StateSwap: Probing Support–Elimination Hidden States in Multiple-Choice Questions

Chao Gao[1], Haijiang Liu[2,3], Qiyuan Li[2,3], Caicai Guo[2,3], Frank van Harmelen[1], Jinguang Gu[2,3]

[1]Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[2]School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China
[3]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan, China

Correspondence: c.gao@vu.nl, simon@wust.edu.cn

Abstract

Large language models often answer the same multiple-choice question inconsistently when it is posed under support-oriented and elimination-oriented framings. We investigate whether these discrepancies arise from different internal representations induced by the two framings. We introduce a dual-framing protocol with minimally varied prompts that use either support- or elimination-oriented framing while keeping the evaluation target fixed. To probe the internal computation, we append an untrained special token, [STATE], and treat its residual-stream activation as an intervention interface. Across both models, the two framings induce separable [STATE] activations concentrated in intermediate layers. Swapping these activations between paired prompts systematically changes predictions and improves cross-framing agreement, providing intervention-based evidence that the activations are behaviorally relevant. Beyond instance-level substitution, mean-difference steering directions derived from the dual-framing contrast exhibit more bounded layer-wise responses than matched contrastive activation addition directions under the evaluated protocol. Code is available at https://github.com/Cha0Ga0/SWAPSTATE

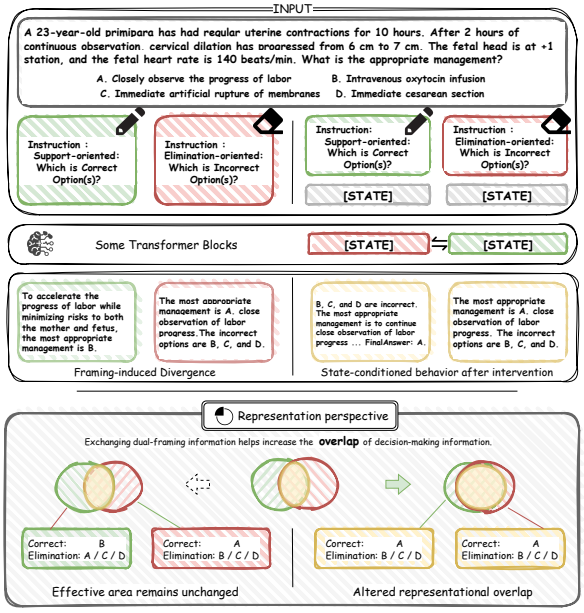


Figure 1: **Overview of StateSwap.** StateSwap swaps [STATE] activations between paired SUP and ELIM prompts while keeping the textual input fixed. The figure shows an idealized agreement-increasing outcome for illustration only; substitution may improve, reduce, or leave unchanged accuracy and agreement. We interpret systematic prediction changes, including but not limited to improvements, as evidence of behavioral relevance under intervention.

1 Introduction

Recent work has shown that the process-of-elimination (PoE) strategy, widely used by humans for solving multiple-choice questions (MCQs), can improve the performance of large language models (LLMs) (Ma and Du, 2023). This approach is motivated by the fact that support-oriented (SUP) prompts ("which option is correct?") and elimination-oriented (ELIM) prompts ("which options are incorrect?") are logically tied to the same answer key. For example, some studies first prompt the model to identify and eliminate incorrect options before deriving the final answer (Zhu et al., 2025; Fu et al., 2025; Wang et al., 2026; Zhao and Zhang, 2025). However, empirical studies also report that elimination-style prompting can instead impair accuracy (Balepur et al., 2024). At first glance, this appears contradictory: if SUP and ELIM prompts encode the same task objective and semantic content for decision-making, they should not lead to different predictions. This motivates the following question: do the two framings induce distinguishable internal representations?

Inspired by recent mechanistic studies (Heimersheim and Nanda, 2024; Zhang et al., 2025; Lu et al., 2025), we investigate prompt framing under

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 6343–6364
October 24-29, 2026 ©2026 Association for Computational Linguistics

controlled inference settings, using deterministic decoding and paired prompt evaluation to isolate representation-level effects from decoding variability. We test whether framing-dependent behavioral differences are accompanied by distinguishable activation patterns in the model's hidden-state space. Separability identifies a representational correlate of prompt framing, and activation substitution evaluates its behavioral relevance.

Based on this perspective, we investigate how SUP and ELIM prompts shape the internal representations of the model, and whether targeted activation substitution can provide intervention-based evidence on the role of these representations. Specifically, we use a newly registered, untrained special token with no pretrained lexical semantics, `[STATE]`, to analyze the geometry of the hidden states elicited by the two framings and assess their representational distinguishability. We then introduce activation-level interventions that replace selected internal states while keeping the textual input fixed, allowing us to examine how changes in the induced representation affect the model's output distribution over answer indices. Comparisons with dual-prompt ensemble baselines distinguish activation substitution from simple answer aggregation. Finally, we test whether the intervention remains effective when `[STATE]` is moved away from its canonical final position.

The main contributions of this work are threefold:

1. We identify distinguishable `[STATE]` representations induced by SUP and ELIM prompts that encode the same underlying decision problem.

2. We show that substituting framing-sensitive `[STATE]` activations changes model predictions and is associated with higher accuracy and cross-framing agreement in the evaluated settings.

3. We find that mean-difference steering directions derived from the two-framing prompt contrast exhibit lower layer-wise variability than matched contrastive activation addition directions under the evaluated protocol.

2 Related Work

LLM reasoning has been studied from multiple complementary perspectives, including prompting, fine-tuning, symbolic tools, model collaboration, and output verification (He et al., 2025). Sparse-autoencoder analyses characterize activation patterns associated with chain-of-thought reasoning (Chen et al., 2025), while causal tracing and model editing link localized computations to factual predictions (Meng et al., 2022). Circuit-level studies localize induction-like computations to particular attention heads (Olsson et al., 2022), and activation patching provides a framework for testing the behavioral relevance of localized components (Heimersheim and Nanda, 2024). Synthetic state-tracking experiments further show that transformers can develop localized mechanisms, including late-layer MLP neurons, for representing and updating latent states (Zhang et al., 2025).

This interventionist view is further supported by recent work on inference-time intervention and activation steering demonstrating that high-level behaviors in LLMs can be influenced by manipulating directions in activation space (Li et al., 2023; Turner et al., 2024; Zou et al., 2023). Methods such as contrastive activation addition (CAA) (Rimsky et al., 2024), conditional activation steering (Lee et al., 2025), and dynamic steering approaches (Li et al., 2025) show that activation shifts derived from dataset-level statistics can systematically alter model outputs, including truthfulness, refusal behavior, and bias-related responses. These results provide strong evidence that internal representations encode behaviorally meaningful information in explorable and manipulable forms.

A parallel line of work explores interface-based control of internal computation via localized representation slots. In encoder-style models, special tokens such as `[CLS]` serve as sequence-level interfaces to global representations (Devlin et al., 2019), while in autoregressive models, parameter-efficient methods such as prompt tuning and prefix tuning learn continuous embeddings that steer model behavior without modifying base parameters (Lester et al., 2021; Li and Liang, 2021; Liu et al., 2022). These approaches highlight the role of localized interfaces as high-leverage access points to latent representations.

Motivated by this line of work, we adopt an intervention-based perspective to investigate how prompt framing shapes internal representations. To operationalize this perspective, we use a fixed, untrained token as a minimal read–write interface to the residual stream. This setup allows us to examine whether semantically equivalent prompts

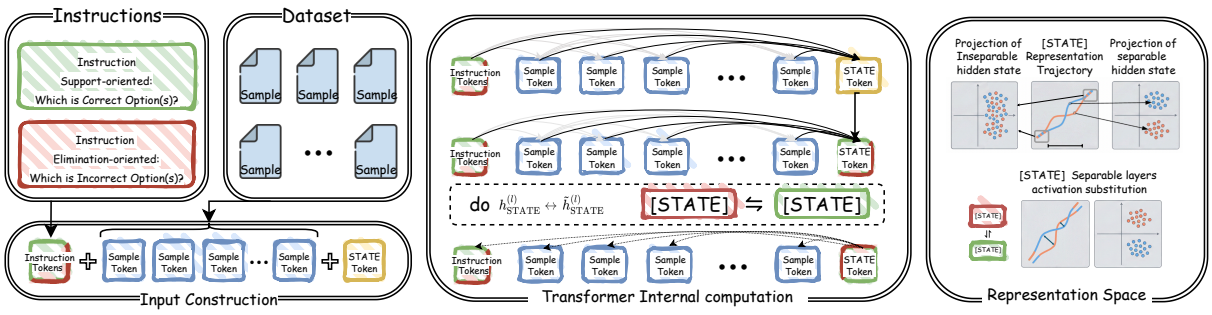


Figure 2: Paired SUP (green) and ELIM (red) prompts place an aligned [STATE] interface (yellow) at the same position. Layer-wise diagnostics localize candidate intervention regions (Section 3.3). StateSwap then replaces the full [STATE] activation between the paired framings within the selected layers (Section 3.2).

with different framings give rise to distinguishable decision-related representations for the same input instance and whether these representations can be aligned through instance-level intervention.

3 Methodology

We investigate whether framing-conditioned [STATE] activations are behaviorally relevant under controlled substitution. Section 3.1 defines the aligned inputs and state interface, Section 3.2 formalizes cross-framing substitution, and Section 3.3 describes how the intervention layers are localized; Appendix A.1 summarizes the notation.

3.1 Input Construction and State Interface

Consider a causal language model f_θ with L Transformer blocks. We append a newly registered [STATE] token to each prompt, yielding an input sequence $\mathbf{x} = (x_1, \ldots, x_T)$ in which [STATE] occupies the final position $t_S \triangleq T$. Let $\mathbf{h}_t^{(l)} \in \mathbb{R}^d$ denote the residual-stream representation at position t after Transformer block l, where $t \in \{1, \ldots, T\}$ and $l \in \{1, \ldots, L\}$. We define the corresponding [STATE] representation as

$$\mathbf{s}^{(l)} \triangleq \mathbf{h}_{t_S}^{(l)}.$$

Although [STATE] has no pretrained lexical semantics and receives no task-specific training, its contextualized representation can aggregate information from the preceding prompt and propagate through later layers to affect subsequent generation. We therefore use $\mathbf{s}^{(l)}$ as a single-token read–write interface for inference-time interventions. Token registration and embedding initialization are described in Appendix A.2.

For each question q, a pair of prompts is constructed, differing only in the framing instruction: eliminating incorrect options (ELIM) versus supporting correct options (SUP). Let $I \in \{\text{ELIM}, \text{SUP}\}$ denote the framing instruction. Each prompt ends with [STATE]:

$$\mathbf{x}_I(q) = \text{Concat}(I, q, \texttt{[STATE]}).$$

To ensure that [STATE] occupies the same token index in both framings, instruction segments are padded to the same token length (Appendix A.2).

3.2 State Substitution and Prediction Sensitivity

Let $W \subseteq \{1, \ldots, L\}$ denote the contiguous intervention region localized by the diagnostic in Section 3.3. For framing I, let $\bar{I}$ denote its complement. Each substitution direction is run independently on $\mathbf{x}_I(q)$, with the cached donor state from $\bar{I}$ substituted over W:

$$\mathbf{s}^{(k)}(I, q) \leftarrow \mathbf{s}^{(k)}_{\text{cache}}(\bar{I}, q), \quad k \in W.$$

For each $k \in W$, the intervention overwrites only the post-block state $\mathbf{s}^{(k)}(I, q)$ during the prompt-processing pass; all other residual-stream vectors at the intervention point remain unchanged, although downstream activations may change through residual propagation. Computation then proceeds normally from the substituted states to standard autoregressive generation. Feature-window slices are used only for diagnostics and localization of W; the intervention itself replaces the full d-dimensional state $\mathbf{s}^{(k)}$.

3.3 Locating a Stable Intervention Window

For $\mathbf{s}^{(l)} \in \mathbb{R}^d$, candidate regions are identified by scanning contiguous feature windows of size w:

$$\mathbf{s}^{(l)}_{[j:j+w)} \in \mathbb{R}^w, \quad j \in \{0, \ldots, d-w\}.$$

For each layer l and window (j, w), a paired Cohen's d statistic is computed from the paired question-wise ELIM–SUP differences (Appendix A.3). Two scalarizations are used.

Random-direction diagnostic Once per run, a single random unit vector $\mathbf{u} \in \mathbb{R}^d$ is sampled and projected onto the corresponding window slice:

$$\phi^{\text{rand}}_{j,w}(\mathbf{z}) = \mathbf{u}^{\top}_{[j:j+w)}\mathbf{z},$$

yielding $d^{\text{rand}}_l(j, w)$.

Mean-difference diagnostic For each (l, j, w), we form the mean-difference direction

$$\mathbf{v}_l(j, w) = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{s}^{(l)}_{\text{ELIM},i,[j:j+w)} - \mathbf{s}^{(l)}_{\text{SUP},i,[j:j+w)}\right),$$

stabilize its normalization and project:

$$\phi^{\text{md}}_{l,j,w}(\mathbf{z}) = \hat{\mathbf{v}}_l(j, w)^{\top}\mathbf{z},$$
$$\hat{\mathbf{v}}_l(j, w) = \frac{\mathbf{v}_l(j, w)}{\max(\|\mathbf{v}_l(j, w)\|, \epsilon_v)},$$

where $\epsilon_v > 0$ is small, yielding $d^{\text{md}}_l(j, w)$.

Window-level separability scores are summarized into a per-layer intensity for each layer. Layers are ranked by this intensity, and adjacent selected layers are merged into a contiguous intervention region W; the aggregation and selection procedure is given in Appendix A.3.

4 Experimental Setup and Evaluation Protocol

4.1 Datasets

To study prompt-framing effects under controlled conditions, we evaluate on two multiple-choice benchmarks: (i) a reasoning-focused subset of MMLU covering 17 subject categories (MMLU-17) (Hendrycks et al., 2021), and (ii) MedQA-CH, a Chinese medical benchmark (Jin et al., 2021), which require non-trivial reasoning and allow unambiguous comparison across equivalent formulations. To avoid data leakage, we use 50 calibration examples from the training split solely to localize a candidate intervention window; these examples do not enter downstream test evaluation, and the selected window is fixed for all reported experiments. After fixing the window, we conduct a post hoc robustness analysis using 1,000 examples and 500 bootstrap resamples. Dataset statistics are provided in Appendix B.1. All downstream inference and reported evaluations use the official test splits.

4.2 Models and decoding

We evaluate on reasoning-focused multiple-choice benchmarks and open-weight LLMs Qwen-2.5-7B-Instruct (Yang et al., 2024) and GLM-4-9B (GLM et al., 2024) in a zero-shot setting. All generations use deterministic decoding (greedy; no sampling) to avoid confounding intervention effects with sampling variance. We fix the chat template, context length, and decoding hyperparameters across all runs; details are provided in Appendix B.2.

4.3 Prompt structure and [STATE]

All prompts consist of an instruction segment, the question with options, and a final [STATE] token. The exact instruction templates are provided in Appendix B.3. To verify robustness to random [STATE] initialization, we rerun our method on randomly sampled instances with different random seeds while keeping all other settings fixed, and observe negligible variation in the generated responses (Appendix B.4).

4.4 Inference

We consider two framings: a SUP instruction, which asks the model to predict the correct option(s), and an ELIM instruction, which asks the model to predict the incorrect option(s). For each framing, we evaluate a baseline condition and an intervention condition in which the [STATE] representation is substituted over a selected layer region W using the other framing. All other computations are held fixed. Outputs from both framings are parsed into a shared option-set representation, as detailed in Appendix B.5.1.

For reported ensemble results, we use paired deterministic prompt realizations within each framing. Thus, SUP and ELIM each consist of two prompt-instantiated runs, yielding two framing-specific voting groups. We use strict-majority voting within each group.

4.5 Metrics

We report accuracy (ACC) for each framing and cross-framing decision overlap measured by the Jaccard index. We further analyze correctness to distinguish symmetric and asymmetric outcomes. Response-level stability is evaluated using Length-Aware Cosine Similarity (LAC) and BERTScore-F1 computed on the full generated responses under deterministic decoding. Cosine similarity captures global embedding-level alignment, while

BERTScore-F1 reflects token-level semantic consistency between responses. Metric definitions and parsing rules are provided in Appendix B.5.2.

5 Results

We evaluate the proposed framework through the following evaluation questions (EQs):

EQ1 (Representation) Do the two framings induce separable internal representations?

EQ2 (Effects of State Substitution) Does exchanging the final-position [STATE] representation transfer decision-relevant information between the two framings?

EQ3 (Mechanism Specificity) Are these effects specific to the [STATE] interface rather than generic perturbations?

We additionally conduct a separate extension analysis to examine whether the underlying SUP–ELIM contrast remains useful without the [STATE] interface in a population-level steering setting.

5.1 EQ1: Separable Decision-Related Representations

Before evaluating the downstream effects of [STATE] substitution, we examine whether the two framings produce systematically different internal representations at the [STATE] interface. In this diagnostic step, representation-level separability localizes a plausible layer window W for subsequent interventions.

Separability diagnostics Figure 3 visualizes the [STATE] representations across representative layers. The SUP and ELIM representations are largely mixed in the early layers, become clearly separated in the intermediate layers, and show weaker separation in the later layers. This pattern suggests that framing-dependent structure at the [STATE] position is concentrated in a subset of intermediate layers rather than uniformly distributed across depth. We further quantify this trend with layer-wise separability diagnostics for both Qwen-2.5-7B and GLM-4-9B in Appendix D.3.

Localization of the intervention window W Based on the intermediate-layer concentration observed in Figure 3 and the layer-wise separability diagnostics, we select a compact intervention window W for [STATE] substitution. For Qwen-2.5-7B, we use layers 11–20 as the intervention window.

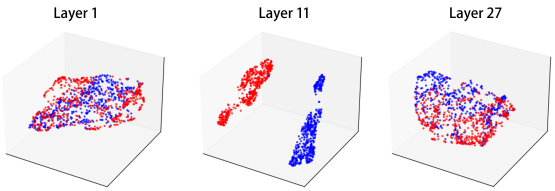


Figure 3: PCA visualization of [STATE] representations for Qwen-2.5-7B at Layers 1, 11, and 27. Red and blue points denote SUP and ELIM prompts, respectively. The two framings are largely mixed in the early layers, become clearly separated in the intermediate layers, and show weaker separation in the later layers. Additional layer-wise PCA visualizations are provided in Appendix D.1.

For GLM-4-9B, the same separability-based criterion selects layers 21–40; full cross-model diagnostics are reported in Appendix D.3. These windows are determined from representation-level separability diagnostics rather than downstream task performance. A post-hoc comparison across candidate Qwen layer ranges, including the magnitude and stability of output-distribution shifts under substitution, is provided in Appendix D.2. The stability of the localized intermediate-layer region under a larger calibration sample and bootstrap resampling is evaluated in Appendix D.4.

Random-label control To assess whether the observed layer-wise separability could arise from spurious correlations or artifacts of the analysis procedure, we perform a random-label control in which the binary framing labels are randomly permuted while all representations and projection settings are held fixed. This provides a baseline for estimating the level of separability expected under chance alignment.

As shown in Figure 4, the separability obtained from 10 random label permutations remains consistently low across layers and window sizes, in contrast to the structured intermediate-layer peak observed under the true labels. The gap between the two conditions persists across all window sizes considered, indicating that the localization pattern is not an artifact of windowing, random projection, or label imbalance.

5.2 EQ2: Effects of State Substitution

Decision-related behavioral effects of [STATE] substitution Table 1 examines whether substituting the localized [STATE] representation leads to observable downstream changes in decision behav-

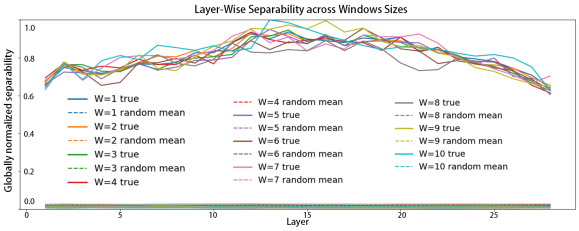


Figure 4: Random-label control for layer-wise separability across window sizes. Solid lines: true labels. Dashed lines: shuffled-label mean across repeated shuffles. Shaded regions: $\pm 2\sigma$ across shuffles. For each layer and feature-window size w, the separability score is the Top-K mean of the absolute paired Cohen's d values across feature-window positions. Scores are jointly min–max normalized using the union of the true-label values and the shuffled-label mean $\pm 2\sigma$ bounds across all feature-window sizes.

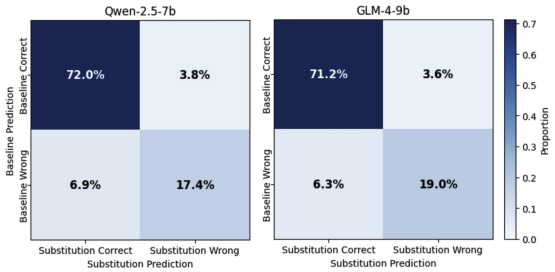


Figure 5: Base-to-substitution accuracy transition matrices. Each matrix shows the proportion of examples transitioning between baseline and substituted predictions, aggregated over both framings. Rows indicate baseline outcomes (Correct / Wrong), and columns indicate outcomes after [STATE] substitution. The four cells correspond to Correct→Correct, Correct→Wrong, Wrong→Correct, and Wrong→Wrong transitions. Results are shown for Qwen-2.5-7B (left) and GLM-4-9B (right).

ior. Greedy decoding removes sampling randomness from the BASE→SUB comparison. Across both models, benchmarks, and framings, [STATE] substitution consistently changes the model's prediction outcomes. In particular, both framing-specific accuracy and cross-framing Jaccard overlap increase under substitution. These changes provide behavioral evidence that the substituted [STATE] activations affect decision-related computation in addition to exhibiting linear separability. More fine-grained accuracy breakdowns and ensemble comparisons are provided in Appendix C.1.

Framing asymmetry under intervention Figure 5 summarizes the accuracy transitions induced by [STATE] substitution for both models. Each matrix reports how predictions shift between correct and incorrect outcomes when replacing the original decision state with its counterpart from the other framing.

Setting	ACC (SUP)	ACC (ELIM)	Jaccard
Qwen-2.5-7B			
MMLU-17 (base)	65.52	65.38	75.46
MMLU-17 (Sub)	**66.53**	**66.31**	**76.63**
MedQA-CH (base)	77.12	74.36	78.10
MedQA-CH (Sub)	**80.26**	**77.45**	**81.58**
GLM-4-9B			
MMLU-17 (base)	66.47	68.74	72.96
MMLU-17 (Sub)	**69.21**	**72.36**	**78.21**
MedQA-CH (base)	75.91	73.54	78.67
MedQA-CH (Sub)	**76.82**	**77.92**	**82.48**

Table 1: **Behavioral outcomes under cross-framing [STATE] substitution.** Base denotes inference with the aligned [STATE]-augmented prompt but without activation replacement. Sub denotes cross-framing replacement over the selected layer window W. For ACC (SUP), the recipient run uses a SUP prompt and receives the cached [STATE] activation from its paired ELIM prompt; ACC (ELIM) uses the reverse direction. Each value is computed by within-framing strict-majority voting over two deterministic prompt realizations, with ties counted as incorrect. SUP and ELIM predictions are not combined. Jaccard is the intersection-over-union of the question-index sets answered correctly under SUP and ELIM. All values are percentages; higher values indicate higher accuracy or greater overlap between successful instances. All values are percentages; higher is better. Per-task MMLU accuracy is reported in Table C.3.

Across both Qwen-2.5-7B and GLM-4-9B, substantial off-diagonal mass shows that substitution systematically alters model behavior. A non-trivial fraction of previously incorrect predictions becomes correct after substitution, while a smaller fraction of correct predictions degrades. The imbalance between these transitions reveals an asymmetric behavioral effect across the two framing directions. Additional analysis in Appendix C.2 shows that [STATE] substitution more often increases than decreases eliminated-incorrect mass, suggesting enhanced error-elimination determinacy.

5.3 EQ3: Interface, Content, and Position Specificity

Ablation settings We conduct a series of controlled ablations to isolate which aspects of the intervention are responsible for the observed be-

havioral effects. For each ablation, we compare the intervened output against the model's original greedy-decoding output without activation intervention. All interventions are applied within the same layer window W identified in Section 5.1. The ablations are organized along three orthogonal dimensions:

1. **Intervention object** We compare StateSwap at [STATE] with a final-content-token control that applies the same swap to the original prompt's final content token without [STATE]. This comparison measures interface specificity.

2. **Substitution content** We vary what replaces the original [STATE] representation, comparing structured cross-framing substitutions with content-agnostic controls. This comparison measures the dependence of model behavior on substitution content.

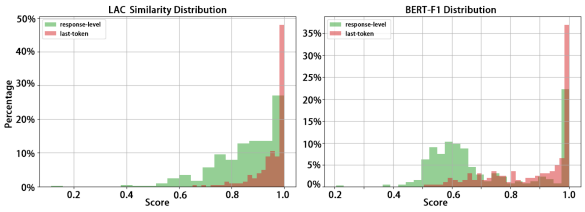
3. **Interface position** We move [STATE] from its canonical final position to earlier positions relative to the question and option segments while keeping the substitution operation and layer window fixed. This comparison measures position dependence.

Detailed implementations of each ablation, including zeroing, random replacement, framing-conditioned substitution, and alternative interface positions, are provided in Appendix E.

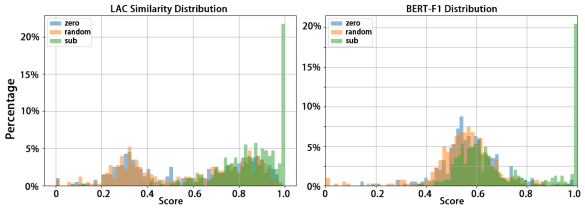
Interface specificity: [STATE] versus final-content-token substitution We assess whether the observed effects depend on the intervention location or can be reproduced by modifying late-stage token representations more generally. We apply the same substitution procedure to the original final content token, while keeping the intervention window and decoding configuration fixed.

As shown in Figure 6a, substitution at the final content token produces outputs that remain close to the baseline, with similarity scores concentrated near 1.0. Under the tested conditions, the [STATE] interface is therefore more effective than the final-content-token control at altering model behavior. Quantile summaries for both conditions are reported in Table E.3.

Effect of substitution content We examine how the nature of the substituted representation affects model behavior. Figure 6b reports response-level



(a) Interface specificity. Substituting [STATE] vs. the final content token under the same window W and greedy decoding.



(b) Ablations at [STATE]. Zeroing and random replacement vs. cross-framing substitution.

Figure 6: Response-level stability controls. (a) shows that late-position content-token states are substantially less effective as intervention targets than the [STATE] interface. (b) distinguishes unstructured residual injections (zero/random), which degrade stability, from structured cross-framing substitution, which preserves high-similarity regimes.

similarity between the baseline output and outputs produced under different intervention types, measured using length-aware cosine similarity (LAC) and BERTScore-F1. Quantile statistics for both metrics are provided in Table E.2.

Unstructured interventions (e.g., zeroing or random replacement) substantially reduce response similarity, whereas cross-framing substitution preserves high similarity across most examples. The distinct response regimes associate StateSwap's effects with the content of the substituted activation.

[STATE] token ablation We also evaluate the effect of inserting the [STATE] token on MedQA-CH accuracy by comparing the original prompt with an otherwise identical prompt that includes the token. Adding [STATE] changes accuracy negligibly across both models (Appendix E.1), separating token insertion from the subsequent activation-substitution effect.

Position sensitivity under misaligned substitution We further examine whether cross-framing substitution depends on the structural placement of [STATE]. More than 95% of generations under the tested misaligned placements either become incoherent or fail to produce a valid task-relevant response; the evaluated configurations, failure cri-

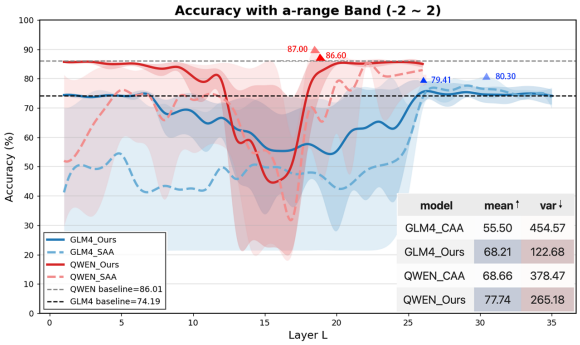


Figure 7: Accuracy under activation steering across intermediate layers. Solid lines denote steering directions derived from the dual-framing contrast between SUP and ELIM prompts, while dashed lines denote the original CAA baseline. Red and blue curves correspond to Qwen-2.5-7B and GLM-4-9B, respectively. Horizontal dashed lines indicate unsteered baseline accuracies.

terion, and canonical-position comparison are provided in Appendix E.3.

5.4 Dual Framing as a General Intervention Contrast

The preceding experiments characterize StateSwap within the dual-framing design. We next evaluate the SUP–ELIM contrast independently of the introduced interface by removing [STATE] and instantiating the same contrast in a population-level CAA-style mean-difference steering framework. This analysis compares the direction derived from matched SUP–ELIM prompts with the original CAA direction under otherwise matched settings.

As shown in Figure 7, the dual-framing direction produces systematic layer-dependent effects in both models even without the [STATE] token. Detailed aggregate statistics and complete layer-wise results, including comparisons with the original CAA direction and failures under strong coefficients, are provided in Appendix C.3.

6 Discussion

The choice of SUP and ELIM is central to what can be inferred from StateSwap. In preliminary analyses, we considered other complementary prompt pairs, including strictly translated Chinese and English instructions and prompts requesting a direct answer versus explicit reasoning. Such diversity can benefit prediction aggregating alternative reasoning paths improves accuracy (Wang et al., 2022; Lin et al., 2024). Language contrasts remained strongly classifiable from shallow through deep layers, plausibly because they introduce pervasive lexical and syntactic differences. Direct-answer and reasoning-first prompts were most distinguishable in intermediate and late layers, potentially reflecting different response strategies and generation plans. In both cases, a classifier could exploit factors broader than the decision distinction of interest.

The SUP–ELIM contrast was selected to reduce these confounds. The paired prompts preserve the question, options, language, decoding procedure, and answer key while changing whether the model identifies the correct option or eliminates the incorrect ones. Although this does not guarantee that framing is the only information encoded at [STATE], it supports a more specific conclusion: logically equivalent decision objectives can induce distinguishable intermediate representations whose exchange affects downstream predictions.

Additional, SUP and ELIM are broad observational labels that may each contain finer-grained states associated with domain, uncertainty, response strategy, or required deliberation. A model facing a medical question, for example, may adopt a more cautious state, potentially expressed through longer chains of thought, while remaining within the same SUP or ELIM class. Our analysis marginalizes over such within-class structure because it asks whether one controlled framing distinction is detectable and behaviorally relevant. Hierarchical or multi-factor analyses could instead separate framing from domain, confidence, and reasoning strategy, and test which finer-grained states remain independently substitutable.

7 Conclusion

StateSwap provides a training-free framework for analyzing framing-sensitive representations through a dedicated residual-stream interface. Across two evaluated LLMs, support- and elimination-oriented prompts induce separable intermediate-layer [STATE] activations. Exchanging these activations systematically changes downstream predictions, while the controls distinguish structured substitution from token insertion, final-content-token replacement, unstructured perturbations, and prompt ensembling. These findings establish the behavioral relevance of framing-conditioned activations within the tested models, multiple-choice framing protocol, and deterministic decoding setting.

Limitations

We conduct the experiments in a controlled setting with four-option multiple-choice questions and deterministic greedy decoding. This design helps isolate activation-level effects from sampling variance and makes BASE-to-SUB differences easier to attribute to the intervention. However, it also limits the scope of our claims. We do not evaluate whether the same framing-sensitive states persist under sampling-based decoding, longer multi-step generation, or open-ended tasks.

The current formulation is also tied to the structure of multiple-choice questions. In particular, SUP and ELIM prompts define complementary option sets, and our evaluation relies on parsing outputs into correct and incorrect options. Extending StateSwap beyond MCQs would require paired prompt formulations with a shared evaluation target, as well as task-specific agreement metrics for open-ended outputs.

Our method further relies on an untrained [STATE] token inserted at a fixed position. Although different random initializations of this token produce negligible variation in our experiments, the inserted token may still introduce subtle distributional or representational shifts. We therefore do not claim that the observed effects are independent of the interface design or intervention site.

StateSwap also assumes white-box access to model activations and parameters, leaving its applicability to black-box models untested. Majority voting over paraphrased prompts is used only to reduce prompt-realization noise during evaluation, and is not part of the intervention method itself.

Ethical Statements

This work does not involve human subjects, user studies, or the collection of personal data. All experiments are conducted offline on publicly available multiple-choice benchmarks, including MMLU and MedQA-CH, using openly released language models. The generated responses are synthetic benchmark outputs and are not associated with real individuals.

StateSwap is proposed as a diagnostic method for studying framing-sensitive internal representations under controlled inference settings. Although the intervention changes intermediate activations, our experiments are limited to semantically equivalent multiple-choice prompts and deterministic decoding. We do not present the method as a deployed alignment or decision-support system.

A potential risk of activation-level interventions is that similar techniques could be adapted for broader model steering, including biased or undesired behavioral changes. We therefore restrict our claims to offline analysis and report ablations to distinguish structured state substitution from generic perturbations. Since one benchmark is medical, we emphasize that the results should not be interpreted as evidence of clinical reliability or suitability for real-world medical decision making. Any use in high-stakes settings would require separate domain validation and safety evaluation.

We used AI-assisted tools responsibly to support language editing and manuscript preparation; all research ideas, analyses, results, and conclusions were developed and verified by the authors, who take full responsibility for the content of this paper.

References

Nishant Balepur, Shramay Palta, and Rachel Rudinger. 2024. It's not easy being wrong: Large language models struggle with process of elimination reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10143–10166.

Xi Chen, Aske Plaat, and Niki van Stein. 2025. How does chain of thought think? mechanistic interpretability of chain-of-thought reasoning with sparse autoencoding. *arXiv preprint arXiv:2507.22928*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Qihang Fu, Yongbin Qin, Ruizhang Huang, Yanping Chen, Yulin Zhou, and Lintao Long. 2025. Exclusion of thought: Mitigating cognitive load in large language models for enhanced reasoning in multiple-choice tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21673–21686.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Feijuan He, Han Lai, Jiaqi Liu, Bo Wang, Haoran Chen, Haohan Liu, and Chenxi Zhang. 2025. Solving mathematical problems using large language models: A survey. *Data Intelligence*, 7(4):907–946.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025. Programming refusal with conditional activation steering. In *International Conference on Learning Representations*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. 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. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. 2025. Fairsteer: Inference time debiasing for llms with dynamic activation steering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11293–11312.

Lei Lin, Jiayi Fu, Pengli Liu, Qingyang Li, Yan Gong, Junchen Wan, Fuzheng Zhang, Zhongyuan Wang, Di Zhang, and Kun Gai. 2024. Just ask one more time! self-agreement improves reasoning of language models in (almost) all scenarios. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3829–3852.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68. Association for Computational Linguistics.

Wenquan Lu, Yuechuan Yang, Kyle Lee, Yanshu Li, and Enqi Liu. 2025. Latent chain-of-thought? decoding the depth-recurrent transformer. *arXiv preprint arXiv:2507.02199*.

Chenkai Ma and Xinya Du. 2023. Poe: Process of elimination for multiple choice reasoning. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 4487–4496.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *CoRR*, abs/2209.11895.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522. Association for Computational Linguistics.

{Alexander Matt} Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. 2024. Activation addition: Steering language models without optimization. Working paper, arXiv.org, United Kingdom.

Mingsen Wang, Zhaoqiang Li, Xuegang Zhao, and Qihan Guo. 2026. Eliminate-then-select: A human-centric reasoning framework for educational question answering with llms. *Information Processing & Management*, 63(2):104422.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yifan Zhang, Wenyu Du, Dongming Jin, Jie Fu, and Zhi Jin. 2025. Finite state automata inside transformers with chain-of-thought: A mechanistic study on state tracking. *arXiv preprint arXiv:2502.20129*.

Qianli Zhao and Mei Zhang. 2025. Elimination-based reasoning with llm for multiple-choice educational question answering. *Journal of King Saud University Computer and Information Sciences*, 37(7):204.

Zhenhao Zhu, Bulou Liu, Qingyao Ai, and Yiqun Liu. 2025. Option-id based elimination for multiple choice questions. *arXiv preprint arXiv:2501.15175*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Methodological Details

A.1 Notation Summary

For clarity, we summarize the StateSwap-specific notation below. Standard Transformer notation follows Section 3.1; arguments (I, q) are omitted when they are clear from context.

Term / Symbol	Object (this paper)
framing and instance	$I \in \{\text{SUP}, \text{ELIM}\}$; $\bar{I}$ is the complementary framing and q is the shared question
framing-conditioned input	$\mathbf{x}_I(q) = \text{Concat}(I, q, \texttt{[STATE]})$
interface position	$t_S \triangleq T$, the final position in $\mathbf{x}_I(q)$
interface state	$\mathbf{s}^{(l)}(I, q) = \mathbf{h}^{(l)}_{t_S}(I, q) \in \mathbb{R}^d$
intervention region	contiguous $W \subseteq \{1, \ldots, L\}$
feature-window indices	length w, start $j \in \{0, \ldots, d - w\}$, and evaluated size set $\mathcal{W} \subseteq \{1, \ldots, d\}$
feature-window slice	$\mathbf{s}^{(l)}_{[j:j+w)} \in \mathbb{R}^w$
window-level diagnostic	$d_l(j, w)$, the paired Cohen's d after scalarization; d_l^{rand} and d_l^{md} denote the two diagnostics
per-size layer intensity	$\tilde{d}_l(w)$, the mean of the K largest values in $\{\lvert d_l(j, w)\rvert\}_{j=0}^{d-w}$
aggregated layer intensity	$\tilde{d}_l = \lvert\mathcal{W}\rvert^{-1} \sum_{w \in \mathcal{W}} \tilde{d}_l(w)$
state substitution	$\mathbf{s}^{(k)}(I, q) \leftarrow \mathbf{s}^{(k)}_{\text{cache}}(\bar{I}, q)$ for $k \in W$

A.2 Construction and Alignment of the [STATE] Interface

We introduce [STATE] as an additional special token in the tokenizer vocabulary. Concretely, we add [STATE] via the tokenizer interface: `tokenizer.add_special_tokens(·)`. After extending the tokenizer vocabulary, we resize the model's input embedding matrix to match the updated vocabulary size. All experiments are inference-only: model parameters are not updated, and the embedding vector corresponding to [STATE] remains at its initialization value throughout.

Token-Level Alignment Our methodology requires [STATE] to occupy the same absolute token index t_S across ELIM and SUP prompts for the same question q. We ensure this by token-level alignment of the instruction segment. Let the full prompt be

$$\mathbf{x}_I(q) = \text{Concat}(I, Q, \texttt{[STATE]}).$$

We construct the instruction segment so that it has the same token length for all instructions under the model tokenizer. For practical guidance, please refer to Appendix B.3.

Algorithm 1 Tokenize and Align Instructions

Require: Instructions list $\mathcal{I} = \{I_1, I_2, \ldots, I_n\}$
Ensure: Aligned token lists $\mathcal{T}_{\text{aligned}}$, padding sizes $\mathcal{P}$, target instruction length L_{target}

```
Initialize empty lists T_raw and L_base
for each instruction I in 𝓘 do
    tokens ← Tokenize(I)
    Append tokens to T_raw
    Append |tokens| to L_base
end for
L_target ← max(L_base)
Initialize empty lists T_aligned and 𝓟
for i ← 1 to n do
    pad_i ← L_target − L_base[i]
    aligned_i ← Concat([PAD]^{pad_i}, T_raw[i])
    Append aligned_i to T_aligned
    Append pad_i to 𝓟
end for
return T_aligned, 𝓟, L_target
```

Because the question block and the final [STATE] token are identical in both framings, equalizing the instruction-segment length guarantees that the absolute index t_S of [STATE] is matched between ELIM and SUP for each paired question q.

A.3 Window Diagnostics and Candidate Region Selection

For window size w, we use start indices $j \in \{0, \ldots, d - w\}$ and the slice notation $[j : j + w)$. For a scalarization $\phi : \mathbb{R}^w \to \mathbb{R}$, define paired differences for each question i:

$$\Delta_i^{(l,j,w)} = \phi(\mathbf{s}^{(l)}_{\text{ELIM},i,[j:j+w)}) - \phi(\mathbf{s}^{(l)}_{\text{SUP},i,[j:j+w)}).$$

We report the paired Cohen's d statistic:

$$d_l(j, w) = \frac{\overline{\Delta^{(l,j,w)}}}{\text{std}(\Delta^{(l,j,w)}) + \epsilon}, \tag{1}$$

where the mean and standard deviation are computed over $\{\Delta_i^{(l,j,w)}\}_{i=1}^N$.

Candidate Region Selection We compute layer-level intensities from absolute window-level separability scores using a Top-K mean aggregation. Because both diagnostics are reported in experiments, the same procedure is applied independently to $d_l^{\mathrm{rand}}(j, w)$ and $d_l^{\mathrm{md}}(j, w)$.

For a given diagnostic $d_l(j, w)$ and window size w, we define

$$\tilde{d}_l(w) = \mathrm{TopKMean}\Big(\{|d_l(j, w)|\}_{j=0}^{d-w};\, K\Big),$$

where $\mathrm{TopKMean}(\cdot; K)$ averages the K largest values. Unless stated otherwise, we further average over a small set of window sizes $\mathcal{W}$:

$$\tilde{d}_l = \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \tilde{d}_l(w).$$

Layers are then ranked by $\tilde{d}_l$. To determine a threshold for selecting candidate layers, we rely on an empirical consistency criterion guided by an independent PCA analysis. Specifically, we observe that layers whose $\tilde{d}_l$ values exceed approximately 80% of the maximum separability score consistently coincide with the region showing the clearest separation in the PCA visualization. Based on this observation, we select layers satisfying

$$\tilde{d}_l \geq 0.8 \cdot \max_{l'} \tilde{d}_{l'}$$

For each evaluated model, the layers satisfying this criterion form a single contiguous interval, which we denote by W. This threshold is not tuned for downstream performance. It is fixed across models and datasets and serves only as a diagnostic criterion for localizing layers that exhibit stable, representation-level differences between the two framings.

B Experimental and Evaluation Details

B.1 Datasets and Data Splits

We evaluate our approach on two four-option multiple-choice benchmarks: MedQA-CH and a reasoning-focused subset of MMLU. All downstream inference and reported task evaluations are conducted on the official test splits. A separate set of 50 examples from the training split is used only to localize the candidate intervention window; these examples do not enter the reported test evaluation. Appendix D.4 provides the full calibration protocol and a larger-scale bootstrap robustness analysis.

MedQA-CH MedQA-CH is the Chinese subset of MedQA, a medical question answering benchmark designed to evaluate professional-level clinical reasoning. It consists of 1,015 multiple-choice questions covering clinical knowledge and diagnostic reasoning. All questions are presented in Chinese and evaluated independently from English benchmarks.

MMLU (Reasoning-Focused Subset) We construct a reasoning-oriented subset of MMLU covering 17 subject categories that emphasize logical reasoning, mathematical thinking, legal analysis, ethical judgment, and domain-specific inference. The selected subsets span formal logic, mathematics, law, philosophy, medicine, and social sciences. All MMLU evaluations are conducted in English using the official test splits.

Table B.1 summarizes the detailed dataset composition. In total, our evaluation comprises 5,622 multiple-choice questions. For bilingual evaluation, English and Chinese benchmarks are tested separately rather than mixed within a single evaluation setting.

Dataset / Subset	**# Questions**
MedQA	
MedQA-CH	1,015
MMLU (Reasoning-Focused Subset)	
Formal Logic	125
Logical Fallacies	163
College Math Logic	100
Abstract Algebra (Proofs)	100
High School Math	270
Professional Law	1,534
Jurisprudence	108
Moral Disputes	346
Business Ethics	100
Philosophy	311
Sociology	201
World History	237
U.S. History	204
World Religions	171
Bioethics	272
Clinical Knowledge (Reasoning)	265
Medical Genetics	100
Total (MMLU subset)	4,607
Total (All datasets)	**5,622**

Table B.1: Dataset statistics. All datasets are evaluated using their official test splits. MedQA-CH is evaluated in Chinese, while MMLU subsets are evaluated in English.

B.2 Decoding and Inference Configuration

All experiments are conducted in a zero-shot setting using deterministic greedy decoding. Specifically, we set `do_sample=False` and

max_new_tokens=1024. As greedy decoding is used, sampling-related parameters such as temperature and top_k are not applicable and are ignored by the generation backend. All random seed values are 123.

The maximum context length is fixed across all experiments, and inputs exceeding this limit are truncated accordingly. Right padding is applied with a maximum of 64 padding tokens. All generations use the official chat templates provided by the respective models without modification.

For reproducibility, all runs are executed with fixed random seeds and deterministic computation settings. Inference is performed using torch_dtype=float16, with key–value caching enabled to ensure consistent decoding behavior across runs.

B.3 Detailed Prompt Templates

This section provides the prompt templates used in the experiments.

Prompt Components Each input prompt consists of three components: an instruction segment I, a question segment Q that contains the question text, and an option segment O that enumerates all candidate answers in a fixed format. In addition, a special interface token [STATE] is appended to the end of the prompt to mark the position at which the model is expected to produce its output. For models that employ a beginning-of-sequence token, the tokenizer's default behavior is preserved.

Instruction Templates We use two groups of framing-specific instruction templates with aligned task objectives. One group instructs the model to reason step by step and identify **correct** options (Group 1), while the other instructs the model to reason step by step and identify **incorrect** options (Group 2). Within each group, multiple paraphrased variants are used to reduce dependence on a specific wording. All instruction templates are fixed across datasets and models.

> **Group 1 - Sup prompt**
> Instruction: Proceed step by step: first list the complete chain of reasoning, then pick out **Correct** options. Question: [QUESTION]
> Instruction: Take a careful, step-by-step approach. First, reason in detail, then find **Correct** choices. Question: [QUESTION]

> **Group 2 - Elim prompt**
> Instruction: Proceed step by step: first list the complete chain of reasoning, then pick out **Wrong** options. Question: [QUESTION]
> Instruction: Take a careful, step-by-step approach. First, reason in detail, then find **Incorrect** choices. Question: [QUESTION]

For each framing, we instantiate two instruction templates, yielding two deterministic greedy predictions for SUP and two for ELIM. Unless otherwise specified, reported ensemble results aggregate these paired prompt realizations by strict-majority voting within the corresponding framing or StateSwap variant; a 1–1 tie is counted as an incorrect ensemble prediction. This protocol separates prompt-template variation from decoding randomness, since all runs use greedy decoding.

B.4 Robustness to [STATE] Initialization

The purpose of this experiment is to verify that StateSwap does not rely on incidental randomness from the untrained [STATE] token. Table B.2 shows that output similarity remains near-perfect across different random initializations, supporting the claim that [STATE] functions as a deterministic state interface rather than a source of stochastic variation.

B.5 Output Parsing and Evaluation Metrics

B.5.1 Parsing to Option Sets

To ensure consistency across languages and facilitate automatic evaluation, model outputs are post-processed into a standardized representation consisting of two fields: Correct options:[. . .] and Wrong options:[. . .]. This step preserves the semantic meaning of the original responses while enforcing a fixed, machine-readable output format.

Options Answer extraction Let $\mathcal{O} = \{A, B, C, D\}$ denote the option alphabet. We parse each field into a set $S \subseteq \mathcal{O}$ by extracting all uppercase letters in $\mathcal{O}$ from the field, ignoring punctuation and separators (e.g., commas, Chinese enumeration marks).

B.5.2 Metric Definitions

Accuracy (ACC) For each question with ground-truth correct option $y \in \mathcal{O}$, the prediction under a given projection is considered correct if $y \in S$,

Swap direction	Seed$_a$	Seed$_b$	LA-cos ↑		BERT-F1 ↑	
			mean	std	mean	std
Elim_Baseline ↓ *Sup_Baseline*	1	2	9.99999986589e-01	7.81e-08	9.99999991059e-01	5.65e-08
	1	3	9.99999986589e-01	7.81e-08	9.99999991059e-01	5.65e-08
	1	12	9.99999989569e-01	7.38e-08	9.99999991059e-01	5.65e-08
	1	123	9.99999989569e-01	7.38e-08	9.99999991059e-01	5.65e-08
	2	3	9.99999986589e-01	7.81e-08	9.99999991059e-01	5.65e-08
	2	12	9.99999989569e-01	7.38e-08	9.99999991059e-01	5.65e-08
	2	123	9.99999989569e-01	7.38e-08	9.99999991059e-01	5.65e-08
	3	12	9.99999989569e-01	7.38e-08	9.99999991059e-01	5.65e-08
	3	123	9.99999989569e-01	7.38e-08	9.99999991059e-01	5.65e-08
	12	123	9.99999991059e-01	7.34e-08	9.99999991059e-01	5.65e-08
Sup_Baseline ↓ *Elim_Baseline*	1	2	9.99999982119e-01	7.04e-08	9.99999976158e-01	8.19e-08
	1	3	9.99999982119e-01	7.04e-08	9.99999976158e-01	8.19e-08
	1	12	9.99999984354e-01	6.62e-08	9.99999976158e-01	8.19e-08
	1	123	9.99999984354e-01	6.62e-08	9.99999976158e-01	8.19e-08
	2	3	9.99999982119e-01	7.04e-08	9.99999976158e-01	8.19e-08
	2	12	9.99999984354e-01	6.62e-08	9.99999976158e-01	8.19e-08
	2	123	9.99999984354e-01	6.62e-08	9.99999976158e-01	8.19e-08
	3	12	9.99999984354e-01	6.62e-08	9.99999976158e-01	8.19e-08
	3	123	9.99999984354e-01	6.62e-08	9.99999976158e-01	8.19e-08
	12	123	9.99999985844e-01	6.58e-08	9.99999976158e-01	8.19e-08

Table B.2: Detailed robustness results for random `[STATE]` initialization. We sample 40 instances and rerun inference with five random seeds (1, 2, 3, 12, 123), keeping all other settings fixed. For each swap direction and seed pair, we report the mean and standard deviation of LA-cos and BERT-F1 computed over the 40 instances, showing negligible variation across seeds.

where S denotes the set of options predicted as correct by the model. Formally, accuracy is defined as:

$$\mathrm{ACC} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[y_i \in S_i\right],$$

where N is the total number of questions, y_i is the ground-truth correct option for the i-th question, S_i denotes the predicted set of correct options, and $\mathbb{I}[\cdot]$ is the indicator function.

Jaccard Let C_{Sup} and C_{Elim} denote the sets of question indices for which the model predicts the correct answer under the support- and elimination-oriented framings, respectively. For each question, correctness is determined solely based on the `Correct options` field in the normalized output, as described above.

We define cross-framing agreement using the Jaccard index over these two sets:

$$J = \frac{|C_{\text{Sup}} \cap C_{\text{Elim}}|}{|C_{\text{Sup}} \cup C_{\text{Elim}}|}.$$

Intuitively, a question contributes to the intersection if and only if the correct option appears in the `Correct options` list under both framings. For example, if the SUP output is `Correct: [1,0,0,0]` and the ELIM output is also `Correct: [1,0,0,0]`, the question is counted as overlapping; whereas if the two outputs differ (e.g., `[1,0,0,0]` vs. `[0,1,0,0]`), it is not. This metric therefore measures agreement in successful decisions between the two framings, rather than overlap between option sets for individual questions.

Relation Between ACC and Jaccard While ACC measures whether the model produces the correct answer under a single framing, the Jaccard index captures the consistency of correct decisions between the two framings. In particular, ACC reflects per-framing performance, whereas Jaccard evaluates cross-framing agreement conditioned on correctness. A high ACC with a low Jaccard score indicates that the model can answer correctly under either framing but does so inconsistently, while a high Jaccard score reflects stable decision-making between the two framings. The two metrics therefore provide complementary views of model behavior.

Length-Aware Cosine Similarity Given a baseline response r and its perturbed counterpart $\tilde{r}$, we compute the cosine similarity between their sentence embeddings. To account for potential length mismatches, we apply a length-aware scaling factor:

$$\mathrm{LAC}(r, \tilde{r}) = \cos(\mathbf{e}_r, \mathbf{e}_{\tilde{r}}) \cdot \frac{\min(|r|, |\tilde{r}|)}{\max(|r|, |\tilde{r}|)},$$

where $\mathbf{e}_r$ and $\mathbf{e}_{\tilde{r}}$ denote normalized sentence embeddings, and $|r|$ denotes the character length of the response.

This metric captures semantic similarity while penalizing excessive length divergence caused by perturbations. The length-aware scaling term is introduced to account for cases where perturbations to the hidden state cause the model to generate excessively long or verbose outputs that are only weakly related to the original response. In such cases, standard cosine similarity alone may overestimate semantic consistency due to shared topical content. By down-weighting similarity scores when large length discrepancies occur, this metric penalizes uncontrolled or drifting generations and better reflects whether the core semantic content is preserved.

BERTScore We additionally compute BERTScore between the perturbed and baseline outputs, reporting F1 scores. BERTScore measures token-level semantic similarity using contextualized embeddings and is sensitive to both content preservation and phrasing changes. In our experiments, the baseline output is treated as the reference and the perturbed output as the candidate.

Relation Between LAC and BERTScore These two metrics are computed by comparing each perturbed output against its corresponding baseline output without any state modification. They are therefore intended to measure the sensitivity of the generation to perturbations applied to the `[STATE]` representation, rather than absolute output quality.

Length-aware cosine similarity reflects global semantic consistency between the perturbed and baseline responses, while BERTScore captures fine-grained lexical and contextual deviations. Importantly, higher scores do not necessarily indicate better performance: values close to 1 suggest that the perturbation has little effect on the output, whereas lower scores indicate that modifying the internal state leads to substantial changes in the generated content. Extremely low scores may further reflect degenerate or incoherent outputs caused by the perturbation.

B.5.3 Error-Elimination Metric

For instance i and framing $p \in \{\text{SUP}, \text{ELIM}\}$, let $W_i^{(p)} \subseteq \mathcal{O}$ be the options parsed from the `Wrong options` field. We define the eliminated-incorrect count as

$$E_i^{(p)} = \left|W_i^{(p)} \setminus \{y_i\}\right| \in \{0, 1, 2, 3\}.$$

Let $E_{i,\text{base}}^{(p)}$ and $E_{i,\text{sub}}^{(p)}$ denote the values before and after substitution. We define

$$\Delta_i^{(p)} = E_{i,\text{sub}}^{(p)} - E_{i,\text{base}}^{(p)}.$$

An instance is counted as a RIGHT SHIFT if $\Delta_i^{(p)} > 0$, and as a LEFT SHIFT if $\Delta_i^{(p)} < 0$. We report the proportions of right- and left-shifted instances under each framing and their macro-average across SUP and ELIM. A higher right-shift frequency indicates stronger incorrect-option elimination.

C Additional Behavioral Results

C.1 Ensemble and Per-Task Results

Our main reported results are based on deterministic ensemble inference rather than a single prompt realization. Specifically, each single-framing result is averaged over two greedy runs with different prompt realizations. To further examine whether the observed gains can be explained by simple aggregation effects, we compare several ensemble variants: single-framing baselines (SUP-only and ELIM-only), StateSwap variants (SUP(ELIM State) and ELIM(SUP State)), a naive cross-framing ensemble (SUP+ELIM), framing-plus-StateSwap ensembles, and a dual StateSwap ensemble. As shown in Table C.1, ensembles involving swapped `[STATE]` predictions consistently outperform ensembles formed only from the original framings. In particular, the dual StateSwap ensemble achieves the best accuracy for both Qwen-2.5-7B and GLM-4-9B, suggesting that StateSwap transfers useful decision-state information beyond ordinary variance reduction from ensembling.

Since combining two framings can produce an even number of votes, we further evaluate a weighted majority rule for tie-breaking. We use a linear vote of the form $\text{vote} = \text{SUP} + \alpha \cdot \text{ELIM}$ and search over α for 30 trials, reporting the best achievable accuracy for each ensemble. As shown in Table C.2, even under this optimized aggregation scheme, the dual StateSwap ensemble remains stronger than direct SUP–ELIM ensembling. This indicates that the StateSwap gains cannot be explained away by a better voting heuristic alone.

Method	Qwen-2.5-7B Acc.	GLM-4-9B Acc.
2-Run Ensemble		
ELIM-only	0.7436	0.7354
ELIM(SUP State)	**0.7745**	**0.7792**
SUP-only	0.7712	0.7591
SUP(ELIM State)	**0.8026**	**0.7682**
4-Run Ensemble		
SUP + SUP(ELIM State)	0.7939	0.7692
ELIM + ELIM(SUP State)	0.7844	0.7861
SUP + ELIM	0.8020	0.7890
SUP(ELIM State) + ELIM(SUP State)	**0.8156**	**0.7949**

Table C.1: Ensemble ablation results. The 2-run setting reports deterministic greedy ensembles within a single framing or StateSwap condition. The 4-run setting combines two 2-run groups. StateSwap-based ensembles outperform the naive SUP+ELIM ensemble on both models.

Method	Qwen-2.5-7B Acc.	GLM-4-9B Acc.
4-Run Ensemble with Weighted Voting		
SUP + SUP(ELIM State)	0.8008	0.7732
ELIM + ELIM(SUP State)	0.7943	0.7861
SUP + ELIM	0.8059	0.7939
SUP(ELIM State) + ELIM(SUP State)	**0.8225**	**0.8018**

Table C.2: Weighted-vote ensemble ablation. We search over the linear tie-breaking weight α for 30 trials and report the best accuracy. StateSwap-based ensembling remains stronger than direct SUP+ELIM aggregation.

Per-task MMLU accuracy Table C.3 reports per-task accuracy under the SUP and ELIM framings. Compared with the Base setting, the Sub variant achieves consistent or improved performance across most MMLU categories.

C.2 Decision Determinacy Behavior

Beyond answer accuracy, we analyze whether [STATE] substitution alters the model's determinacy in eliminating incorrect options, using the error-elimination metric defined in Appendix B.5.3.

As shown in Table C.4, right shifts consistently occur more frequently than left shifts across models and datasets, indicating that [STATE] substitution more often strengthens the model's ability to eliminate incorrect options, rather than relaxing it.

This asymmetric shift aligns with the gains observed in ACC Elim and Jaccard (Table 1), suggesting that the observed accuracy improvements are driven by enhanced error-elimination capability, manifested as more decisive and systematic suppression of incorrect options, rather than random variation or superficial answer changes.

C.3 Detailed Steering Results

This experiment isolates how a population-level steering direction changes when the contrast is constructed from instructions rather than answer options. It contains no [STATE] token and does not perform instance-level state substitution.

Matched contrast construction For calibration example i, let p_i^{SUP} request the correct option, let p_i^{ELIM} request an incorrect option, and let y_i be the correct answer letter. We pair y_i with a deterministic incorrect letter $\widetilde{y}_i$, chosen as the first letter in $\{A, B, C, D\}$ that differs from y_i. Let $h_l(p, y)$ denote the residual-stream activation at the appended answer token y after layer l. Standard CAA holds the instruction fixed and changes the completion:

$$v_l^{\text{CAA}} = \frac{1}{N} \sum_{i=1}^{N} \left[h_l(p_i^{\text{SUP}}, y_i) - h_l(p_i^{\text{SUP}}, \widetilde{y}_i) \right]. \tag{2}$$

The dual-framing construction holds the completion fixed and changes the instruction:

$$v_l^{\text{DF}} = \frac{1}{N} \sum_{i=1}^{N} \left[h_l(p_i^{\text{SUP}}, y_i) - h_l(p_i^{\text{ELIM}}, y_i) \right]. \tag{3}$$

We use $N = 50$ calibration examples for both directions. The two contrasts share the positive pair (p_i^{SUP}, y_i). Their negative members are complementary counterfactuals: CAA preserves the instruction but forces an inconsistent option, whereas dual framing preserves the option but forces an inconsistent instruction–completion pair. Neither negative member is treated as a valid answer to its instruction. Both serve only to define a matched activation difference.

Steering and evaluation protocol We independently ℓ_2-normalize each layer-wise direction, giving $\widehat{v}_l = v_l / \|v_l\|_2$. For one layer at a time, we modify the residual stream at the answer boundary as

$$h_l \leftarrow h_l + a\widehat{v}_l, \qquad a \in \{-2, -1, 0, 1, 2\}. \tag{4}$$

Task	Qwen-2.5-7B				GLM-4-9B			
	Base-Elim	Sub-Elim	Base-Sup	Sub-Sup	Base-Elim	Sub-Elim	Base-Sup	Sub-Sup
abstract_algebra_proofs	63.27	61.62	63.64	66.00	65.31	65.31	59.60	70.71
bioethics	76.01	79.04	78.31	80.88	81.18	86.40	82.35	87.50
business_ethics	71.72	71.00	68.00	68.00	69.39	73.47	68.69	74.75
clinical_knowledge	75.19	74.05	69.96	69.81	74.90	78.03	73.86	76.23
college_math	70.97	69.47	61.05	63.00	63.04	72.34	58.95	61.46
formal_logic	64.46	63.64	62.81	63.93	61.29	68.25	64.80	65.08
hs_math	86.92	88.67	75.00	77.61	81.42	83.90	62.20	70.45
jurisprudence	71.30	71.96	69.44	68.52	80.37	79.63	75.93	75.00
logical_fallacies	81.51	80.13	81.87	84.28	77.07	85.19	83.23	83.44
medical_genetics	84.38	84.85	80.00	80.00	85.00	86.00	80.00	81.00
moral_disputes	61.70	62.39	63.85	65.51	69.01	72.09	63.58	65.90
philosophy	64.61	65.48	71.70	73.31	68.93	71.06	63.67	69.45
professional_law	47.22	48.92	50.20	50.36	54.46	57.67	53.82	55.71
sociology	73.06	71.94	71.28	71.43	76.02	84.26	72.73	75.00
us_history	85.05	85.20	80.20	82.14	84.92	89.11	85.93	87.13
world_history	81.62	79.24	84.55	82.63	84.85	86.38	86.38	86.44
world_religions	85.71	84.52	82.53	82.63	84.02	86.90	81.66	82.25

Table C.3: Extended per-task results on the MMLU benchmark. We report accuracy under the two framings (Elim and Sup), each evaluated with the original model (Base) and with [STATE] substitution (Sub). This table complements the main results by providing fine-grained, task-level comparisons. All values are percentages(%); higher is better.

Model	**Dataset**	Right Shift		Left Shift
Qwen-2.5-7B	MMLU	7.25%	>	7.17%
	MedQA	11.31%	>	9.29%
GLM-4-9B	MMLU	9.75%	>	7.52%
	MedQA	15.24%	>	11.39%

Table C.4: Directional changes in eliminated-incorrect mass from baseline to substitution (proportions over all instances). Right Shift indicates an increase in eliminated-incorrect mass, while Left Shift indicates a decrease.

Both methods use identical hooks, prompt templates at evaluation, coefficient values, and next-token scoring over $\{A, B, C, D\}$. We evaluate all 1,015 MedQA-CH test questions with Qwen-2.5-7B and GLM-4-9B. Accuracy at $a = 0$ is the unsteered reference and is omitted from Table C.5. We sweep Layers 1–26 for Qwen and Layers 1–35 for GLM. Consequently, the comparison changes only whether the calibration contrast perturbs the answer option or the instruction.

Aggregate comparison Averaged across the evaluated layers, the dual-framing direction retains higher accuracy than CAA at $a = -2$, -1, and 2. The respective dual-framing and CAA means are 59.03 vs. 37.16, 69.67 vs. 48.85, and 68.15 vs. 60.35 for GLM. They are 71.21 vs. 52.26, 81.61 vs. 75.97, and 74.77 vs. 62.90 for Qwen. At $a = 1$, the means are nearly identical: 74.05 vs. 74.07 for GLM and 83.37 vs. 83.52 for Qwen.

The clearest stability difference occurs at $a = -1$. Relative to CAA, the layer-wise range narrows from 50.54 to 19.90 points for GLM and from 60.89 to 28.08 points for Qwen. At $a = 1$, the range is similar for GLM (10.93 for CAA and 9.16 for dual framing) and narrower under dual framing for Qwen (27.00 vs. 12.41). Strong coefficients still cause sharp failures at individual layers. Higher retained accuracy also measures robustness to intervention, not greater steering strength by itself. The evidence therefore supports a more bounded layer-wise response under this matched protocol, rather than general superiority over CAA. Table C.5 reports every layer–coefficient result.

D Representation and Window Analyses

D.1 Additional Visualization of Layer-wise Separability

We visualize [STATE] representations using principal component analysis (PCA) to illustrate their layer-wise geometry under the two framings.

For Qwen-2.5-7B, we extract residual-stream activations at the [STATE] position using the same contiguous feature window ($w = 6$) as in the separability analysis. At each layer, PCA is applied independently to pooled representations from both framings, and the first three principal components are visualized.

Figure D.1 shows substantial overlap in early layers, clearer separation in intermediate layers, and reduced separation in later layers, matching the non-monotonic depth-wise trend in Figure D.2.

Layer	Two-framing				CAA			
	$a=-2$	$a=-1$	$a=1$	$a=2$	$a=-2$	$a=-1$	$a=1$	$a=2$
GLM-4-9B								
1	74.68	74.19	74.38	74.38	21.48	46.60	69.95	26.90
2	73.40	73.89	74.68	74.68	22.46	58.13	72.22	48.87
3	73.40	73.50	74.29	0.00	22.17	50.15	75.67	0.00
4	74.88	74.48	74.09	73.79	21.48	37.14	75.76	69.75
5	74.19	74.19	74.19	73.89	23.94	50.44	74.38	67.98
6	75.76	75.37	73.20	73.00	21.48	33.20	74.09	48.87
7	74.48	75.37	73.69	73.30	21.48	35.67	70.54	38.03
8	61.48	69.56	72.91	69.46	21.48	32.91	71.43	47.88
9	60.79	69.26	74.38	71.13	21.48	32.91	68.77	45.62
10	59.51	68.37	74.98	71.63	21.48	25.91	69.75	52.91
11	55.27	70.05	72.61	60.99	21.48	30.94	68.77	49.75
12	59.70	71.23	71.63	63.94	21.48	40.30	72.12	65.02
13	45.81	67.88	73.10	64.83	21.48	37.04	69.46	55.17
14	42.17	67.98	73.30	62.56	21.48	40.49	71.92	67.00
15	33.79	60.30	71.13	61.67	21.48	37.24	73.79	66.80
16	40.59	61.38	67.98	51.43	21.48	36.16	73.60	67.68
17	36.26	65.42	69.16	52.71	21.48	36.75	73.69	58.33
18	45.12	67.29	70.54	47.88	21.48	35.96	73.00	61.38
19	40.00	69.06	70.05	41.87	21.58	33.30	73.10	58.82
20	28.08	56.95	71.92	64.14	21.48	32.71	72.02	44.53
21	41.38	55.47	77.04	73.99	21.48	37.04	75.47	42.17
22	38.62	63.74	76.35	74.09	21.48	37.93	77.24	57.24
23	43.05	66.11	75.47	76.55	21.87	38.72	77.44	63.35
24	38.13	60.89	76.55	76.45	25.42	39.90	79.70	61.97
25	55.96	68.08	77.14	79.11	40.69	54.48	71.13	66.80
26	71.72	74.09	76.65	79.41	69.75	71.13	79.01	74.19
27	71.72	73.99	75.96	78.72	75.86	72.71	79.11	79.41
28	70.74	73.50	75.86	78.62	73.10	72.71	78.23	80.20
29	73.60	74.38	75.57	78.03	77.83	76.45	76.75	79.41
30	72.22	73.89	75.67	77.14	74.68	74.58	77.14	80.30
31	71.63	73.50	75.67	77.34	73.40	73.99	77.83	80.10
32	71.53	73.50	75.67	77.14	74.38	74.68	75.07	76.75
33	72.91	73.89	75.07	77.54	72.81	73.89	74.19	76.16
34	72.32	74.09	75.57	77.34	74.38	74.88	74.78	75.96
35	71.23	73.60	75.37	76.35	69.75	72.61	75.17	76.95
Qwen-2.5-7B								
1	85.52	85.71	86.21	85.42	29.66	67.19	81.28	29.85
2	85.81	86.11	85.81	85.12	37.93	67.78	79.90	35.76
3	86.21	85.91	85.71	85.02	33.20	78.13	82.76	62.17
4	86.31	86.11	84.33	83.35	44.04	80.79	84.43	73.89
5	85.62	85.62	84.93	83.94	59.51	83.74	83.98	77.34
6	85.22	85.42	85.52	84.24	58.42	82.76	82.56	73.69
7	84.33	85.12	84.53	82.17	54.58	81.67	83.94	74.98
8	81.18	84.63	85.32	82.56	38.82	66.40	83.94	74.19
9	83.25	84.33	85.71	82.27	53.69	79.80	85.12	76.45
10	73.89	84.14	84.04	80.49	54.78	80.10	85.02	71.23
11	73.69	83.35	82.17	77.93	57.34	81.48	84.73	75.07
12	70.25	81.38	82.56	82.17	57.14	80.39	85.42	70.64
13	26.40	72.12	76.45	63.25	47.39	72.71	84.73	65.62
14	24.53	71.82	77.44	61.38	21.87	76.45	83.55	39.80
15	25.62	61.38	77.83	22.76	22.86	68.18	85.71	45.62
16	21.97	58.23	76.95	23.84	21.48	57.44	80.30	23.94
17	39.70	73.50	74.09	23.35	22.86	25.81	60.00	24.43
18	56.65	83.25	84.63	77.24	55.96	70.25	86.31	59.31
19	76.85	83.25	86.50	84.53	45.42	77.14	86.50	52.32
20	84.73	85.81	85.42	84.63	82.86	83.94	87.00	61.28
21	84.63	85.81	85.62	84.93	83.84	85.32	86.70	48.08
22	86.01	86.01	84.53	83.94	83.55	85.81	86.50	83.65
23	86.50	86.31	84.73	84.24	68.57	85.42	86.50	86.70
24	86.11	86.11	85.42	85.02	69.66	86.21	84.73	83.45
25	86.60	85.62	85.42	84.83	75.17	86.70	84.43	81.58
26	83.84	84.93	85.71	85.52	78.23	83.65	85.42	84.43

Table C.5: Detailed MedQA-CH accuracy under mean-difference activation steering. DUAL holds the answer option fixed and contrasts the SUP and ELIM instructions. CAA holds the instruction fixed and contrasts the correct and paired incorrect options. Results use steering coefficients $a \in \{-2, -1, 1, 2\}$. The upper block reports GLM-4-9B results, while the lower block reports Qwen-2.5-7B results. The unsteered case $a = 0$ is omitted for clarity.

Layers	Cos ↓	IQR	p90	KL ↑	IQR	p90
3–5	0.130	0.070	0.275	1.634	1.319	6.832
13–15	0.108	0.085	0.295	1.000	0.943	3.745
11–20	0.108	0.066	0.257	1.487	1.198	6.253
25–28	0.131	0.069	0.275	1.636	1.317	6.832

Table D.1: Window comparison under representation substitution. Lower cosine distance indicates smaller global distortion in logit space; when paired with non-trivial KL shifts, we use it as a proxy for reduced collateral perturbation. IQR and p90 summarize dispersion and tail heaviness. Values are averaged across swap directions.

PCA is performed independently at each layer; hence the resulting axes are not comparable across layers.

D.2 Analysis of the Intervention Window

This subsection compares candidate layer windows using output-level statistics under representation substitution as a post-hoc assessment of the separability-selected intervention window.

Analysis Metrics We measure (i) cosine distance between the original and intervened logit vectors (global geometric change), and (ii) $\mathrm{KL}(q\|p)$ between the corresponding output distributions (probability mass reallocation). For each metric we report the mean and two stability summaries: IQR and the 90th percentile (p90), where larger p90 indicates heavier tails. p90 corresponds to the 90th percentile of per-example distances; larger p90 implies more extreme cases among the top 10% most affected examples.

Layer Windows Discussion Table D.1 aggregates results across swap directions (direction-specific trends are qualitatively consistent with the conclusions). Shallow (3–5) and late (25–28) layers behave nearly identically on both metrics, yielding larger cosine distances (0.130–0.131) and the largest KL values (1.634–1.636) with heavier KL tails (p90=6.832). This suggests strong distributional shifts accompanied by more extreme probability reallocations among the most affected examples.

The intermediate-layer peak (13–15) exhibits the largest dispersion and tail heaviness in cosine distance (IQR=0.085, p90=0.295) while producing the smallest KL magnitude and tails (KL mean=1.000, p90=3.745), indicating a mismatch between geometric logit changes and distributional reallocation under this window.

In contrast, the intermediate window (11–20) shows the most favorable stability–effect profile under our criteria: mean cosine distance is similar for 11–20 and 13–15 (both ≈ 0.108), but 11–20 achieves the smallest cosine dispersion and lightest cosine tails (IQR=0.066, p90=0.257), while maintaining substantial KL shifts (mean=1.487) and a lighter KL tail compared to 3–5/25–28 (p90 6.253 vs. 6.832). A smaller KL p90 implies less extreme probability reallocations among the top 10% most affected instances, which is desirable for a robust and predictable intervention interface.

Conclusion Overall, the separability-selected window 11–20 also provides the best stability–effect trade-off under output-level statistics: it avoids large and heavy-tailed geometric perturbations (cosine) while still inducing meaningful distributional changes (KL) without the heaviest tails observed at the layer-range extremes. This analysis is post hoc: the output-level statistics are not used to select the intervention window or tune it for task performance.

D.3 Cross-Model Separability Diagnostics at the [STATE] Interface

To complement the PCA visualization in Section 5.1, we report the full layer-wise separability diagnostics for both Qwen-2.5-7B and GLM-4-9B. For each model, we use the same paired Sup/Elim construction, compute Cohen's d over residual-stream features at the [STATE] position, and apply the same sliding-window procedure over feature dimensions. We report both the label-free and label-informed variants of the normalized separability heatmaps.

Figure D.2 shows that the two models exhibit a consistent layer-wise pattern. In both Qwen-2.5-7B and GLM-4-9B, separability is weak in early layers, becomes most pronounced in intermediate layers, and attenuates toward later layers. This cross-model consistency supports the conclusion that framing-dependent structure at the [STATE] interface is concentrated in an intermediate-layer region rather than uniformly distributed across network depth.

Following the same criterion used in the main text, we select model-specific intervention windows based solely on the observed separability patterns: layers 11–20 for Qwen-2.5-7B and layers 21–40 for GLM-4-9B. These windows are not tuned based on downstream task performance;

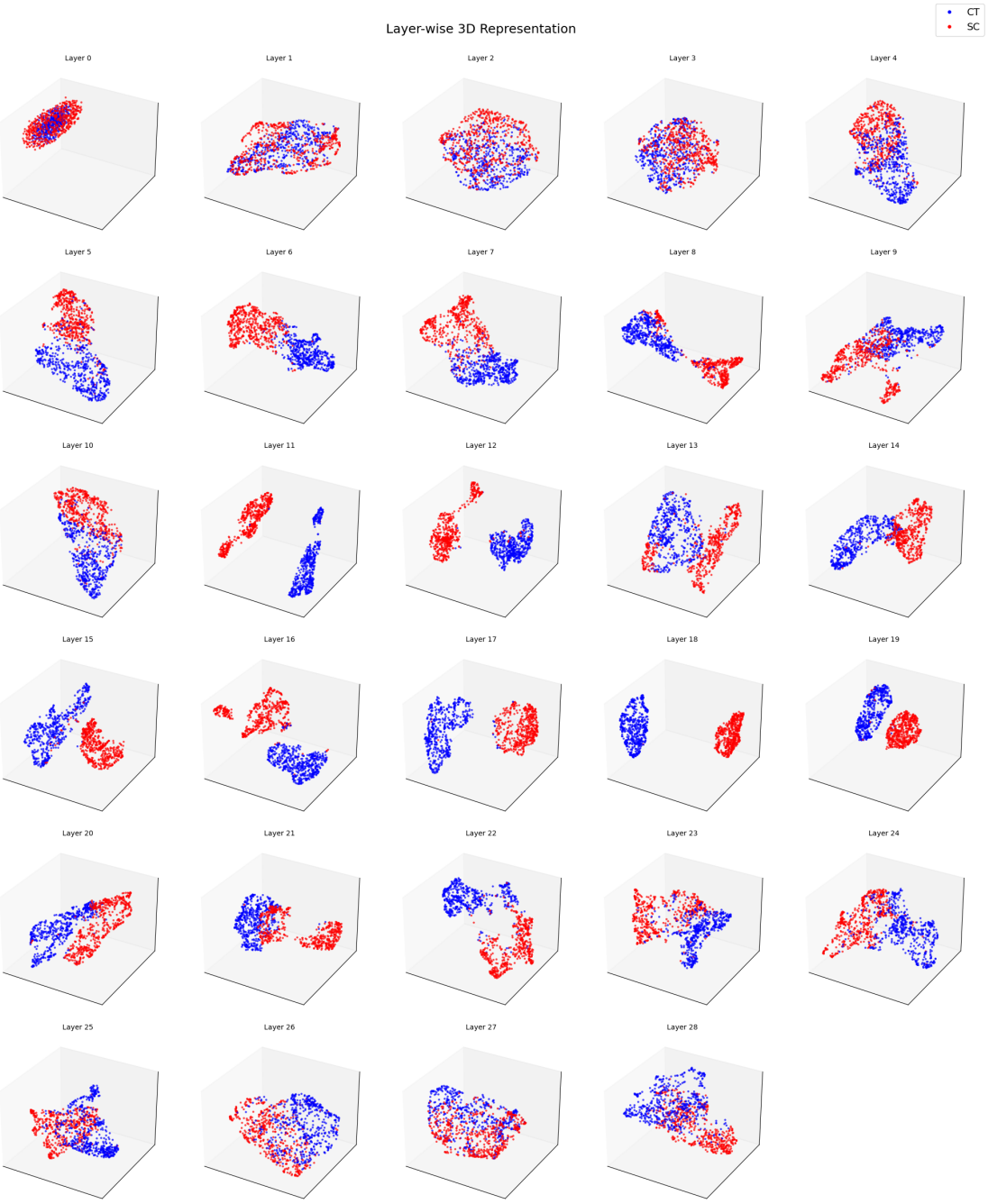


Figure D.1: **Layer-wise PCA of framing-conditioned [STATE] representations.** For each layer of Qwen-2.5-7B, PCA is applied to pooled representations using the same feature window ($w = 6$) as in the separability analysis. Points are individual instances colored by framing. Intermediate-layer representations show clearer separation than early- or late-layer representations, consistent with Figure D.2.

they are used in subsequent experiments to test whether [STATE] substitution within the separable intermediate-layer region produces systematic behavioral changes.

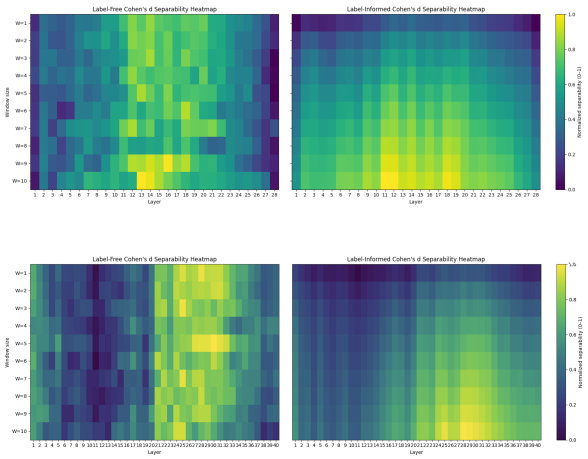


Figure D.2: Layer-wise normalized separability at the [STATE] position for Qwen-2.5-7B (top) and GLM-4-9B (bottom). For each model, the heatmaps report label-free Cohen's d and the corresponding label-informed variant. Both models show weak early-layer separability, a pronounced intermediate-layer concentration, and attenuation toward later layers.

D.4 Robustness of Window Localization

Window localization uses a calibration set of 50 examples. Each example is instantiated with two prompt realizations for each of the SUP and ELIM framings, yielding 200 framing-conditioned representations. The resulting model-specific intervention window is fixed before evaluation on the test split.

We assess the stability of this localization for Qwen-2.5-7B on an expanded sample of 1,000 examples. We generate 500 bootstrap resamples and recompute the layer-wise separability profile for each resample. Table D.2 reports the mean normalized score and the width of the 95% bootstrap confidence interval for the ten highest-scoring layers under each diagnostic.

Under both diagnostics, the highest-scoring layers cluster in the intermediate region, with most falling between Layers 11 and 20. The bootstrap results show that localization to this region is consistent across resamples, although the exact layer rankings differ between the two diagnostics.

Label-free separability				Label-informed separability			
Rank	Layer	Mean	95% CI width	Rank	Layer	Mean	95% CI width
1	19	0.898	0.142	1	12	0.730	0.205
2	16	0.891	0.139	2	14	0.716	0.181
3	14	0.882	0.137	3	11	0.705	0.152
4	20	0.868	0.142	4	18	0.705	0.200
5	18	0.853	0.139	5	19	0.682	0.178
6	12	0.839	0.135	6	16	0.674	0.175
7	13	0.807	0.135	7	13	0.665	0.186
8	9	0.775	0.145	8	20	0.660	0.180
9	11	0.767	0.124	9	15	0.636	0.173
10	17	0.747	0.123	10	17	0.635	0.174

Table D.2: Bootstrap summaries of layer-wise separability for Qwen-2.5-7B. For each diagnostic, the table lists the ten layers with the highest mean normalized scores; confidence-interval widths are estimated from 500 bootstrap resamples of the 1,000-example sample.

E Ablation Details and Controls

This section provides implementation details for the ablation experiments described in the main text. All interventions operate on the residual-stream activations at the [STATE] position and are applied layerwise within the intervention window W, unless otherwise specified.

E.1 Token and Interface Controls

Setting	Correct / Total	Accuracy
GLM-4-9B w/o token	812 / 1015	80.00%
With [STATE] token	810 / 1015	79.80%
Qwen-2.5-7B w/o token	882 / 1015	86.90%
With [STATE] token	885 / 1015	87.19%

Table E.1: MedQA-CH accuracy with and without the [STATE] token.

To test interface specificity, we compare substitution at [STATE] with otherwise identical substitution at the original prompt's final content token. Substitution content and the layer window are held fixed in the interface comparison.

E.2 Substitution content

We consider the following substitution types when modifying the [STATE] representation:

- **Zeroing.** The residual vector at the [STATE] position is replaced with a zero vector at each layer in W.
- **Random replacement.** The residual vector is replaced by a randomly sampled vector drawn from the empirical distribution of [STATE] activations at the same layer. Sampling is performed independently for each layer, and the sampled vector is fixed across forward passes within a run.

- **Framing-conditioned substitution.** The residual vector at the [STATE] position is replaced with a cached activation obtained from the complementary framing (Sup↔Elim) for the same input question. Substitution is applied layerwise within W, preserving the original activation outside the intervention window.

E.3 Interface Position

To test whether substitution depends on the structural role of the interface, we compare the canonical final-position configuration with two misaligned placements:

$$\mathbf{x}_1 = [\mathsf{PAD}]^n + I + Q + O + [\mathsf{STATE}],$$

$$\mathbf{x}_2 = [\mathsf{PAD}]^n + I + Q + [\mathsf{STATE}] + O,$$

$$\mathbf{x}_3 = [\mathsf{PAD}]^n + I + [\mathsf{STATE}] + Q + O,$$

where I, Q, and O denote the instruction, question, and answer-option segments, respectively. For all three configurations, we apply the same framing-conditioned substitution over the same layer window W.

We count a generation as failed when it is incoherent or cannot be parsed into the required task-level answer representation. Across the two misaligned configurations, more than 95% of generations fail this criterion. By contrast, the canonical configuration preserves coherent, task-relevant generation for the majority of inputs. This result indicates that [STATE] is not an arbitrary writable slot: its effectiveness depends on being placed after the complete question and option context.

E.4 Quantitative Similarity Summaries for Ablations

This appendix reports distributional summaries for the response-level similarity analyses shown in Figures 6a and 6b. For each condition, we compute length-aware cosine similarity (LAC) and BERTScore-F1 between the baseline output and the output under intervention, and report the median, interquartile range (IQR), and upper quantiles (P90/P95). The number of evaluated runs is denoted by N.

Substitution content: structured vs. unstructured interventions Table E.2 summarizes the similarity distributions for different substitution contents at the [STATE] position (Figure 6b), contrasting framing-conditioned substitution with content-agnostic controls.

Metric	Median	IQR	P90	P95
LAC				
zeroing	0.6620	0.4913	0.9005	0.9351
random	0.6242	0.4933	0.8931	0.9141
cross-framing	0.8766	0.1925	1.0000	1.0000
BERTScore-F1				
zeroing	0.5631	0.1196	0.6954	0.7554
random	0.5560	0.1289	0.6855	0.7460
cross-framing	0.6400	0.3449	1.0000	1.0000

Table E.2: Quantile summaries of response-level similarity for different substitution contents at the [STATE] position (Figure 6b).

Metric	Median	IQR	P90	P95
LAC				
[STATE]	0.8766	0.1925	1.0000	1.0000
Final content token	0.9780	0.0660	1.0000	1.0000
BERTScore-F1				
[STATE]	0.6400	0.3449	1.0000	1.0000
Final content token	0.9444	0.2330	1.0000	1.0000

Table E.3: Quantile summaries comparing substitutions at the [STATE] position and the final content token.

Intervention location: [STATE] vs. final-content-token substitution Table E.3 reports quantile summaries for substitutions applied at [STATE] versus at the final content token position (Figure 6a), using the same intervention window W and decoding procedure.