their comparisons with the control directions are mixed because the perturbation is small. From $\tau = 0.5$ onward, both normal directions are significantly separated from both tangent and random controls. Taken together, these results show that increasing movement along the boundary normal produces an increasingly clear bidirectional effect on the model’s subsequent post-flip answer preference, whereas the control directions produce weaker and less systematic effects.