ANSWER CANDIDATES RESHAPE LANGUAGE MODEL DECISIONS: LAYER-WISE DYNAMICS AND CAUSAL STATE INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple-choice benchmarks usually treat answer options as a passive interface for measuring what a language model knows. We show that answer candidates actively shape how models form answers. For each question, we construct paired inputs that preserve the question and correct answer while replacing only wrong options with unrelated alternatives from the same task. Across six Llama and Qwen models ranging from 3B to 14B parameters and fifteen tasks, we find that benchmark distractors delay (the top-ranked candidate) the last change in the model's favored option in 86 of 90 model-task settings, shorten the span over which the correct answer maintains a consistent lead (persistent correct-answer preference), and reduce accuracy by 20.2 percentage points. We then test whether these internal differences causally affect predictions. Specifically, we transfer the hidden state at the final token of the strongest wrong answer between paired inputs. This intervention improves accuracy by 8.78 percentage points and exceeds norm-matched random and cross-question controls. A separate 360-question panel further shows that reversing the transfer direction reverses the aggregate behavioral effect. These results show that answer candidates are not passive evaluation interfaces. Instead, they actively shape language model computation and causally influence final decisions. *The code is provided in the supplementary material.*

1 INTRODUCTION

Multiple-choice question answering (MCQA) evaluates language models based on their final selection among candidate options. However, these options are also part of the input context, changing them can alter both the available alternatives and the computation leading to the final choice. This dual role complicates the interpretation of model behavior, yet how candidate composition shapes answer preferences across intermediate layers remains poorly understood.

Recent work provides behavioral and mechanistic foundations for this question. (Cho et al., 2026) shows that changing wrong options can alter final predictions, while the Tuned Lens (Belrose et al., 2023) maps intermediate hidden states to token predictions for layer-wise inspection of model preferences. (boyu shi et al., 2026) further uses decision margins, the score gap between the correct answer and its strongest alternative, to study when a sustained correct-answer lead emerges across layers. Building on these insights, we ask **how a controlled change in candidate context alters answer preferences across layers, and whether candidate-dependent internal states can causally change the model's final choice.**

To handle this question, as shown in Figure 1(A), we construct paired inputs that differ only in their wrong options. For each question, the benchmark input retains the original candidate set, whereas the unrelated input replaces wrong options with candidates from other questions within the same task, selected for similar length and low word overlap with the question. The question text, correct answer, correct-answer position, and total number of options remain fixed. Then, as shown in Figure 1(B), we characterize the resulting layer-wise trajectories using two complementary measurements. **Last revision depth** records the layer at which the model's top-ranked candidate changes for the final time, indicating how late its current choice remains open to revision. **Persistent correct preference** identifies when the correct candidate begins to maintain a sustained lead over the

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

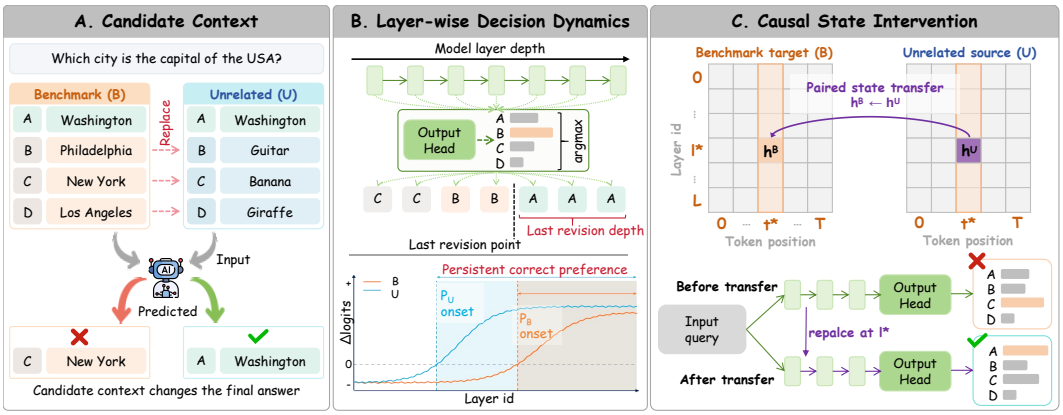


Figure 1: The overall framework of the study. (A) constructs paired inputs. (B) characterizes the resulting layer-wise answer formation using last revision depth and persistent correct preference measurements. $\Delta logits$ is the correct candidate advantage over strongest competitor. (C) tests whether candidate-dependent internal states causally affect final choices through full-state transfers.

alternatives across the remaining layers. Together, the two measurements capture complementary aspects of answer formation: the timing of final candidate revision and the emergence of a sustained correct-answer advantage.

Finally, we test whether candidate-dependent internal states contribute causally to the final selection. As illustrated in Figure 1(C), we fix the highest-scoring wrong option in the unmodified benchmark run as the competitor and define its boundary as the final token of that option text. At a selected layer, we replace the benchmark state at this boundary with the state from the corresponding option slot in the paired unrelated input. This full-state transfer changes a single contextual representation while leaving the benchmark input and model parameters unchanged. We evaluate the paired transfer against norm-matched random updates and cross-question state differences to test whether the effect depends on the paired state source. We further examine the directionality of this effect by reversing the transfer direction on an independent question panel.

Across six Llama (Grattafiori et al., 2024) and Qwen (Qwen et al., 2025; Yang et al., 2025) models and 15 tasks, candidate composition consistently changes answer formation. Benchmark inputs delay final revision in 86 of 90 model-task settings, reduce the fraction of network depth with sustained correct preference, and lower final accuracy by 20.2 percentage points. Seven readout ablations preserve the persistent-preference contrast across changes in normalization, label form, calibration, and option rotation. These results show that candidate context affects both when layer-wise preferences settle and how long the correct answer maintains a sustained advantage.

The intervention experiments further show a causal contribution of candidate-dependent contextual states. On a separate 180-question confirmation panel, paired boundary-state transfer improves accuracy by 8.78 percentage points. It also increases the score gap between the correct answer and the selected wrong competitor more than either norm-matched random or cross-question controls. On a separate 360-question panel, we test the transfer in both directions. Transferring unrelated states into benchmark inputs improves accuracy, whereas the reverse transfer produces an aggregate decrease. These experiments move from an observational candidate-context contrast to causal evidence that paired contextual boundary states influence final choices and are accompanied by corresponding changes in layer-wise answer formation.

2 Related Work

2.1 Candidate context and intermediate predictions

Multiple-choice predictions are sensitive to candidate presentation and content, including option position and distractor changes (Zheng et al., 2024; Cho et al., 2026). Beyond this output-level sensitivity, recent work has examined how answers emerge internally. Layer-wise readout meth-

ods recover evolving predictions from hidden states (Belrose et al., 2023; Ghandeharioun et al., 2024), while activation-based analyses localize representations involved in multiple-choice answer formation (Wiegreffe et al., 2025). Decision-level analyses further characterize when correct-answer preferences emerge and stabilize across layers (boyu shi et al., 2026). Our work connects these directions by testing how controlled changes in candidate content reshape layer-wise answer trajectories while the question and correct answer remain fixed.

2.2 Causal states and interventions

Mechanistic interpretability increasingly uses internal-state interventions to test whether representations contribute functionally to model behavior (Conmy et al., 2023; Li et al., 2023). Activation patching provides a standard framework for such tests, although its conclusions depend on choices such as the evaluation metric and source construction. Related frameworks extend intervention-based analysis to hidden representations and causal variables (Ghandeharioun et al., 2024; Mueller et al., 2025), while circuit studies identify functionally relevant components underlying knowledge and reasoning behaviors (Yao et al., 2024; Hong et al., 2026). Following this perspective, we intervene on candidate-dependent boundary states and compare them with matched alternative updates to isolate the causal contribution of the paired state source.

3 Methodology

3.1 Paired Candidate Inputs and Evaluation Data

For each multiple-choice question, we construct benchmark and unrelated inputs that differ only in their wrong options. The benchmark condition B retains the original candidate set. The unrelated condition U replaces each wrong option with a candidate from another question in the same task. Replacements are length-matched to reduce token-length effects, have low lexical overlap with the question to avoid direct question-option cues, and remain within task to preserve task-specific option format and content distribution. The question text, correct answer text and position, and option count are fixed across the pair.

We evaluate this contrast on six Llama and Qwen models from 3B to 14B parameters and 15 tasks, yielding 40,890 paired model-question records. The intervention experiments use six tasks and disjoint panels of 36 discovery questions, 180 first-confirmation questions, and 360 bidirectional-confirmation questions. Question identities are shared across models within each panel. Model, task, and dataset details are provided in the Appendix A.

3.2 Layer-wise Answer Readout and Trajectory Measures

Layer-wise readout. At each layer ℓ, we apply the model's final normalization and output projection to the hidden state at the final prompt position. Answer labels such as "A" may be encoded as different single tokens depending on surrounding whitespace. To account for this variation, we consider all valid single-token forms of each label. For each option content a, $\lambda(a)$ denotes its corresponding option (e.g., "A", "B", "C", or "D"), and $\mathcal{T}(\lambda(a))$ denotes the set of valid single-token forms of that label. We then score each option using the largest logit among these forms and select the option with the highest score.

$$s_\ell(a) = \max_{v \in \mathcal{T}(\lambda(a))} z_\ell(v), \qquad \hat{a}_\ell = \arg\max_a s_\ell(a). \tag{1}$$

Here, $z_\ell(v)$ is the logit of token v at layer ℓ, $s_\ell(a)$ is the resulting score for candidate a, and $\hat{a}_\ell$ is the top-ranked candidate. When the positions of option content are rotated, we continue tracking the content rather than its option. This ensures that comparisons are not affected by position changes. Because the score is directly computed from the model's own output head, this readout requires no additional training.

Trajectory measures. To characterize how candidate context shapes layer-wise answer formation in multiple-choice tasks, we define two complementary measurements. Last revision depth captures

when the model's preferred option becomes fixed during forward computation. It records the final layer at which the top-ranked option changes. For a model with L layers, we define the normalized depth R as follows.

$$R = \frac{1}{L} \max\left(\{1\} \cup \{\ell \in \{2, \ldots, L\} \mid \hat{a}_\ell \neq \hat{a}_{\ell-1}\}\right). \tag{2}$$

The set identifies layers where the top-ranked option changes, and the maximum selects the final revision event before normalization by model depth. The term $\{1\}$ handles trajectories without any change, giving $R = 1/L$. The larger R indicates that the model revises its preferred option until later layers. This measure captures revision timing independently of whether the final option is correct.

Persistent correct preference measures when the correct option begins to establish a sustained advantage over alternatives. At each layer ℓ, we define whether the correct option leads all alternatives and the fraction of subsequent layers where this lead is maintained:

$$b_\ell = \mathbf{1}\left[s_\ell(g) > \max_{a \neq g} s_\ell(a)\right], \qquad S_\ell = \frac{1}{L - \ell + 1} \sum_{t=\ell}^{L} b_t. \tag{3}$$

Here, g denotes the correct option, a denote any answer option, b_ℓ indicates whether the correct option has the highest score at layer ℓ, and S_ℓ measures how often this advantage is preserved from layer ℓ to the final layer. We define the onset layer P as the earliest layer where the correct option leads and remains ahead in at least 80% of subsequent layers, and define the corresponding persistent span A as:

$$P = \min\left(\{\ell \mid b_\ell = 1,\ S_\ell \geq 0.8\} \cup \{L + 1\}\right), \qquad A = 1 - \frac{\min(P, L)}{L}. \tag{4}$$

If no layer satisfies the criterion, $P = L + 1$ and $A = 0$. A larger A indicates that the correct option maintains a sustained advantage over a larger fraction of the network. This criterion does not require an uninterrupted lead. The correct option only needs to lead at onset and in at least 80% of the remaining layers.

3.3 Paired State Transfer and Intervention Analysis

Paired state transfer. The trajectory measures reveal observational differences, but do not establish whether candidate-dependent states affect final choices. We therefore perform a full-state transfer intervention. For each question, we identify the fixed competitor r as the highest-scoring wrong option in the unmodified benchmark input. Its boundary is defined as the final token of the option text. At the discovery-selected layer, let h^B and h^U denote the boundary states from the benchmark and unrelated inputs. We transfer the state difference from the unrelated condition into the benchmark condition:

$$\delta = h^U - h^B, \qquad h^{B \leftarrow U} = h^B + \delta = h^U. \tag{5}$$

This operation changes only one hidden state while keeping the benchmark input and model parameters unchanged. Later layers are then recomputed using the transferred state. The transferred vector represents a complete contextual state and may contain information from both the option and preceding prompt. We evaluate the forward transfer and its reverse direction on an independent question panel.

Matched controls and analysis. To isolate the source effect of the transferred state, we compare paired transfer with norm-matched random updates and cross-question state differences. We further evaluate position and layer controls to examine where the transfer effect is localized. The primary intervention outcome is the final score gap between the correct option and the fixed competitor:

$$G = \frac{s_L(g) - s_L(r)}{\sigma_m}, \tag{6}$$

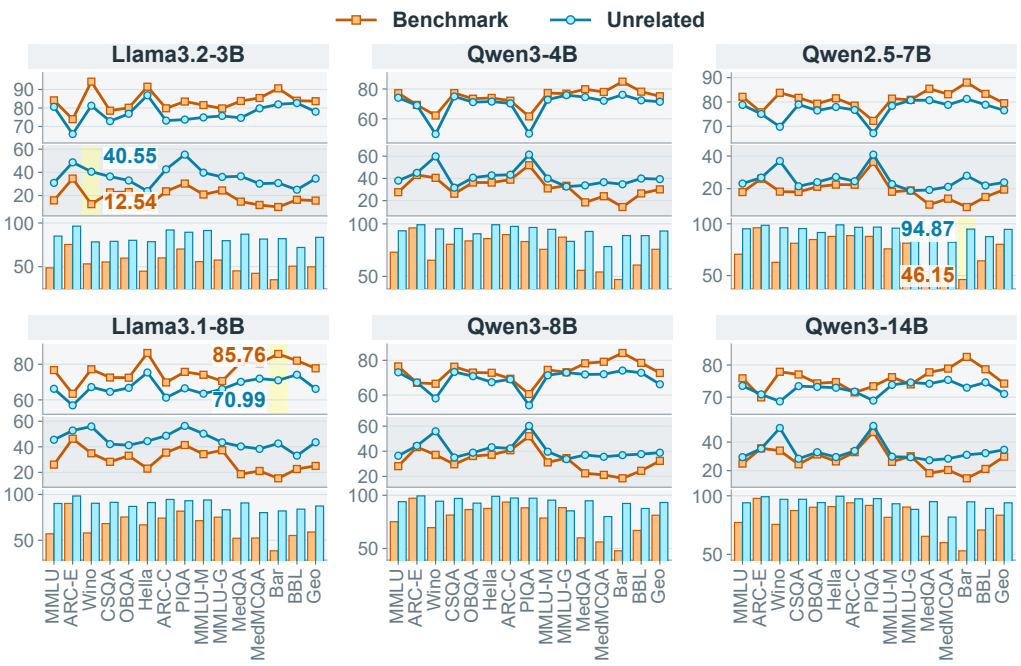


Figure 2: Candidate-context effects across six models and 15 tasks. Benchmark (B) and unrelated (U) inputs are compared for last revision depth (Top), persistent correct-preference span (Middle), and final accuracy (Bottom).

Here, g denotes the correct option, r denotes the fixed competitor, and $s_L(\cdot)$ represents the final-layer score of an option. The scaling constant σ_m is determined on the discovery set and fixed before confirmation. The larger G indicates a stronger preference for the correct option over the competitor. We additionally report final accuracy, correct-answer probability, and last revision depth as complementary outcomes. Aggregation and uncertainty estimation details are provided in Appendix C.

4 Experiments

In this section, we first study how replacing only the wrong options reshapes layer-wise answer formation and test the robustness of this pattern across readout variants. We then perform causal interventions with paired candidate-dependent states, analyzing their impact on final decisions, the layers most responsible for this effect, and its reversibility across transfer directions. The detailed models, tasks, and implementation settings are provided in Appendices A and B.

4.1 Candidate Context Shifts Answer Revision Later

We compare benchmark and unrelated candidate sets across six models and 15 tasks using last revision depth, persistent correct-preference span, and final accuracy. Figure 2 summarizes all 90 model-task combinations. Across the three measurements, benchmark inputs show later final revisions (top), shorter sustained correct-answer preference (middle), and lower accuracy (bottom). Mean last revision depth increases from 72.14% to 77.49%, with later revisions in 86 of 90 cells. Persistent correct-preference span decreases from 37.13% to 26.68%, with a later onset in 85 of 90 cells. Final accuracy decreases from 91.06% to 70.89%.

Together, these aligned but distinct outcomes show that candidate context changes both when answer preferences settle and how long the correct answer maintains a sustained advantage. Because replacing wrong options can simultaneously alter several content properties, including their compatibility with the question, we interpret this as a candidate-context effect under the controlled replacement procedure.

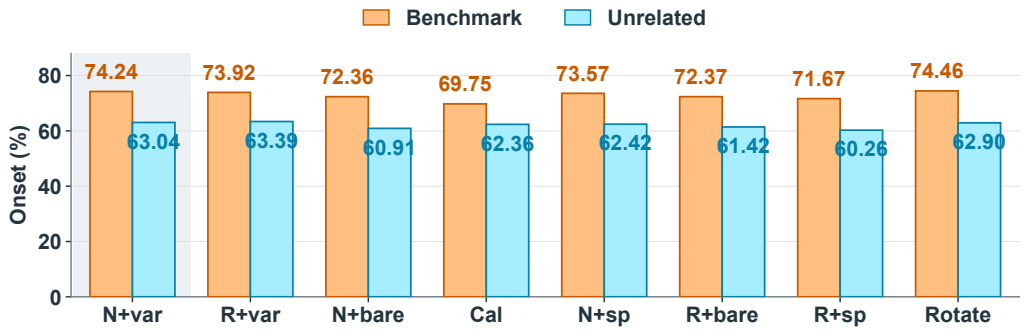


Figure 3: Robustness of persistent correct-preference onset to readout choices. The x-axis shows eight readout variants: N/R denote native normalization/raw residual projection, var/bare denote label variants/bare labels, sp denotes space-prefixed labels, Cal denotes logistic calibration, and Rotate denotes cyclic option rotations. Lower onset indicates earlier sustained correct preference.

Table 1: Accuracy before and after paired boundary-state transfer on the 180-question confirmation panel. Δ denotes the accuracy change in percentage points (↑: larger values indicate greater improvement).

	Llama3.2-3B			**Qwen3-4B**			**Qwen2.5-7B**		
Task	Base	Transfer	Δ	Base	Transfer	Δ	Base	Transfer	Δ
MMLU	49.17	65.00	+15.83↑	64.17	72.50	+8.33↑	69.17	75.00	+5.83↑
ARC-C	60.00	67.50	+7.50↑	92.50	94.17	+1.67↑	85.83	92.50	+6.67↑
CSQA	58.67	71.33	+12.67↑	77.33	86.00	+8.67↑	71.33	81.33	+10.00↑
PIQA	58.33	76.67	+18.33↑	85.00	96.67	+11.67↑	83.33	95.00	+11.67↑
MedQA	37.33	49.33	+12.00↑	50.00	62.67	+12.67↑	59.33	60.67	+1.33↑
Geo	47.50	56.67	+9.17↑	75.00	85.00	+10.00↑	81.67	88.33	+6.67↑
Avg.	51.83	64.42	+12.58↑	74.00	82.83	+8.83↑	75.13	82.14	+7.03↑
	Llama3.1-8B			**Qwen3-8B**			**Qwen3-14B**		
Task	Base	Transfer	Δ	Base	Transfer	Δ	Base	Transfer	Δ
MMLU	55.83	63.33	+7.50↑	70.83	77.50	+6.67↑	73.33	82.50	+9.17↑
ARC-C	72.50	80.00	+7.50↑	93.33	95.83	+2.50↑	98.33	99.17	+0.83↑
CSQA	67.33	81.33	+14.00↑	80.00	92.00	+12.00↑	82.67	89.33	+6.67↑
PIQA	65.00	78.33	+13.33↑	88.33	96.67	+8.33↑	88.33	95.00	+6.67↑
MedQA	46.00	53.33	+7.33↑	59.33	70.00	+10.67↑	59.33	76.67	+17.33↑
Geo	55.83	61.67	+5.83↑	80.00	83.33	+3.33↑	84.17	90.00	+5.83↑
Avg.	60.42	69.67	+9.25↑	78.64	85.89	+7.25↑	81.03	88.78	+7.75↑

4.2 The Persistent-Preference Contrast Is Robust to Readout Choices

The broad analysis shows that benchmark inputs produce a shorter persistent correct-preference span. We next test whether this contrast depends on the readout procedure. We repeat the analysis under seven alternative readout variants that modify normalization, answer-label representation, position calibration, and option rotation. Figure 3 compares persistent-preference onset across the reference readout and these variants.

Across all eight readouts, benchmark inputs consistently show later persistent-preference onset than unrelated inputs. Under the reference setting, the mean onset is 74.24% of network depth for benchmark inputs and 63.04% for unrelated inputs. Although alternative readouts shift the absolute onset values, the direction of the contrast remains unchanged in every setting. Complete task-level, model-level, and model-task results are provided in Appendix D. These analyses show that the observed candidate-context effect is robust to the tested normalization, label representation, calibration, and option-order choices.

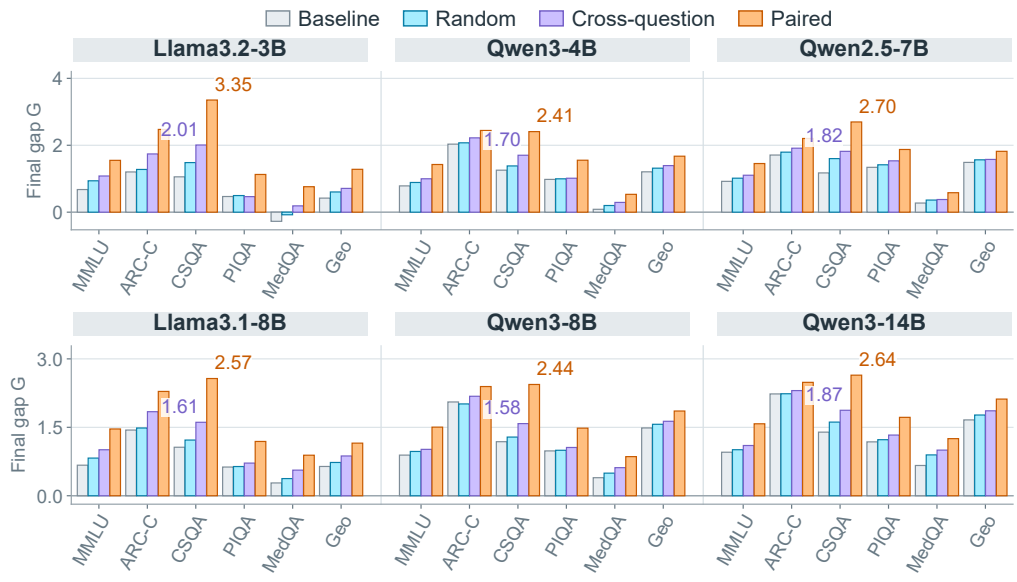


Figure 4: Source specificity of paired boundary-state transfer on the 180-question confirmation panel. Each model block compares four intervention conditions across six tasks: Baseline (no state replacement), Paired transfer (replacement with the same question's paired unrelated state), Random (norm-matched random state updates), and Cross-question (norm-matched state differences from another question in the same task). Values are absolute fixed-competitor score gaps (G) between the correct candidate and the selected wrong competitor, with larger values indicating stronger preference for the correct candidate.

4.3 Paired Boundary-State Transfer Causally Changes Final Choices

The previous analyses establish that candidate context alters layer-wise answer formation, but they do not determine whether the corresponding internal states causally affect the final decision. We therefore perform a state intervention on the 180-question confirmation panel. At the discovery-selected layer, we replace the benchmark state at the fixed competitor boundary with its paired state from the unrelated input, while keeping the benchmark text and model parameters unchanged.

Paired boundary-state transfer changes final predictions and improves accuracy across models and tasks. To determine whether this improvement depends on information contained in the paired state rather than an arbitrary hidden-state perturbation, Figure 4 compares the paired transfer with three control conditions: the unmodified baseline, norm-matched random updates, and norm-matched cross-question updates. The random control tests whether an update of the same magnitude is sufficient to produce the effect, whereas the cross-question control tests whether a structured state difference from another question in the same task is sufficient.

Using the fixed-competitor score gap as a continuous measure of preference for the correct candidate, the gap increases from 1.017 to 1.752 after paired transfer. The paired update exceeds random and cross-question controls by 0.622 and 0.466 gap units on average, respectively. The corresponding effects are positive in all 36 model-task combinations. These comparisons show that the intervention effect depends on the source of the transferred state rather than update magnitude or a generic structured perturbation. Because the transferred vector is a complete contextual boundary state, these results provide evidence that candidate-dependent contextual states contribute causally to final choice formation. Complete model-task intervention outcomes are provided in Appendix E. We next examine where this transfer effect is strongest across input positions and network depths.

4.4 The Transfer Effect Depends on Position and Layer

Having established that paired boundary-state transfer has a source-dependent causal effect, we next examine where this effect is localized. We repeat the transfer with alternative input positions and

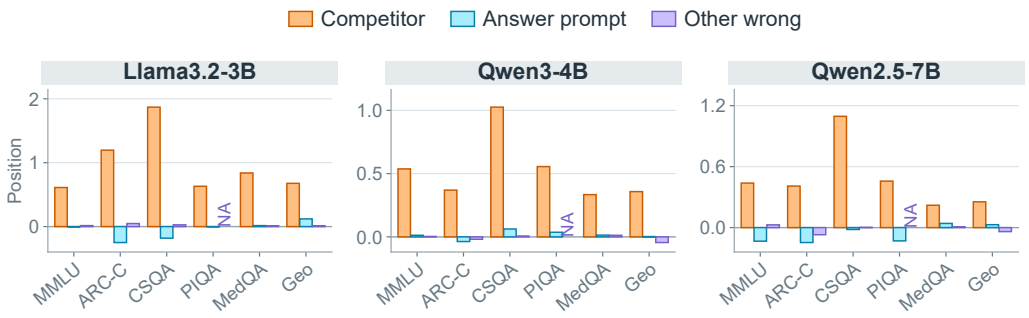


Figure 5: Position dependence of paired state transfer. Each panel compares the competitor boundary with the answer-prompt position and another wrong-option boundary. Values show source effects on the scaled fixed-competitor gap after subtracting matched random effects.

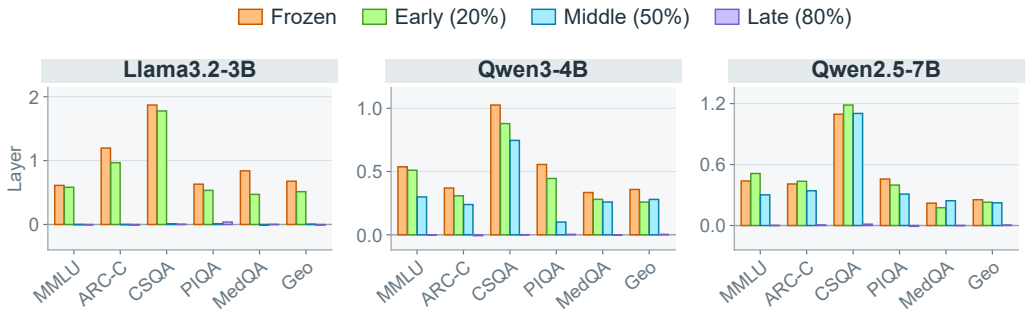


Figure 6: Layer dependence of paired state transfer. Each panel compares the frozen selected layer with representative early, middle, and late depths. Values show source effects on the scaled fixed-competitor gap after subtracting matched random effects.

network depths, reporting source effects after subtracting matched random updates. Figures 5 and 6 summarize the position and layer comparisons.

The position analysis identifies the competitor boundary as the strongest tested location. Its effect exceeds the answer-prompt position by 0.652 gap units on average and another wrong-option boundary by 0.659 gap units. This pattern suggests that the transferred state is most effective when introduced at the boundary of the candidate that directly competes with the correct answer. In contrast, states from the answer prompt or another wrong option provide less useful contextual information for shifting the final preference, even when the update magnitude is matched. The ordering is consistent across models and tasks.

The layer analysis shows that the discovery-selected layer performs better than representative early, middle, and late alternatives by 0.058, 0.413, and 0.623 gap units, respectively. The relatively smaller difference from the early layer indicates that useful candidate-dependent information is already present before the selected depth. However, the larger separation from middle and late alternatives suggests that the selected layer provides a more effective point for transferring this information into downstream computation. These results localize the strongest transfer effect to the competitor boundary and the selected intervention depth among the tested positions and layers. Complete localization results are provided in Appendix E.3.

4.5 Full-State Effects Extend to Both Transfer Directions

The initial confirmation panel evaluates only transfer from unrelated states into benchmark inputs. We therefore test the same frozen intervention configuration in both directions on a separate 360-question panel with disjoint question identities. Table 2 summarizes the resulting accuracy changes.

Table 2: Accuracy changes under bidirectional full-state transfer on the 360-question confirmation panel. Base denotes the target condition before transfer, and Δ reports the accuracy change in percentage points.

	Llama3.2-3B						Qwen3-4B					
	$U \to B$			$B \to U$			$U \to B$			$B \to U$		
Task	Base	Transfer	Δ	Base	Transfer	Δ	Base	Transfer	Δ	Base	Transfer	Δ
MMLU	53.75	60.83	+7.08↑	90.42	83.33	-7.08↓	73.75	76.67	+2.92↑	94.17	87.50	-6.67↓
ARC-C	68.00	81.33	+13.33↑	92.50	87.50	-5.00↓	89.58	96.67	+7.08↑	99.17	97.50	-1.67↓
CSQA	52.33	68.00	+15.67↑	81.00	68.33	-12.67↓	77.00	88.00	+11.00↑	93.67	82.33	-11.33↓
PIQA	65.83	75.00	+9.17↑	87.50	88.33	+0.83↑	76.67	90.00	+13.33↑	95.83	97.50	+1.67↑
MedQA	45.00	58.00	+13.00↑	80.00	75.67	-4.33↓	63.67	70.33	+6.67↑	88.33	83.33	-5.00↓
Geo	56.25	64.58	+8.33↑	80.42	79.58	-0.83↓	78.33	89.17	+10.83↑	89.58	88.33	-1.25↓
Avg.	**56.86**	**67.96**	**+11.10↑**	**85.31**	**80.46**	**-4.85↓**	**76.50**	**85.14**	**+8.64↑**	**93.46**	**89.42**	**-4.04↓**
	Qwen2.5-7B						**Llama3.1-8B**					
	$U \to B$			$B \to U$			$U \to B$			$B \to U$		
Task	Base	Transfer	Δ	Base	Transfer	Δ	Base	Transfer	Δ	Base	Transfer	Δ
MMLU	69.17	78.75	+9.58↑	97.08	91.67	-5.42↓	55.83	66.25	+10.42↑	90.00	86.25	-3.75↓
ARC-C	92.50	97.92	+5.42↑	98.33	97.92	-0.42↓	77.92	87.50	+9.58↑	94.58	93.75	-0.83↓
CSQA	75.00	88.33	+13.33↑	92.00	79.00	-13.00↓	64.00	81.33	+17.33↑	93.67	80.00	-13.67↓
PIQA	84.17	90.83	+6.67↑	95.83	96.67	+0.83↑	72.50	75.00	+2.50↑	90.83	93.33	+2.50↑
MedQA	61.00	69.67	+8.67↑	85.00	78.33	-6.67↓	58.67	66.00	+7.33↑	83.67	83.00	-0.67↓
Geo	73.75	82.08	+8.33↑	90.00	88.33	-1.67↓	57.08	62.92	+5.83↑	82.50	80.83	-1.67↓
Avg.	**75.93**	**84.60**	**+8.67↑**	**93.04**	**88.65**	**-4.39↓**	**64.33**	**73.17**	**+8.83↑**	**89.21**	**86.19**	**-3.02↓**
	Qwen3-8B						**Qwen3-14B**					
	$U \to B$			$B \to U$			$U \to B$			$B \to U$		
Task	Base	Transfer	Δ	Base	Transfer	Δ	Base	Transfer	Δ	Base	Transfer	Δ
MMLU	75.83	81.25	+5.42↑	94.58	91.67	-2.92↓	80.83	87.08	+6.25↑	95.00	91.67	-3.33↓
ARC-C	95.42	97.92	+2.50↑	99.17	98.33	-0.83↓	94.58	95.83	+1.25↑	97.50	97.50	0.00
CSQA	78.67	89.67	+11.00↑	95.33	84.33	-11.00↓	79.67	95.67	+16.00↑	98.33	90.00	-8.33↓
PIQA	89.17	92.50	+3.33↑	96.67	96.67	0.00	82.50	92.50	+10.00↑	95.83	97.50	+1.67↑
MedQA	68.33	75.67	+7.33↑	89.00	81.00	-8.00↓	69.00	81.00	+12.00↑	88.33	80.33	-8.00↓
Geo	82.08	92.08	+10.00↑	89.17	83.33	-5.83↓	85.00	91.67	+6.67↑	90.42	87.92	-2.50↓
Avg.	**81.58**	**88.18**	**+6.60↑**	**93.99**	**89.22**	**-4.76↓**	**81.93**	**90.62**	**+8.70↑**	**94.24**	**90.82**	**-3.42↓**

Transferring unrelated states into benchmark inputs increases accuracy from 72.86% to 81.61%, with improvements in all 36 model-task cells. Reversing the transfer direction yields the opposite aggregate trend: transferring benchmark states into unrelated inputs decreases accuracy from 91.54% to 87.46% on average.

The continuous score-gap measure shows the same directional pattern as the accuracy results. Under $U \to B$, the mean fixed-competitor gap increases from 1.138 to 1.875 across all 36 model-task cells, whereas $B \to U$ decreases the mean gap from 2.278 to 1.817 in 34 of 36 cells (Appendix E.4). These changes indicate that the transferred state modifies not only the final accuracy but also the model's relative preference between the correct candidate and its strongest competitor. Revision timing shows corresponding shifts on the same panel: forward transfer moves the final preference revision earlier, whereas reverse transfer moves it later (Appendix E.4). These results show that candidate-dependent contextual states influence final choices in a direction-dependent manner and that this effect generalizes to new question identities.

5 Conclusion

Multiple-choice benchmarks often treat answer candidates as passive interfaces for evaluating model knowledge. Our findings show that candidates actively shape language model computation. By controlling the question and correct answer while changing only wrong options, we reveal that candidate context alters layer-wise answer formation and that the corresponding contextual states can causally influence final choices. These results establish that answer options are not merely outputs to be selected among, but also inputs that participate in the decision process being measured. Our paired candidate analysis and state intervention framework provide a systematic approach for studying how input composition affects internal decision dynamics in language models. More broadly, these findings suggest that understanding model behavior requires analyzing not only what models predict but also how candidate contexts shape the path toward those predictions.

AI USE STATEMENT

Generative AI tools are used to aid or polish manuscript writing, retrieve and discover related literature, and draft sections of the manuscript. All AI-assisted content is reviewed and revised by the authors. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human subjects, private data collection, or applications with direct social risks. We study language model decision mechanisms under controlled benchmark settings. The released materials are intended to support reproducible research and further understanding of model behavior.

REPRODUCIBILITY STATEMENT

We provide code and implementation details to facilitate reproduction of our experiments. The main paper describes the paired candidate construction, layer-wise readout, state transfer intervention, and evaluation metrics. Appendix A summarizes the evaluated models, tasks, and question panels. Appendix B provides implementation details for candidate readout and intervention procedures, including the frozen intervention configuration. Appendix C describes aggregation procedures and uncertainty estimation. The complete source code is provided in the supplementary materials.

REFERENCES

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

boyu shi, Chang Liu, Chuanbao Gao, Xu Yang, and Xin Geng. Understanding performance collapse in layer-pruned large language models via decision representation transitions. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=bPnEyHvZSu.

Gyeongje Cho, Yeonkyoung So, and Jaejin Lee. Choices speak louder than questions. In *International Conference on Learning Representations*, volume 2026, pp. 75943–75962, 2026.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Vijay Devane, Mohd Nauman, Bhargav Patel, Aniket Mahendra Wakchoure, Yogeshkumar Sant, Shyam Pawar, Viraj Thakur, Ananya Godse, Sunil Patra, Neha Maurya, et al. Bhashabench v1: A comprehensive benchmark for the quadrant of indic domains. *arXiv preprint arXiv:2510.25409*, 2025.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*, 2024.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Guan Zhe Hong, Nishanth Dikkala, Enming Luo, Cyrus Rashtchian, Xin Wang, and Rina Panigrahy. A implies b: Circuit analysis in llms for propositional logical reasoning. *Advances in Neural Information Processing Systems*, 38:30749–30803, 2026.

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

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, et al. Mib: A mechanistic interpretability benchmark. *arXiv preprint arXiv:2504.13151*, 2025.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Rodrigo Ferreira Rodrigues, Karim Radouane, Jose G Moreno, and Lynda Tamine. Geobench-llm: A comprehensive benchmark for evaluating llms on geo-related tasks. *arXiv preprint arXiv:2608.07411*, 2026.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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)*, pp. 4149–4158, 2019.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hanna Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how lms answer multiple choice questions. In *International Conference on Learning Representations*, volume 2025, pp. 78293–78318, 2025.

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.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. *Advances in Neural Information Processing Systems*, 37:118571–118602, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pp. 19426–19454, 2024.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D Manning, Peter Henderson, and Daniel E Ho. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pp. 169–193, 2025.

APPENDIX

Appendix A describes the evaluated models, benchmark tasks, and question panels.

Appendix B provides implementation details for layer-wise readout and intervention procedures.

Appendix C describes aggregation, uncertainty estimation, and additional statistical analyses.

Appendix D reports additional robustness analyses for candidate context effects.

Appendix E provides complete intervention results and control comparisons.

A MODELS, TASKS, AND EVALUATION PANELS

A.1 MODEL AND TASK

We evaluate the six Llama and Qwen models, including Llama3.2-3B, Llama3.1-8B, Qwen2.5-7B, Qwen3-4B, Qwen3-8B, and Qwen3-14B, spanning 3B to 14B parameters. Here, we also list all fifteen tasks (Hendrycks et al., 2020; Clark et al., 2018; Talmor et al., 2019; Bisk et al., 2020; Zellers et al., 2019; Mihaylov et al., 2018; Jin et al., 2021; Sakaguchi et al., 2021; Pal et al., 2022; Zheng et al., 2025; Devane et al., 2025; Rodrigues et al., 2026). We use the following task names and abbreviations throughout the paper. The general knowledge and commonsense tasks include Massive Multitask Language Understanding (MMLU), the Easy and Challenge sets of the AI2 Reasoning Challenge (ARC-E and ARC-C), WinoGrande (Wino), and CommonsenseQA (CSQA). They also include OpenBookQA (OBQA), HellaSwag (Hella), and Physical Interaction Question Answering (PIQA). The domain-specific tasks comprise the medical MMLU subset (MMLU-M), the high-school geography MMLU subset (MMLU-G), MedQA, MedMCQA, Bar Exam QA (Bar), BhashaBench-Legal (BBL), and GeoBenchLLM (Geo). MedQA and MedMCQA retain their original dataset names in all figures and tables. To examine whether our findings hold across knowledge domains, we report the two MMLU subsets separately from the remaining subjects. Throughout the paper, MMLU denotes the remaining subjects after excluding MMLU-M and MMLU-G, ensuring that the three evaluation sets are disjoint. Interventions use MMLU, ARC-C, CSQA, PIQA, MedQA, and Geo.

A.2 QUESTION PANELS

We divide the evaluation questions into separate panels according to their experimental roles. The Broad panel supports the main candidate-context analysis by comparing benchmark and unrelated candidate sets across models and tasks. The Discovery panel is used only to determine the frozen intervention configuration, including the intervention layer and score normalization constants.

The two confirmation panels use question identities disjoint from the Discovery panel and provide independent evaluations of the state-transfer effect. The Forward-transfer panel evaluates the transfer from unrelated states into benchmark inputs ($U \rightarrow B$), while the Bidirectional-transfer panel evaluates both $U \rightarrow B$ and $B \rightarrow U$ to examine whether the behavioral effect follows the direction of state transfer. Table 3 summarizes the evaluation panels.

Table 3: Evaluation panels and their experimental roles.

Panel	Tasks	Questions	Role
Broad	15	–	Candidate-context analysis
Discovery	6	36	Intervention selection
Forward-transfer	6	180	Forward state transfer evaluation
Bidirectional-transfer	6	360	Directionality evaluation

A.3 FROZEN INTERVENTION CONFIGURATION

The intervention configuration is selected exclusively on the Discovery panel and fixed before the state-transfer evaluations. This procedure prevents the intervention location and score normalization

from being optimized on evaluation questions. For each model, the selected configuration includes the intervention block and the model-specific score normalization constant used for fixed-competitor gap comparisons.

Table 4 summarizes the frozen configurations for all evaluated models. The hook index follows the zero-based indexing convention used in implementation, whereas the reported intervention block uses one-based indexing for readability. The score scale denotes the fixed normalization constant determined on the Discovery panel and applied consistently during subsequent evaluations.

Table 4: Frozen intervention configurations for state-transfer experiments. Hook indices follow zero-based implementation indexing, while intervention blocks use one-based indexing. Score scales are model-specific normalization constants determined on the Discovery panel and fixed before evaluation.

Model	Total Blocks	Hook Index	Intervention Block	Score Scale
Llama3.2-3B	28	7	8	0.609884
Qwen3-4B	36	9	10	1.369867
Qwen2.5-7B	28	7	8	1.300801
Llama3.1-8B	32	8	9	0.746601
Qwen3-8B	36	9	10	1.411753
Qwen3-14B	40	10	11	1.592929

B Readout and Intervention Implementation

B.1 Layer-wise Readout

We extract the residual state after each decoder block at the final prompt position and apply the model's native final normalization and output projection. The embedding output is excluded. Because answer labels may correspond to different single tokens depending on surrounding whitespace, we evaluate all valid single-token forms of each option label and use the maximum logit among them. Correct-answer probability is computed by applying softmax over the candidate scores rather than the full vocabulary.

For option rotations, scores are mapped back to the original option identities before comparison. A tied correct-answer score does not satisfy the strict lead criterion used for persistent correct preference. Last revision depth records the final layer at which the top-ranked candidate changes. If no such change occurs, the first layer is assigned. These procedures ensure that layer-wise comparisons are based on candidate identities rather than formatting choices.

B.2 State Transfer Implementation

The state transfer procedure follows the frozen configuration selected on the Discovery panel. For each intervention, the fixed competitor is defined as the highest-scoring wrong option in the benchmark input, and its boundary is the final token of the option text. We replace the residual state at this boundary with the corresponding state from the paired input while keeping the benchmark input and model parameters unchanged. Later layers are then recomputed after the replacement.

For control analyses, random updates are normalized to match the paired state difference, and cross-question controls use state differences from another question within the same task. Position controls compare the competitor boundary with the final prompt position and another wrong-option boundary. Layer controls compare the selected intervention block with representative early, middle, and late blocks.

C Aggregation and Statistical Analysis

All analyses are based on paired comparisons between conditions. For the broad candidate-context analyses, we first average results across cyclic option rotations within each question when applicable. We then compute paired contrasts for each model-task combination, with each model-task

cell contributing equally to the aggregate results. For intervention analyses, uncertainty intervals are estimated using bootstrap resampling over models, tasks, and question identities. Each bootstrap sample resamples the evaluated models and tasks with replacement, followed by resampling question identities within each sampled task. The same resampled questions are retained across paired conditions. We report the 2.5th and 97.5th percentiles of the bootstrap distribution. The forward-transfer evaluation uses 10,000 bootstrap replicates, and the bidirectional-transfer evaluation uses 5,000 replicates.

The primary intervention comparison evaluates paired transfer against the baseline, norm-matched random updates, and cross-question controls using the fixed-competitor score gap. Accuracy, correct-answer probability, revision timing, and localization analyses are reported as complementary outcomes. For all intervention analyses, changes are calculated as transfer minus baseline. Direction counts summarize the consistency of effects across model-task cells rather than statistical significance for individual cells.

D Robustness of Candidate Context Effects

We test whether the persistent-preference contrast is robust to alternative readout choices. Table 5 summarizes the reference readout and seven variants that modify normalization, label representation, calibration, or option rotation. The corresponding model-task results are shown in Figures 7 and 8.

Across all eight readouts, benchmark inputs consistently show later persistent-preference onset than unrelated inputs. Under the reference readout, the mean onset is 74.24% for benchmark inputs and 63.04% for unrelated inputs. The absolute onset values vary across readouts, but the direction of the contrast remains unchanged. All 48 model averages and 120 task averages are positive, and 694 of 720 model-task contrasts show the same direction. Figures 7 and 8 provide the complete model-task breakdown.

Table 5: Persistent-preference onset across eight readout variants. B and U denote benchmark and unrelated inputs. Values show the onset layer as a percentage of model depth. Δ denotes B minus U, and positive counts indicate the number of model-task cells with later onset under B.

Setting	Readout variant	B	U	Δ	Positive
N+var	Native norm + label variants	74.24	63.04	+11.19	86/90
R+var	Raw state + label variants	73.92	63.39	+10.53	85/90
N+bare	Native norm + bare labels	72.36	60.91	+11.45	88/90
Cal	Logistic calibration	69.75	62.36	+7.39	89/90
N+sp	Native norm + space-prefixed labels	73.57	62.42	+11.15	86/90
R+bare	Raw state + bare labels	72.37	61.42	+10.95	88/90
R+sp	Raw state + space-prefixed labels	71.67	60.26	+11.41	86/90
Rotate	Cyclic option rotations	74.46	62.90	+11.56	86/90

E Complete Intervention Results

E.1 Additional Outcomes on the Forward Transfer Panel

Table 6 provides complementary outcomes for the 180-question panel, including correct-answer probability, fixed-rival gap, and last revision depth. Paired transfer increases mean probability from 57.54% to 63.36% and mean gap from 1.02 to 1.75. Revision depth changes are also reported for completeness.

E.2 Source-Control Results

Table 7 reports absolute final gaps under baseline, random, cross-question, and same-question transfer. The mean gain over baseline is 0.74, with a 95% interval from 0.46 to 1.13. Same-question transfer exceeds both matched controls in all thirty-six cells. Its mean advantage is 0.62 over random directions, with an interval from 0.39 to 0.95. The corresponding advantage over cross-question

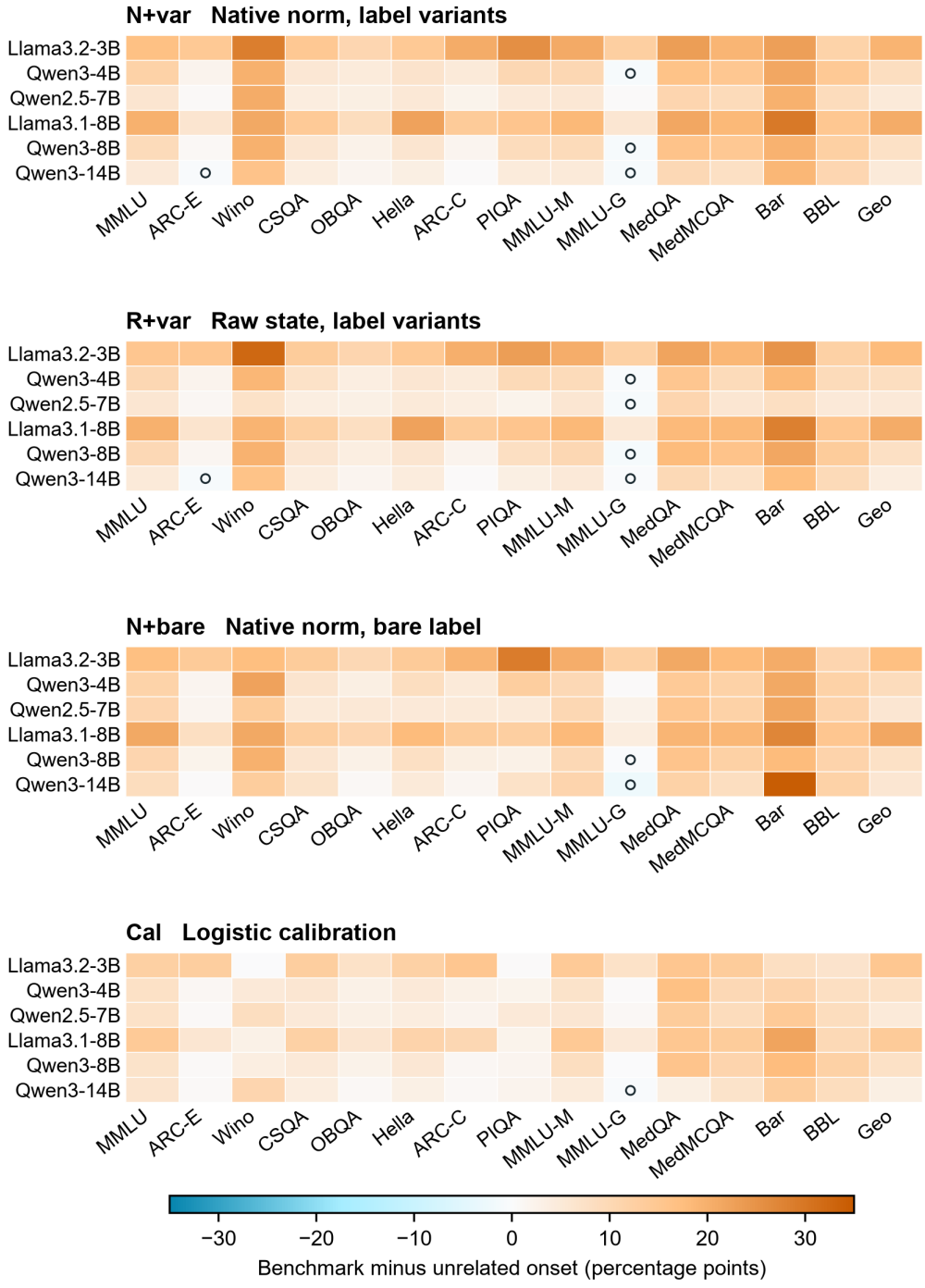


Figure 7: Model-task readout contrasts for the first four variants. Each cell shows benchmark minus unrelated persistent-preference onset. Positive values indicate later onset under benchmark inputs. Open circles mark nonpositive contrasts.

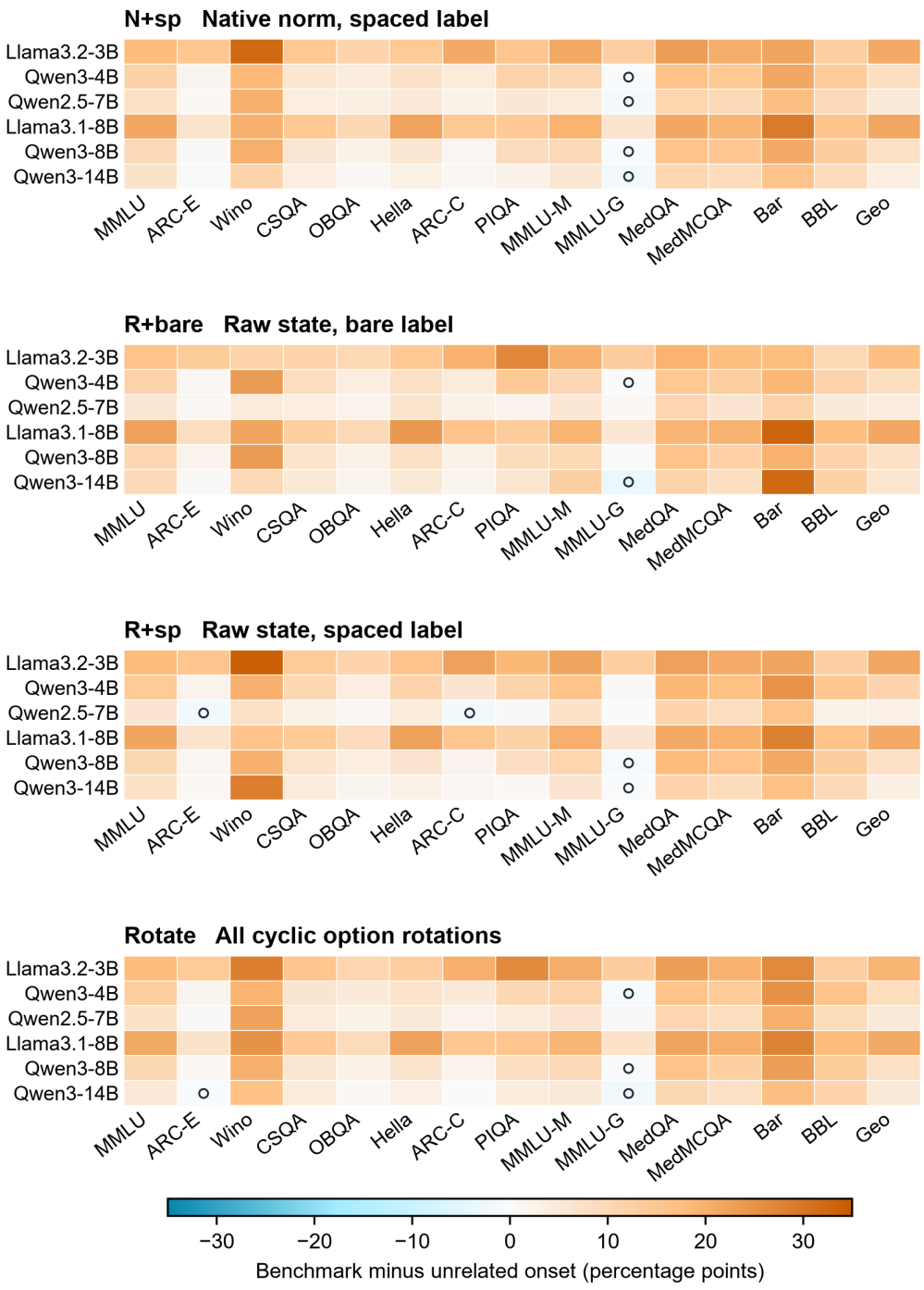


Figure 8: Model-task readout contrasts for the remaining four variants. The layout and scale follow Figure 7. Together, the two figures show all 720 model-task-variant contrasts.

Table 6: Complete non-accuracy outcomes on Confirmation I. Each model–task cell contains 30 questions. Accuracy is reported in main-text Table 1.

	Probability (%)			Fixed-rival gap			Revision depth (%)		
Task	Base	Transfer	Δ	Base	Transfer	Δ	Base	Transfer	Δ
Llama3.2-3B									
MMLU	37.94	41.93	+3.99	0.68	1.55	+0.87	84.73	84.32	-0.42
ARC-C	45.19	50.62	+5.44	1.20	2.47	+1.27	79.32	76.96	-2.35
CSQA	43.39	51.76	+8.37	1.06	3.35	+2.30	76.48	73.17	-3.31
PIQA	55.94	64.05	+8.10	0.47	1.13	+0.66	83.15	78.87	-4.29
MedQA	26.77	30.71	+3.94	-0.27	0.76	+1.03	87.10	84.64	-2.45
Geo	36.13	40.29	+4.16	0.42	1.28	+0.86	81.13	82.89	+1.76
Qwen3-4B									
MMLU	53.59	60.31	+6.72	0.79	1.43	+0.64	78.59	78.38	-0.21
ARC-C	79.72	83.27	+3.55	2.03	2.44	+0.41	73.26	72.57	-0.69
CSQA	61.78	69.90	+8.12	1.25	2.41	+1.15	77.78	76.17	-1.61
PIQA	74.31	84.86	+10.56	0.98	1.55	+0.57	62.36	59.17	-3.19
MedQA	36.73	40.78	+4.05	0.09	0.53	+0.45	83.87	85.44	+1.57
Geo	64.23	70.01	+5.78	1.21	1.67	+0.47	71.41	72.94	+1.53
Qwen2.5-7B									
MMLU	56.40	61.67	+5.28	0.92	1.45	+0.53	81.31	79.58	-1.73
ARC-C	72.43	77.13	+4.70	1.71	2.20	+0.49	78.15	78.27	+0.12
CSQA	60.13	70.16	+10.04	1.17	2.70	+1.52	79.69	80.24	+0.55
PIQA	77.68	85.65	+7.97	1.34	1.87	+0.53	65.89	66.67	+0.77
MedQA	38.18	40.48	+2.31	0.27	0.58	+0.31	82.93	84.00	+1.07
Geo	68.75	72.31	+3.56	1.49	1.82	+0.33	78.30	76.85	-1.46
Llama3.1-8B									
MMLU	40.33	45.10	+4.77	0.67	1.46	+0.79	81.38	80.00	-1.38
ARC-C	50.56	55.15	+4.59	1.44	2.29	+0.85	75.94	74.51	-1.43
CSQA	44.78	52.69	+7.91	1.06	2.57	+1.51	77.50	74.46	-3.04
PIQA	59.35	67.33	+7.98	0.63	1.19	+0.56	82.92	79.69	-3.23
MedQA	29.22	31.43	+2.21	0.28	0.89	+0.61	86.54	86.58	+0.04
Geo	38.09	41.54	+3.45	0.64	1.15	+0.51	81.77	80.36	-1.41
Qwen3-8B									
MMLU	56.79	62.23	+5.44	0.89	1.50	+0.61	74.70	72.82	-1.87
ARC-C	81.47	84.22	+2.74	2.06	2.39	+0.34	66.25	66.25	+0.00
CSQA	62.00	71.41	+9.41	1.18	2.44	+1.25	74.17	70.89	-3.28
PIQA	74.33	83.93	+9.60	0.98	1.48	+0.50	66.11	58.15	-7.96
MedQA	45.95	50.13	+4.18	0.40	0.86	+0.46	74.04	75.11	+1.07
Geo	70.52	74.32	+3.80	1.49	1.86	+0.37	67.75	67.94	+0.19
Qwen3-14B									
MMLU	62.69	68.70	+6.02	0.95	1.58	+0.62	77.33	77.12	-0.21
ARC-C	88.97	88.83	-0.13	2.23	2.49	+0.25	70.40	70.69	+0.29
CSQA	69.30	78.82	+9.52	1.39	2.64	+1.25	76.78	74.40	-2.38
PIQA	78.01	88.07	+10.06	1.18	1.72	+0.54	75.58	70.79	-4.79
MedQA	51.88	58.71	+6.84	0.66	1.25	+0.59	77.13	77.38	+0.25
Geo	77.89	82.44	+4.55	1.66	2.12	+0.46	75.40	74.23	-1.17

Each row first reports the absolute baseline and transfer values, then their paired change. Probability and depth changes use percentage points. Gap changes use the frozen score scale. Calculations use unrounded values.

transfer is 0.47, with an interval from 0.29 to 0.70. These comparisons show that the paired question's state difference carries information that matters for the final choice beyond the size of the update.

E.3 Complete Localization Results

Figures 9 and 10 provide complete localization results across all six models and six tasks. The main figures retain the first three models in the existing display order for space. These complete views retain every model-task cell, including cases where an early or middle layer matches or exceeds the selected layer. The per-model vertical scales match those in the main figures.

E.4 Additional Outcomes on the 360-Question Panel

Tables 8 and 9 extend the main bidirectional accuracy results. Forward transfer raises mean probability from 59.76% to 66.12% and the mean gap from 1.14 to 1.87. Both increase in all thirty-six cells. Reverse transfer lowers probability from 74.45% to 70.34% and the gap from 2.28 to 1.82, with decreases in thirty-three and thirty-four cells. The mean probability changes are 6.36 and -4.12 percentage points. Their pointwise 95% intervals are 4.19 to 8.89 and -7.66 to -1.48. These results provide additional probability, gap, and revision-depth outcomes corresponding to the bidirectional accuracy analysis reported in the main text.

Forward transfer shifts mean revision depth from 76.44% to 74.62%, with earlier revisions in thirty-two cells. Reverse transfer shifts it from 71.21% to 72.54%, with later revisions in thirty-four cells. The respective changes are -1.82 and 1.33 percentage points, with intervals from -3.49 to -0.72 and 0.49 to 2.21. This second panel supports an aggregate timing shift in the expected direction. The complete cell results preserve its exceptions and the weaker timing evidence in Confirmation I. Together, the panels connect paired boundary states to final preferences and to changes in decision timing under the tested interventions.

Table 7: Absolute final fixed-rival gaps under source controls on Confirmation I. Same uses the paired question. Random uses three norm-matched random directions. Cross uses a fixed donor from another question in the same task. The final columns compare Same with each control.

Task	Base	Random	Cross	Same	Same–R	Same–C
Llama3.2-3B						
MMLU	0.68	0.94	1.08	1.55	+0.61	+0.47
ARC-C	1.20	1.28	1.74	2.47	+1.20	+0.73
CSQA	1.06	1.48	2.01	3.35	+1.87	+1.34
PIQA	0.47	0.50	0.46	1.13	+0.63	+0.66
MedQA	-0.27	-0.08	0.19	0.76	+0.84	+0.57
Geo	0.42	0.60	0.71	1.28	+0.68	+0.57
Qwen3-4B						
MMLU	0.79	0.89	1.00	1.43	+0.54	+0.43
ARC-C	2.03	2.07	2.22	2.44	+0.37	+0.22
CSQA	1.25	1.38	1.70	2.41	+1.03	+0.71
PIQA	0.98	1.00	1.01	1.55	+0.56	+0.54
MedQA	0.09	0.20	0.29	0.53	+0.33	+0.24
Geo	1.21	1.32	1.39	1.67	+0.36	+0.28
Qwen2.5-7B						
MMLU	0.92	1.02	1.10	1.45	+0.44	+0.35
ARC-C	1.71	1.79	1.91	2.20	+0.41	+0.29
CSQA	1.17	1.60	1.82	2.70	+1.10	+0.88
PIQA	1.34	1.42	1.53	1.87	+0.46	+0.34
MedQA	0.27	0.36	0.38	0.58	+0.22	+0.20
Geo	1.49	1.56	1.58	1.82	+0.25	+0.24
Llama3.1-8B						
MMLU	0.67	0.83	1.01	1.46	+0.64	+0.46
ARC-C	1.44	1.48	1.84	2.29	+0.80	+0.44
CSQA	1.06	1.22	1.61	2.57	+1.35	+0.96
PIQA	0.63	0.64	0.72	1.19	+0.55	+0.47
MedQA	0.28	0.38	0.56	0.89	+0.51	+0.33
Geo	0.64	0.73	0.87	1.15	+0.42	+0.28
Qwen3-8B						
MMLU	0.89	0.97	1.02	1.50	+0.53	+0.49
ARC-C	2.06	2.01	2.18	2.39	+0.38	+0.21
CSQA	1.18	1.29	1.58	2.44	+1.15	+0.86
PIQA	0.98	1.00	1.06	1.48	+0.49	+0.42
MedQA	0.40	0.50	0.62	0.86	+0.36	+0.24
Geo	1.49	1.57	1.63	1.86	+0.29	+0.23
Qwen3-14B						
MMLU	0.95	1.01	1.10	1.58	+0.57	+0.48
ARC-C	2.23	2.24	2.31	2.49	+0.25	+0.18
CSQA	1.39	1.61	1.87	2.64	+1.03	+0.77
PIQA	1.18	1.23	1.33	1.72	+0.49	+0.39
MedQA	0.66	0.90	1.00	1.25	+0.36	+0.25
Geo	1.66	1.77	1.86	2.12	+0.35	+0.26

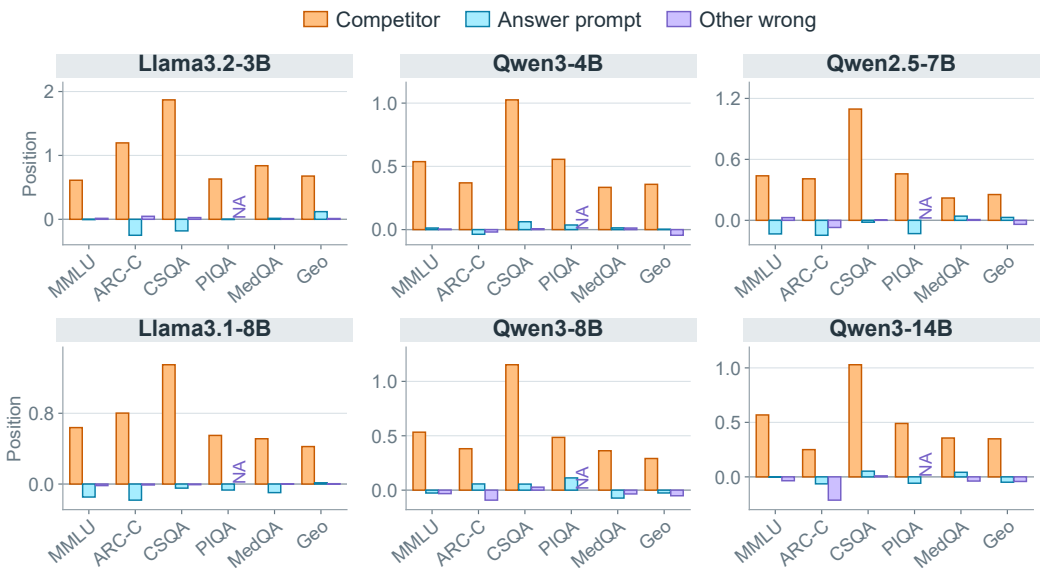


Figure 9: Complete position-control results for all six models and six tasks on Confirmation I. Each panel compares the competitor boundary, the answer-prompt position, and another wrong-option boundary. Values are paired-source effects on the scaled fixed-competitor gap after subtracting the local norm-matched random effect. PIQA has no other wrong option and is marked NA.

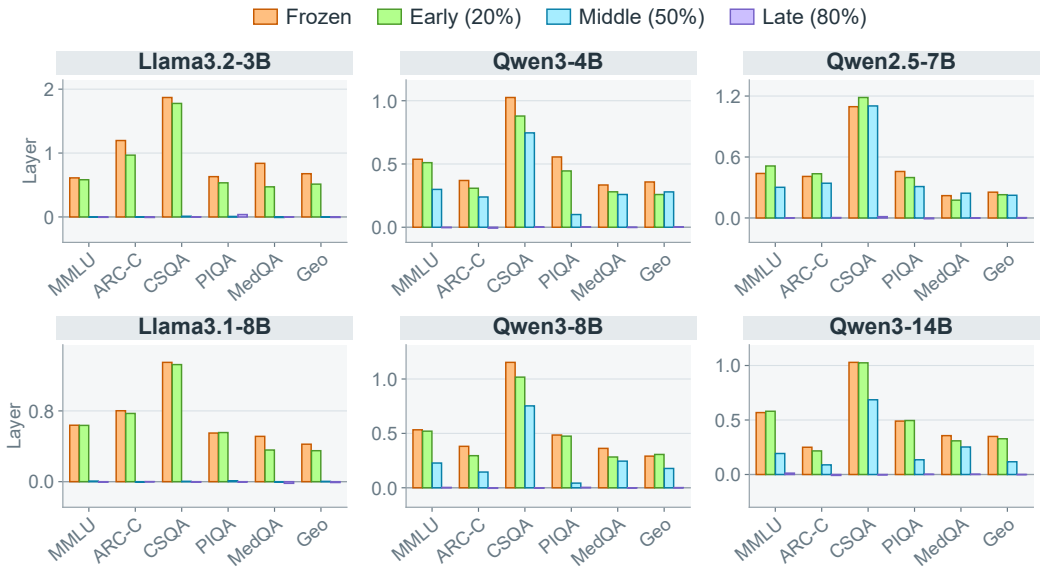


Figure 10: Complete layer-control results for all six models and six tasks on Confirmation I. Each panel compares the frozen selected layer with early, middle, and late depths. Values are paired-source effects on the scaled fixed-competitor gap after subtracting the local norm-matched random effect. Each model uses the same vertical scale as in Figure 9.

Table 8: Additional forward-transfer outcomes on Confirmation II. Each cell contains 60 questions. Base uses benchmark inputs and Transfer inserts the paired unrelated state. Main-text Table 2 reports accuracy.

	Probability (%)			Fixed-rival gap			Revision depth (%)		
Task	Base	Transfer	Δ	Base	Transfer	Δ	Base	Transfer	Δ
Llama3.2-3B									
MMLU	41.10	44.91	+3.81	0.91	1.73	+0.82	81.49	80.89	-0.60
ARC-C	52.30	59.53	+7.22	1.87	3.28	+1.41	76.07	73.29	-2.79
CSQA	41.95	50.78	+8.83	0.98	3.14	+2.16	77.88	75.01	-2.87
PIQA	57.63	64.85	+7.22	0.63	1.34	+0.72	81.73	76.19	-5.54
MedQA	34.65	39.01	+4.36	0.50	1.52	+1.01	83.99	82.55	-1.44
Geo	40.01	43.46	+3.45	0.82	1.53	+0.71	80.39	81.56	+1.18
Qwen3-4B									
MMLU	59.53	62.09	+2.56	1.06	1.44	+0.38	77.50	77.34	-0.16
ARC-C	79.02	85.65	+6.62	1.95	2.50	+0.55	72.32	71.75	-0.57
CSQA	60.57	72.45	+11.88	1.02	2.28	+1.26	76.75	76.07	-0.68
PIQA	68.13	78.95	+10.82	0.75	1.29	+0.53	66.67	61.37	-5.30
MedQA	46.49	51.85	+5.35	0.54	1.15	+0.60	83.68	81.94	-1.73
Geo	66.50	74.27	+7.77	1.21	1.79	+0.58	71.94	69.62	-2.33
Qwen2.5-7B									
MMLU	56.58	60.66	+4.08	0.99	1.36	+0.37	82.05	81.25	-0.80
ARC-C	79.56	84.22	+4.66	2.02	2.54	+0.52	76.35	76.09	-0.27
CSQA	60.81	71.34	+10.53	1.27	2.49	+1.21	79.81	79.44	-0.37
PIQA	73.70	83.14	+9.44	1.13	1.67	+0.54	67.50	65.42	-2.08
MedQA	44.09	49.52	+5.43	0.53	1.03	+0.49	84.02	83.60	-0.43
Geo	62.37	68.22	+5.85	1.12	1.65	+0.52	78.56	78.04	-0.52
Llama3.1-8B									
MMLU	43.80	47.14	+3.35	1.03	1.53	+0.51	80.83	78.57	-2.27
ARC-C	56.17	61.35	+5.18	1.80	2.63	+0.83	73.37	70.80	-2.57
CSQA	43.75	53.03	+9.29	0.86	2.54	+1.67	77.50	74.30	-3.20
PIQA	62.90	67.20	+4.30	0.88	1.28	+0.41	83.67	78.70	-4.97
MedQA	38.71	42.40	+3.69	0.87	1.57	+0.70	83.10	81.71	-1.40
Geo	40.97	45.37	+4.40	0.72	1.37	+0.65	80.77	81.05	+0.29
Qwen3-8B									
MMLU	62.51	65.00	+2.49	1.10	1.43	+0.32	74.14	73.48	-0.66
ARC-C	83.39	85.82	+2.43	2.08	2.41	+0.32	65.03	63.47	-1.55
CSQA	59.90	71.82	+11.92	0.99	2.26	+1.27	71.70	69.57	-2.13
PIQA	70.77	79.52	+8.74	0.80	1.20	+0.40	69.88	61.04	-8.84
MedQA	52.41	58.87	+6.46	0.77	1.47	+0.70	75.35	74.96	-0.39
Geo	70.42	79.00	+8.58	1.38	2.07	+0.69	69.35	65.88	-3.47
Qwen3-14B									
MMLU	68.53	70.47	+1.94	1.25	1.63	+0.38	77.23	77.26	+0.03
ARC-C	86.50	88.60	+2.09	2.06	2.43	+0.37	69.68	69.75	+0.07
CSQA	70.88	83.04	+12.15	1.33	2.49	+1.15	76.27	74.12	-2.16
PIQA	74.65	83.88	+9.23	1.01	1.46	+0.44	72.65	70.35	-2.29
MedQA	61.85	68.04	+6.18	1.08	1.84	+0.76	77.79	76.85	-0.94
Geo	78.25	84.83	+6.58	1.62	2.18	+0.56	74.72	73.00	-1.72

Each row first reports the absolute baseline and transfer values, then their paired change. Probability and depth changes use percentage points. Gap changes use the frozen score scale. Calculations use unrounded values.

Table 9: Additional reverse-transfer outcomes on Confirmation II. Each cell contains 60 questions. Base uses unrelated inputs and Transfer inserts the paired benchmark state. Negative probability and gap changes indicate reduced support for the correct answer. Positive depth changes indicate later final revisions.

	Probability (%)			Fixed-rival gap			Revision depth (%)		
Task	Base	Transfer	Δ	Base	Transfer	Δ	Base	Transfer	Δ
Llama3.2-3B									
MMLU	53.81	51.24	-2.57	2.36	1.89	-0.46	76.62	78.62	+1.99
ARC-C	66.82	64.76	-2.06	3.43	3.05	-0.38	67.13	67.74	+0.61
CSQA	59.74	52.78	-6.96	3.57	1.88	-1.69	70.92	73.81	+2.89
PIQA	72.14	72.02	-0.12	2.01	1.99	-0.02	73.36	75.24	+1.88
MedQA	53.21	49.13	-4.09	2.86	2.16	-0.69	74.20	77.39	+3.19
Geo	49.48	48.93	-0.55	1.82	1.66	-0.15	77.56	78.84	+1.28
Qwen3-4B									
MMLU	75.26	70.94	-4.32	1.88	1.58	-0.30	74.05	74.21	+0.16
ARC-C	87.62	86.07	-1.55	2.55	2.38	-0.18	71.03	71.54	+0.51
CSQA	81.35	68.04	-13.30	2.52	1.23	-1.29	73.81	75.55	+1.73
PIQA	89.10	88.20	-0.90	2.05	1.91	-0.15	54.70	55.88	+1.18
MedQA	68.60	63.26	-5.34	1.81	1.35	-0.46	76.61	77.85	+1.24
Geo	78.57	75.34	-3.24	2.11	1.73	-0.37	68.81	69.57	+0.76
Qwen2.5-7B									
MMLU	74.64	69.62	-5.02	1.90	1.55	-0.35	78.27	79.35	+1.07
ARC-C	89.50	87.37	-2.13	2.84	2.59	-0.25	75.73	76.17	+0.44
CSQA	82.26	69.50	-12.76	2.83	1.48	-1.35	78.87	79.56	+0.69
PIQA	86.90	87.69	+0.79	1.86	1.89	+0.03	67.53	65.77	-1.76
MedQA	65.13	59.91	-5.22	1.81	1.42	-0.39	81.92	82.06	+0.14
Geo	75.05	71.26	-3.79	2.01	1.60	-0.41	76.90	77.54	+0.64
Llama3.1-8B									
MMLU	57.72	56.15	-1.57	2.12	1.90	-0.23	70.05	71.17	+1.12
ARC-C	67.52	66.50	-1.02	2.79	2.62	-0.17	64.35	65.82	+1.47
CSQA	64.22	56.70	-7.52	3.00	1.74	-1.26	68.29	72.15	+3.85
PIQA	75.03	75.22	+0.19	1.73	1.75	+0.02	71.72	72.68	+0.96
MedQA	51.55	49.33	-2.23	2.12	1.77	-0.34	74.15	76.14	+1.99
Geo	55.55	53.45	-2.10	2.01	1.69	-0.32	70.94	73.74	+2.80
Qwen3-8B									
MMLU	74.10	70.42	-3.68	1.73	1.48	-0.25	68.00	69.94	+1.94
ARC-C	86.98	85.30	-1.69	2.38	2.19	-0.19	64.05	64.38	+0.32
CSQA	84.65	68.76	-15.89	2.65	1.14	-1.50	68.01	70.49	+2.48
PIQA	87.31	86.52	-0.80	1.70	1.62	-0.08	56.37	58.61	+2.25
MedQA	75.98	69.11	-6.87	2.14	1.58	-0.55	68.56	70.80	+2.24
Geo	78.83	73.21	-5.62	2.20	1.63	-0.57	64.76	67.25	+2.49
Qwen3-14B									
MMLU	80.02	76.15	-3.87	1.93	1.63	-0.30	74.54	75.01	+0.47
ARC-C	90.68	89.42	-1.27	2.47	2.37	-0.10	70.10	71.20	+1.10
CSQA	89.90	78.60	-11.30	2.63	1.53	-1.10	72.48	75.22	+2.73
PIQA	89.78	90.04	+0.26	1.83	1.83	-0.01	71.19	70.04	-1.15
MedQA	77.76	70.90	-6.86	2.15	1.67	-0.48	74.04	75.20	+1.16
Geo	83.58	80.27	-3.31	2.21	1.90	-0.32	74.02	75.00	+0.98

Each row first reports the absolute baseline and transfer values, then their paired change. Probability and depth changes use percentage points. Gap changes use the frozen score scale. Calculations use unrounded values.