

# Beyond Output-Level Correction: Tracing Positional Bias to Early-Layer Commitment in Transformer Multiple-Choice Evaluation

Soyeon Lee<sup>1,2</sup>

Minhyeok Lee<sup>1,3\*</sup>

<sup>1</sup>Department of Intelligent Semiconductor Engineering, Chung-Ang University

<sup>2</sup>GAI System Inc.

<sup>3</sup>School of Electrical and Electronics Engineering, Chung-Ang University

soyeon@gaisys.com

mlee@cau.ac.kr

## Abstract

Language models exhibit positional bias on multiple-choice benchmarks, scoring the same question differently depending on which position holds the correct answer. Existing work measures the bias at the output and corrects it by shuffling option orders, but does not explain where in the network it originates. Using causal activation patching with a formalized Average Indirect Effect (AIE) metric, we trace the layer-by-layer formation of positional commitment across three model families and benchmarks. Positive causal mass concentrates in early-to-mid layers (normalized depth below 0.60) for the higher-capacity models tested, with Pearson correlations above 0.89 across nine model-dataset conditions and three output metrics, and the localization is stable under multi-layer windowed patching. A controlled factorization that permutes option labels and option order independently shows this profile tracks the label-identity component of the bias rather than generic order sensitivity. Sublayer decomposition reveals MLP-dominated early layers (about 59%) and attention-dominated mid-layers (about 76%), consistent with label information being integrated through attention after semantic content is encoded by MLPs. Negative controls show no detectable early-layer separation between robustly correct and bias-manifesting questions, supporting a shared answer-commitment pathway rather than a dedicated bias circuit and explaining why output-level shuffling stabilizes scores without altering internal preference.

## 1 Introduction

When a language model is given a multiple-choice question, the answer it produces should depend on the content of the options, not on whether the correct option happens to be labeled A or D. In practice, this is not what happens. The same model on the same question can answer correctly when the

options are presented in one order and incorrectly when they are cyclically rotated. This phenomenon, known as positional bias, has been documented across MMLU, ARC-Challenge, HellaSwag, and other standard benchmarks (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024; Wei et al., 2024), and it raises a basic concern about whether benchmark scores measure the abilities they are designed to measure.

The standard response in the evaluation literature is to apply corrections at the output, shuffling options across runs, averaging predictions over cyclic permutations, or calibrating output probabilities (Zhao et al., 2021; Xue et al., 2024; Choi et al., 2025; Yoon et al., 2025). Whether such correction is principled depends on where the bias forms. If positional bias is a shallow distortion arising at the final softmax, output-level calibration addresses it at its source; if the bias is integrated into intermediate representations long before the output, calibration masks an internal property of the computation, and a different kind of intervention is needed. Distinguishing these possibilities requires identifying where, within the residual stream, the commitment to a positional preference is formed. We address this question using causal activation patching with the Average Indirect Effect (AIE) as the primary metric (Vig et al., 2020; Meng et al., 2022), which assigns a quantitative causal score to each layer’s contribution to the model’s prediction.

Our contributions are fourfold: (i) a permutation-invariant AIE formulation whose conclusions hold across three output metrics (Pearson  $r > 0.89$ ) and under multi-layer windowed patching; (ii) large-scale all-pairs validation (up to  $N = 1,142$  bias-manifesting questions) showing that early-layer causal dominance is not an artifact of question or pair selection; (iii) a controlled factorization identifying the label-identity component as the primary carrier of the early-to-mid profile, together with a staged MLP-to-attention attribution pattern consis-

\* Corresponding author.

tent across the three tested decoder-only families; and (iv) negative controls establishing that positional bias travels through the general commitment pathway shared with correct prediction, which reframes debiasing as a head-level problem within the mid-layer attention zone, for which a per-head screening supplies initial candidate heads.

## 2 Related Work

Positional bias in multiple-choice evaluation is well documented (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024; Wei et al., 2024). Mitigation has largely operated at the output, through calibration (Zhao et al., 2021), strengthened symbol binding (Xue et al., 2024), auxiliary options (Choi et al., 2025), and order-invariant encodings (Yoon et al., 2025), with a smaller line intervening on internal representations (Yu et al., 2025; Wang et al., 2025b); these approaches quantify or reduce the bias but leave open where in the network it forms.

Closest to our setting, Wiegrefe et al. (2025) attributed answer-symbol prediction in formatted multiple-choice answering to a few middle layers and their self-attention mechanisms. That analysis characterizes answer selection in general; the present work conditions on bias manifestation, contrasting bias-manifesting questions with robustly correct and robustly wrong controls to ask why the selected answer depends on option order, a distinction between shared and dedicated circuitry that an unconditioned analysis of answer selection cannot draw. Our methodology follows the causal mediation and activation-patching tradition (Vig et al., 2020; Meng et al., 2022; Zhang and Nanda, 2024; Heimersheim and Nanda, 2024), and the staged organization we observe connects to accounts of feed-forward layers as key-value memories (Geva et al., 2021, 2022), to head-level circuit analyses such as induction heads (Elhage et al., 2021), and to the causal localization of factual recall in mid-layer MLPs (Meng et al., 2022); Appendix A gives a fuller survey.

## 3 Background and Formal Setup

The experimental logic can be stated before the formalism. For a question the model answers correctly under one cyclic rotation of its options and incorrectly under another, we replace activations of the incorrect run with those of the correct run, one layer at a time; the depth at which restoration succeeds reveals where the decisive commit-

ment forms, the attention/MLP decomposition reveals which component carries it, and comparisons against robustly correct and robustly wrong questions reveal whether that commitment is specific to the bias. The remainder of this section formalizes these quantities.

### 3.1 Multiple-Choice Evaluation and Positional Bias

Let  $\mathcal{Q}$  denote a set of multiple-choice questions. Each question  $q \in \mathcal{Q}$  consists of a prompt  $p_q$  and four answer options  $\{o_1, o_2, o_3, o_4\}$ , exactly one of which is correct. The standard evaluation protocol presents these options in a fixed order and queries the language model for the probability of each option token (Robinson and Wingate, 2023), selecting the highest-probability option as the model’s answer.

**Definition 1** (Positional Preference). *A model  $\mathcal{M}$  exhibits positional preference on question  $q$  if there exists a permutation  $\sigma \in S_4$  of the option ordering such that  $\mathcal{M}$  answers correctly under the original ordering but incorrectly under  $\sigma$ , where the incorrect answer corresponds to the same absolute position that was correct in the original ordering.*

We operationalize this through cyclic rotations. For a question with options  $(o_1, o_2, o_3, o_4)$ , the  $k$ -th cyclic rotation produces the ordering  $(o_{1+k}, o_{2+k}, o_{3+k}, o_{4+k})$  with indices taken modulo 4. A question is bias-manifesting if the model answers correctly under at least one rotation and incorrectly under at least one other rotation. A cyclic rotation jointly permutes the option labels, their relative order, and their surface positions; Section 5.4 separates these factors experimentally and shows that the causal profile measured under cyclic rotation tracks the label-identity component.

Let  $\Pi(q) = \{r_0, r_1, r_2, r_3\}$  denote the set of four cyclic rotations of question  $q$ . Define  $\mathcal{Q}_{\text{bias}} \subseteq \mathcal{Q}$  as the set of bias-manifesting questions.

### 3.2 Residual Stream Decomposition

A transformer (Vaswani et al., 2017) with  $L$  layers processes input tokens through a residual stream. At each layer  $l \in \{0, 1, \dots, L-1\}$ , the hidden state is updated by two sublayer operations.

**Definition 2** (Residual Stream Update). *Let  $\mathbf{h}_l \in \mathbb{R}^d$  denote the residual stream state at the final*

sequence position after layer  $l$ . The update rule is:

$$\mathbf{h}_l^{\text{attn}} = \mathbf{h}_{l-1} + \text{Attn}_l(\mathbf{h}_{l-1}) \quad (1)$$

$$\mathbf{h}_l = \mathbf{h}_l^{\text{attn}} + \text{MLP}_l(\mathbf{h}_l^{\text{attn}}) \quad (2)$$

where  $\text{Attn}_l$  denotes the self-attention sublayer and  $\text{MLP}_l$  denotes the feedforward sublayer at depth  $l$ .

This decomposition is the foundation of the sublayer attribution analysis. By patching either the full residual stream  $\mathbf{h}_l$ , the attention output  $\text{Attn}_l(\mathbf{h}_{l-1})$ , or the MLP output  $\text{MLP}_l(\mathbf{h}_l^{\text{attn}})$ , we can measure the causal effect transmitted through each sublayer type at each depth (Elhage et al., 2021).

The final prediction is obtained by applying the unembedding matrix  $W_U \in \mathbb{R}^{|\mathcal{V}| \times d}$  to  $\mathbf{h}_{L-1}$ :

$$\text{logit}_v = W_U[v, :] \cdot \mathbf{h}_{L-1} \quad (3)$$

where  $v$  indexes vocabulary tokens.

### 3.3 Output Metrics over Answer Options

For the correct answer token  $t^*$  of a question  $q$ , we use three output metrics: the restricted softmax over the four option tokens (M1), the full-vocabulary softmax (M2), and the raw logit (M3),

$$P_{M1}(t^* | q) = \frac{\exp(\text{logit}_{t^*})}{\sum_{j=1}^4 \exp(\text{logit}_{t_j})}, \quad (4)$$

$$P_{M2}(t^* | q) = \frac{\exp(\text{logit}_{t^*})}{\sum_{v \in \mathcal{V}} \exp(\text{logit}_v)}, \quad (5)$$

and  $S_{M3}(t^* | q) = \text{logit}_{t^*}$ , where  $\{t_1, \dots, t_4\}$  are the option tokens of the current rotation. M3 is a score rather than a probability; the label is kept for uniform notation. If M1 inflated early-layer estimates through restricted normalization, the conclusions would be metric-dependent; Section 5.1 addresses this concern. We write  $f_M$  when a derivation applies uniformly across the three metrics.

### 3.4 Causal Activation Patching and the AIE

Causal activation patching is the operation of replacing an intermediate activation in one forward pass (the corrupted run) with the corresponding activation from a different forward pass (the clean run), then measuring the resulting change in output (Zhang and Nanda, 2024; Heimersheim and Nanda, 2024). Because the residual stream is cumulative, restoring an activation at layer  $l$  also changes the input to every subsequent layer; the resulting effect is therefore the total causal effect of that

activation through the model’s downstream computation, rather than a measure of information locally encoded at layer  $l$ . Section 5.3 shows that the resulting localization is robust to patching granularity.

**Definition 3** (Patching Operation). Let  $\mathbf{h}_l^{\text{clean}}$  and  $\mathbf{h}_l^{\text{corr}}$  denote the residual stream states at layer  $l$  under the clean and corrupted forward passes respectively. The patched probability at layer  $l$  is:

$$P_{\text{patch}}(t^* | q, l) = P_{M1}(t^* | \mathbf{h}_l \leftarrow \mathbf{h}_l^{\text{clean}}, \text{corrupted context}) \quad (6)$$

where the arrow denotes substitution of the activation at layer  $l$  while all other activations follow the corrupted forward pass.

**Definition 4** (Average Indirect Effect at Layer  $l$ ). For a bias-manifesting question  $q$  with clean rotation  $r_c$  (a rotation where the model is correct) and corrupted rotation  $r_d$  (a rotation where the model is wrong):

$$\text{AIE}_l(q) = P_{\text{patch}}(t^* | q, l) - P_{M1}(t^* | r_d) \quad (7)$$

The Average Indirect Effect at layer  $l$  across  $\mathcal{Q}_{\text{bias}}$  is:

$$\overline{\text{AIE}}_l = \frac{1}{|\mathcal{Q}_{\text{bias}}|} \sum_{q \in \mathcal{Q}_{\text{bias}}} \text{AIE}_l(q) \quad (8)$$

A positive  $\overline{\text{AIE}}_l$  indicates that patching layer  $l$ ’s residual stream with clean activations causally restores the correct prediction. A negative value indicates that patching suppresses the correct prediction. The suppression phenomenon is analyzed in Section 6.1.

### 3.5 Early Fraction and H1

**Definition 5** (Early Fraction). Let  $\hat{l} = 0.6L$  denote the early-layer threshold at 60% of total depth. The Early Fraction is:

$$\text{EarlyFrac} = \frac{\sum_{l=0}^{\hat{l}} \max(\overline{\text{AIE}}_l, 0)}{\sum_{l=0}^{L-1} \max(\overline{\text{AIE}}_l, 0)} \quad (9)$$

This metric measures the fraction of total positive causal mass that resides in the first 60% of the network. The denominator sums only positive contributions, so  $\text{EarlyFrac} \in [0, 1]$ .

The threshold is expressed in normalized depth so that models with different layer counts are directly comparable, and  $0.6L$  separates the early-to-mid region, where answer commitment could plausibly form, from the final third of the network,

where output construction dominates; Appendix D verifies that the conclusions are stable across a neighborhood of this choice.

**Definition 6** (Hypothesis H1). *H1 states that  $\text{EarlyFrac} \geq \tau$  where  $\tau = 0.30$  is the primary threshold. To guard against degenerate cases where all positive mass is near zero, we additionally require  $\sum_{l=0}^{L-1} \max(\overline{\text{AIE}}_l, 0) \geq 0.01$  (the  $\text{TotalPosAIE}$  guard).*

**Definition 7** (Total Positive AIE).

$$\text{TotalPosAIE} = \sum_{l=0}^{L-1} \max(\overline{\text{AIE}}_l, 0) \quad (10)$$

A value of  $\text{TotalPosAIE} < 0.01$  flags cases where the AIE signal is so small that the  $\text{EarlyFrac}$  estimate is not meaningful.

### 3.6 Sublayer Attribution

To decompose causal mass between attention and MLP components, we define the fractional attribution at each layer.

**Definition 8** (Attention Attribution Fraction).

$$\text{AttnFrac}_l = \frac{|\overline{\text{AIE}}_l^{\text{attn}}|}{|\overline{\text{AIE}}_l^{\text{attn}}| + |\overline{\text{AIE}}_l^{\text{mlp}}|} \quad (11)$$

where  $\overline{\text{AIE}}_l^{\text{attn}}$  is computed by patching only the attention sublayer output and  $\overline{\text{AIE}}_l^{\text{mlp}}$  by patching only the MLP sublayer output. The denominator uses absolute values to measure relative magnitude of contribution regardless of sign.

We aggregate this measure over three depth zones. Let  $Z_E = [0, 0.33L)$ ,  $Z_M = [0.33L, 0.66L)$ , and  $Z_L = [0.66L, L)$  denote the early, mid, and late zones respectively. The zone-level attribution is:

$$\overline{\text{AttnFrac}}_Z = \frac{1}{|Z|} \sum_{l \in Z} \text{AttnFrac}_l \quad (12)$$

### 3.7 Representational Incompatibility

The negative AIE values observed in late layers require a separate mechanistic account. We hypothesize (H-suppress) that late-layer representations are too context-specific to transfer across permutation contexts, creating interference rather than restoration.

**Definition 9** (Cosine Distance between Clean and Corrupted Streams).

$$\text{CosDist}(l, q) = 1 - \frac{\mathbf{h}_l^{\text{clean}}(q) \cdot \mathbf{h}_l^{\text{corr}}(q)}{\|\mathbf{h}_l^{\text{clean}}(q)\| \cdot \|\mathbf{h}_l^{\text{corr}}(q)\|} \quad (13)$$

The H-suppress hypothesis predicts that  $\text{CosDist}(l, q)$  is larger in late layers than in early layers, and that this increased cosine distance is associated with more negative  $\overline{\text{AIE}}_l$ .

**Definition 10** (Cosine Distance Ratio).

$$\text{CDRatio} = \frac{\overline{\text{CosDist}}_{Z_L}}{\overline{\text{CosDist}}_{Z_E}} \quad (14)$$

where  $\overline{\text{CosDist}}_Z$  is the mean cosine distance over all questions and all layers in zone  $Z$ . H-suppress predicts  $\text{CDRatio} > 2$  when late-layer negative AIE is observed.

## 4 Experimental Setup

### 4.1 Models

We evaluate three transformer-based instruction-tuned language models representing distinct architectural families. Llama-3.2-3B-Instruct is a 28-layer grouped-query attention (GQA) model from Meta with  $d = 3072$ . SmolLM3-3B is a 36-layer model from HuggingFace with  $d = 3072$ . Qwen3-1.7B is a 28-layer GQA model from Alibaba with  $d = 2048$ .

These families differ in attention grouping ratio, MLP activation function, and positional-encoding parameterization, enabling cross-architecture comparison. Qwen3-4B additionally enters the patching-granularity, label-factorization, suppression, and negative-control analyses (Sections 5.3, 5.4, 6.1, and 6.2); Qwen3-8B and the base variants of Qwen3-4B and Qwen3-8B enter the scale and post-training analysis in Appendix B. The attention-versus-MLP decomposition uses Qwen3-1.7B, Llama-3.2-3B-Instruct, and SmolLM3-3B; including Qwen3-1.7B, which rejects H1 on MMLU and ARC-Challenge, tests whether the sublayer organization holds independently of where positive causal mass accumulates.

### 4.2 Datasets

We evaluate on three benchmarks. MMLU (Hendrycks et al., 2021) spans 57 academic subjects with four-choice questions; we use a 2,500-question subsample of the full test set for large-scale analysis. ARC-Challenge (Clark et al., 2018) consists of grade-school science questions; we use the complete test partition of 1,172 questions. HelLaSwag (HSwag) (Zellers et al., 2019) evaluates commonsense reasoning with four-choice sentence completion; we use 200-question samples from the

validation set. This selection covers factual recall, scientific reasoning, and commonsense inference across a range of difficulty levels.

### 4.3 Pair Selection and All-Pairs Averaging

A methodological concern with causal patching for positional bias is the choice of which clean/corrupted pair to use per question. For a bias-manifesting question with  $k$  correct rotations and  $4 - k$  incorrect rotations, there are  $k \times (4 - k)$  valid pairs. Taking only the first pair introduces an arbitrary dependency on iteration order.

We address this by averaging AIE profiles over all valid pairs per question:

$$\overline{\text{AIE}}_l(q) = \frac{1}{|\mathcal{P}(q)|} \sum_{(c,d) \in \mathcal{P}(q)} \text{AIE}_l(q, c, d) \quad (15)$$

where  $\mathcal{P}(q)$  is the set of all valid (clean rotation, corrupted rotation) pairs for question  $q$ . This eliminates selection bias and provides a more stable estimate. The final mean profile is  $\overline{\text{AIE}}_l = |\mathcal{Q}_{\text{bias}}|^{-1} \sum_q \overline{\text{AIE}}_l(q)$ .

### 4.4 Bootstrap Confidence Intervals

We compute 95% bootstrap confidence intervals (CIs) for EarlyFrac by resampling questions with replacement  $B = 2,000$  times and recomputing EarlyFrac for each bootstrap sample. Statistical significance for H1 is declared when the lower bound of the 95% CI exceeds  $\tau = 0.30$ .

### 4.5 Implementation Details

All experiments use PyTorch with the HuggingFace Transformers library in float16 precision. Patching intercepts and replaces the relevant activation during the forward pass; full-residual patching replaces the block output, and attention-only and MLP-only patching replace the respective sublayer output before the residual addition (Equations 1 and 2). Multi-seed runs are reported as mean  $\pm$  standard deviation, and full implementation details appear in Appendix L.

## 5 Results

### 5.1 Metric Robustness

Because the restricted softmax metric M1 could in principle inflate early-fraction estimates through normalization effects, we compute AIE profiles under all three metrics from a single patched forward pass per layer, so the profiles are evaluated on identical patched states.

| Model      | Dataset | EF <sub>M1</sub> | EF <sub>M2</sub> | EF <sub>M3</sub> | H1  | Agreement |
|------------|---------|------------------|------------------|------------------|-----|-----------|
| Qwen3-1.7B | MMLU    | 0.194            | 0.193            | 0.145            | No  | Agree     |
| Qwen3-1.7B | ARC-C   | 0.178            | 0.176            | 0.047            | No  | Agree     |
| Qwen3-1.7B | HSwag   | 0.500            | 0.532            | 0.506            | Yes | Agree     |
| SmolLM3-3B | MMLU    | 0.719            | 0.716            | 0.755            | Yes | Agree     |
| SmolLM3-3B | ARC-C   | 0.671            | 0.666            | 0.672            | Yes | Agree     |
| SmolLM3-3B | HSwag   | 0.785            | 0.803            | 0.836            | Yes | Agree     |
| Llama-3.2  | MMLU    | 1.000            | 1.000            | 1.000            | Yes | Agree     |
| Llama-3.2  | ARC-C   | 1.000            | 1.000            | 1.000            | Yes | Agree     |
| Llama-3.2  | HSwag   | 1.000            | 1.000            | 1.000            | Yes | Agree     |

Table 1: EarlyFrac under three output metrics (M1 restricted softmax, M2 full-vocabulary softmax, M3 raw logit) for all nine model-dataset combinations. EF values are means over three seeds.

Figure 1 overlays the three metric profiles for each model, averaged across datasets and seeds; the curves are nearly coincident in all nine model-dataset combinations, so the shape of the AIE profile is consistent across the three metric variants.

Table 1 reports the EarlyFrac values under each metric. All nine model-dataset combinations show three-way agreement on whether H1 is supported. Llama-3.2-3B-Instruct returns EarlyFrac = 1.000 across all three metrics. SmolLM3-3B and Qwen3-1.7B (HellaSwag) support H1 under all three metrics, while Qwen3-1.7B on MMLU and ARC-Challenge rejects H1 under all three. The agreement indicates that any rejection is metric-independent rather than a normalization artifact. Pairwise profile correlations confirm agreement at the layer-location level:  $r(\text{M1}, \text{M2}) > 0.998$  in every condition and  $r(\text{M1}, \text{M3})$  ranges from 0.899 to 0.996. Detailed correlation values are reported in Appendix C.

### 5.2 Large-Scale Validation of Early-Layer Causal Dominance

To establish that early-layer causal dominance holds at scale and under the all-pairs averaging protocol, we evaluate Llama-3.2-3B-Instruct and SmolLM3-3B on 2,500 MMLU questions and the full 1,172-question ARC-Challenge test set, averaging AIE over all valid pair combinations per question as in Equation 15.

The depth-decile AIE profiles (Figure 4, Appendix E) are nearly identical under the two methods, indicating that the first valid-pair approach did not systematically favor the proposed hypothesis; across both models and both datasets, the profiles show near-zero AIE in deciles D1 to D4, a sharp positive peak at D5 or D6, and strongly negative values thereafter.

Table 2 reports the key metrics. EarlyFrac re-

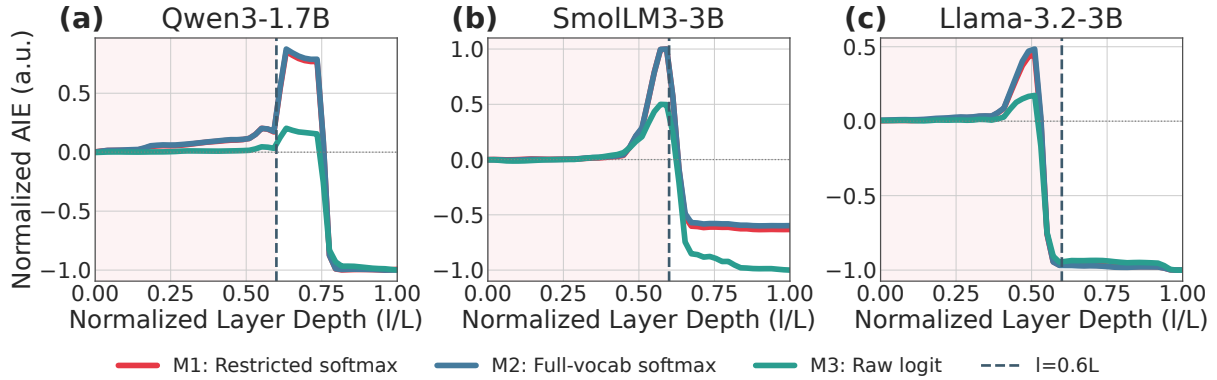


Figure 1: Normalized AIE profiles for Qwen3-1.7B, SmolLM3-3B, and Llama-3.2-3B-Instruct under three metric variants (M1: restricted softmax, M2: full-vocabulary softmax, M3: raw logit), averaged across MMLU, ARC-Challenge, and HellaSwag and normalized to unit maximum absolute value. The dashed vertical line marks the  $l = 0.6L$  threshold.

| Model      | Dataset | $N_{\text{bias}}$ | EF    | 95% CI         | TPA   | H1  |
|------------|---------|-------------------|-------|----------------|-------|-----|
| Llama-3.2  | MMLU    | 1142              | 1.000 | [1.00, 1.00]   | 0.118 | Yes |
| Llama-3.2  | ARC-C   | 351               | 1.000 | [1.00, 1.00]   | 0.163 | Yes |
| SmolLM3-3B | MMLU    | 1031              | 0.685 | [0.673, 0.699] | 0.373 | Yes |
| SmolLM3-3B | ARC-C   | 296               | 0.681 | [0.661, 0.699] | 0.613 | Yes |

Table 2: Large-scale all-pairs averaged results for Llama-3.2-3B-Instruct and SmolLM3-3B. EF is EarlyFrac, TPA is TotalPosAIE.

mains at or above the H1 threshold in every condition, with  $N_{\text{bias}}$  up to 1,142 and tight bootstrap intervals, and Gini coefficients of 0.82 to 0.90 indicate that causal mass concentrates in a small number of layers.

Stratifying bias-manifesting questions by preferred answer position does not reveal significant variation in EarlyFrac with which absolute position the model prefers; full chi-squared results are reported in Appendix G.

### 5.3 Robustness to Patching Granularity

If the early-to-mid localization reflected downstream consumption of an arbitrary early perturbation rather than the depth at which commitment forms, jointly patching wider contiguous windows of layers should shift or dilute the profile. Table 3 reports patching of contiguous two- and three-layer blocks on Qwen3-4B (100 questions per dataset). The windowed profiles remain highly correlated with the single-layer profile ( $r \geq 0.848$ ), and the early-to-mid concentration strengthens as the window widens (EarlyFrac 0.727  $\rightarrow$  0.837 on MMLU, 0.673  $\rightarrow$  0.885 on HellaSwag), so the localization is not an artifact of patching granularity.

| Dataset | Window   | EF    | 95% CI         | Peak  | $r$ w/ single |
|---------|----------|-------|----------------|-------|---------------|
| MMLU    | single   | 0.727 | [0.445, 0.870] | 0.639 | 1.000         |
| MMLU    | 2 layers | 0.794 | [0.553, 0.904] | 0.611 | 0.883         |
| MMLU    | 3 layers | 0.837 | [0.651, 0.921] | 0.583 | 0.848         |
| HSwag   | single   | 0.673 | [0.413, 0.809] | 0.583 | 1.000         |
| HSwag   | 2 layers | 0.806 | [0.641, 0.927] | 0.556 | 0.939         |
| HSwag   | 3 layers | 0.885 | [0.773, 0.956] | 0.528 | 0.897         |

Table 3: Patching-granularity control on Qwen3-4B. EF is EarlyFrac (M1); Peak is the normalized depth of the positive AIE peak;  $r$  is the correlation with the single-layer profile.

### 5.4 Disentangling Label Preference from Order Sensitivity

Cyclic rotation jointly permutes option labels, relative order, and surface positions. Two controlled conditions on Qwen3-4B separate these factors: a label-only condition fixes option content and order while permuting the label assignment, and a position-only condition fixes the labels while permuting the surface order of the options. Table 4 shows that the label-only profile closely tracks the standard cyclic profile ( $r = 0.918$  on MMLU, 0.894 on HellaSwag), whereas the position-only condition produces a qualitatively different late-peaking profile that is strongly anti-correlated with it ( $r = -0.955$  and  $-0.936$ ). The early-to-mid concentration measured under cyclic rotation is therefore primarily aligned with the label-identity component of the bias, so the phenomenon localized here is best described as a label-associated preference expressed under order permutation.

| Dataset | Cyclic | Label | Position | $r(\text{lab})$ | $r(\text{pos})$ |
|---------|--------|-------|----------|-----------------|-----------------|
| MMLU    | 0.727  | 0.622 | 0.048    | 0.918           | -0.955          |
| HSWag   | 0.673  | 0.633 | 0.114    | 0.894           | -0.936          |

Table 4: Label/order factorization on Qwen3-4B: EarlyFrac under the cyclic, label-only, and position-only conditions, and layerwise correlations of the latter two with the cyclic profile.

## 5.5 Staged MLP-to-Attention Attribution Pattern

Decomposing AIE into attention sublayer and MLP sublayer contributions reveals the staged organization of processing across model families (Geva et al., 2021, 2022). Table 5 reports zone-level attribution averaged over datasets and seeds for the three model families used in the decomposition experiment: Qwen3-1.7B, Llama-3.2-3B-Instruct, and SmolLM3-3B.

Figure 2 shows the stacked attribution bars across all model families. The pattern is consistent. Early layers are MLP-dominated, with an average early-zone attention fraction of 0.413 across the three architectures, so MLP accounts for approximately 59% of early-layer causal attribution. Mid-layers shift to attention-dominated processing, with an average mid-zone attention fraction of 0.759 across the nine model-dataset conditions, so attention accounts for approximately 76% of mid-layer attribution. This 35-percentage-point shift from MLP to attention between the early and mid zones holds at the family-averaged level for all three model families, including Qwen3-1.7B. Among the nine individual conditions, Qwen3-1.7B on HellaSwag is the one departure, with near-balanced mid-zone attribution (attention fraction 0.499); the remaining eight conditions show clear mid-zone attention dominance.

Qwen3-1.7B rejects H1 on MMLU and ARC-Challenge, yet its decomposition follows the same MLP-early, attention-mid organization: the staged sublayer pattern appears even where positive causal mass peaks later than the EarlyFrac threshold, so H1 status and sublayer organization vary independently.

In all nine model-dataset combinations, the attention-only peak layer precedes the MLP-only peak layer (0.518L versus 0.929L for Llama-3.2-3B-Instruct, 0.528L versus 0.713L for SmolLM3-3B, and 0.607L versus 0.768L for Qwen3-1.7B on MMLU). This ordering is consistent with attention,

| Model                   | Zone  | AttnFrac     | MLPFrac      | Dom.        |
|-------------------------|-------|--------------|--------------|-------------|
| Qwen3-1.7B <sup>†</sup> | Early | 0.408        | 0.592        | MLP         |
|                         | Mid   | 0.697        | 0.303        | Attn        |
|                         | Late  | 0.420        | 0.580        | MLP         |
| Llama-3.2               | Early | 0.420        | 0.580        | MLP         |
|                         | Mid   | 0.814        | 0.186        | Attn        |
|                         | Late  | 0.353        | 0.647        | MLP         |
| SmolLM3-3B              | Early | 0.410        | 0.590        | MLP         |
|                         | Mid   | 0.772        | 0.228        | Attn        |
|                         | Late  | 0.525        | 0.475        | Attn        |
| Cross-arch. early avg.  |       | <b>0.413</b> | <b>0.587</b> | <b>MLP</b>  |
| Cross-arch. mid avg.    |       | <b>0.759</b> | <b>0.241</b> | <b>Attn</b> |

Table 5: Zone-level attention versus MLP attribution fractions for three model families, averaged over MMLU, ARC-Challenge, and HellaSwag. <sup>†</sup>Qwen3-1.7B rejects H1 on MMLU and ARC-Challenge (EarlyFrac below 0.30). Cross-arch. early/mid avg. denote cross-architecture averages computed across all nine model-dataset conditions.

the sublayer through which positional structure enters the residual stream (Su et al., 2024; Voita et al., 2019), integrating option-label information at mid-depths, after which MLP layers process the integrated representation.

The Pearson correlation between per-layer attention-AIE and MLP-AIE profiles is near zero in all conditions (ranging from -0.244 to +0.179; see Appendix J), indicating that the two sublayer types do not act as linearly redundant copies of the same signal. Sensitivity of H1 to the threshold parameters  $\tau$  and  $c$  is examined in Appendix D; H1 holds across the default neighborhood for SmolLM3-3B and Llama-3.2-3B-Instruct.

## 6 Analysis

### 6.1 Late-Layer Representational Incompatibility

Late-layer patching produces strongly negative AIE values across all models, a pattern consistent with late-layer representations being too context-specific to transplant across permutation contexts (Liu et al., 2024; Hsieh et al., 2024). Late and early layers separate cleanly in the AIE-cosine-distance plane, with late layers combining high clean-corrupted cosine distance with negative AIE and early layers showing the opposite pattern (Figure 5, Appendix H).

Table 6 reports cosine distance ratios (late zone divided by early zone) for each model and dataset. The ratios range from  $1.84\times$  to  $12.45\times$  across the six tested conditions. Five of the six conditions ex-

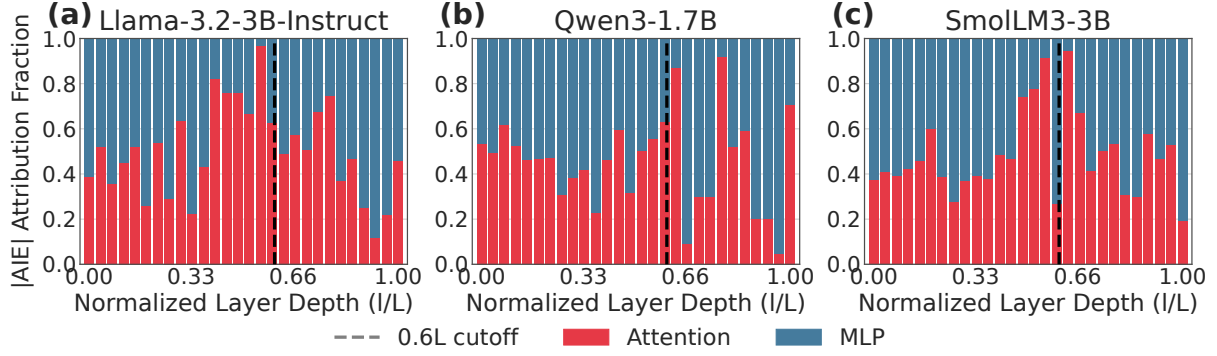


Figure 2: Stacked bar charts showing attention (red) and MLP (blue) fractional attribution of positional bias AIE by normalized layer depth bin, for Llama-3.2-3B-Instruct, Qwen3-1.7B, and SmoLM3-3B averaged across all three datasets. The dashed vertical line marks  $l = 0.6L$ .

| Model     | Data  | $\bar{d}_{\text{early}}$ | $\bar{d}_{\text{late}}$ | Ratio | $\overline{\text{AIE}}_{\text{late}}$ |
|-----------|-------|--------------------------|-------------------------|-------|---------------------------------------|
| Llama-3.2 | MMLU  | $7.1 \times 10^{-3}$     | $3.8 \times 10^{-2}$    | 5.39  | -0.081                                |
| Llama-3.2 | HSwag | $1.7 \times 10^{-2}$     | $5.0 \times 10^{-2}$    | 2.94  | -0.097                                |
| Qwen3-4B  | MMLU  | $1.9 \times 10^{-3}$     | $2.4 \times 10^{-2}$    | 12.45 | -0.079                                |
| Qwen3-4B  | HSwag | $6.1 \times 10^{-3}$     | $2.3 \times 10^{-2}$    | 3.72  | -0.064                                |
| SmoLM3-3B | MMLU  | $6.9 \times 10^{-3}$     | $5.4 \times 10^{-2}$    | 7.81  | -0.074                                |
| SmoLM3-3B | HSwag | $2.3 \times 10^{-2}$     | $4.3 \times 10^{-2}$    | 1.84  | -0.039                                |

Table 6: Cosine distance between clean and corrupted residual streams in early and late zones, with the late-to-early ratio and mean late-zone AIE.

ceed the H-suppress threshold of  $\text{CDRatio} > 2$ ; the sixth (SmoLM3-3B on HellaSwag, ratio  $1.84\times$ ) shows negative late-zone mean AIE of  $-0.039$ , indicating that suppression is also observed at a ratio below 2. Qwen3-4B on MMLU shows the largest observed ratio at  $12.45\times$ , corresponding to a late-zone mean AIE of  $-0.079$ . Llama-3.2-3B-Instruct shows a ratio of  $5.39\times$  on MMLU with late-zone mean AIE of  $-0.081$ .

When a clean late-layer representation is injected into a corrupted context forward pass, the representation is optimized for a different option ordering than the one currently being processed. Rather than restoring the model’s attention to the correct option, it redirects probability mass to wrong options. The redistribution is selective rather than uniform: across the six tested conditions, patching at late layers produces negative  $\Delta P(\text{correct})$  paired with positive  $\Delta P(\text{wrong})$  rather than a simultaneous drop in all option probabilities. Detailed redistribution values are reported in Appendix H.

## 6.2 Negative Controls and the Shared Semantic Commitment Pathway

Whether the early causal mass is specific to positional bias determines its interpretation. A signal present only for bias-manifesting questions would

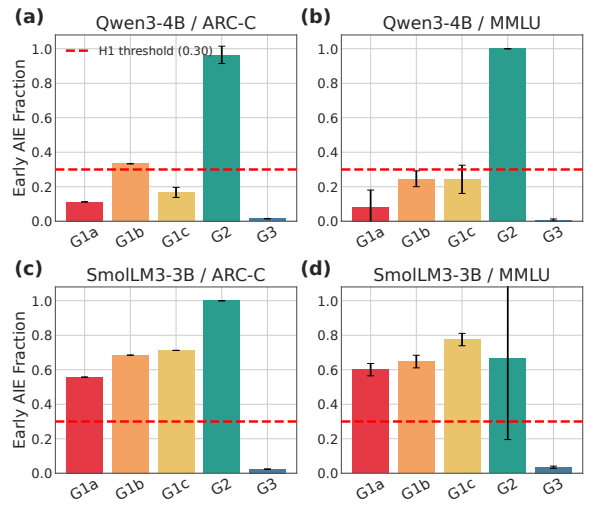


Figure 3: Early AIE fraction per question group for Qwen3-4B and SmoLM3-3B on ARC-Challenge and MMLU. G1a/G1b/G1c are bias-manifesting groups (1/4, 2/4, 3/4 rotations correct), G2 is robustly correct (4/4), and G3 is robustly wrong (0/4). The red dashed line marks the H1 threshold of 0.30. Error bars show standard deviation across seeds.

indicate a dedicated early bias circuit, whereas a signal shared with robustly correct questions would indicate general semantic commitment, of which positional bias is one manifestation.

Figure 3 shows EarlyFrac per group. Table 7 reports the quantitative values. Bonferroni-corrected permutation tests do not detect a significant difference at early layers between bias-manifesting questions (G1, mean EarlyFrac  $\approx 0.63$  to  $0.77$  for SmoLM3-3B) and robustly correct questions (G2, mean EarlyFrac  $\approx 0.67$  to  $1.00$ ), with  $p = 1.000$  at  $l/L \leq 0.40$ . Only robustly wrong questions (G3, mean EarlyFrac =  $0.007$  to  $0.035$ ) show near-zero early AIE, and both G1 and G2 differ strongly from

| Model      | Data  | G1 EF             | G2 EF             | G3 EF             |
|------------|-------|-------------------|-------------------|-------------------|
| SmolLM3-3B | MMLU  | 0.674 $\pm$ 0.019 | 0.667 $\pm$ 0.471 | 0.035 $\pm$ 0.006 |
| SmolLM3-3B | ARC-C | 0.651 $\pm$ 0.000 | 1.000 $\pm$ 0.000 | 0.024 $\pm$ 0.002 |
| Qwen3-4B   | MMLU  | 0.190 $\pm$ 0.075 | 1.000 $\pm$ 0.000 | 0.007 $\pm$ 0.007 |
| Qwen3-4B   | ARC-C | 0.204 $\pm$ 0.010 | 0.964 $\pm$ 0.050 | 0.015 $\pm$ 0.000 |

Table 7: EarlyFrac (mean  $\pm$  std across seeds) for bias-manifesting (G1), robustly correct (G2), and robustly wrong (G3) question groups.

G3 at mid and late layers; full  $p$ -value tables are reported in Appendix I.

The early-layer AIE signal does not track positional preference specifically; it tracks whether the model has a content-based answer to commit to. Questions the model cannot resolve at all (G3) show no early AIE because there is no content-based answer being committed to. Questions the model resolves correctly regardless of option order (G2) show high early AIE for the same reason as bias-manifesting questions, since both involve content commitment. Positional bias arises when this shared early commitment incorporates a label-associated prior (Section 5.4), not from a separate early-layer mechanism dedicated to positional preference.

The decomposition analysis identifies normalized depth 0.33 to 0.66 as the zone where attention mechanisms dominate and where positional integration occurs. Three consequences follow for intervention design. First, output-level shuffling and averaging reduce the visible effect of the bias on benchmark scores but do not change the model, since the positional preference is formed on every forward pass and the averaging only cancels its effect across runs. Output-level corrections are sufficient when the goal is reliable benchmark scores but insufficient when the goal is a model that does not encode positional preference internally (Wang et al., 2024c,b, 2025a). Second, because the mid-layer attention zone is shared with the semantic commitment pathway for correctly answered questions, suppressing this zone wholesale would trade correctness for unbiasedness. Third, these results motivate analysis at the level of individual attention heads within this zone (Wang et al., 2023; Conmy et al., 2023), targeting the attention patterns that encode option labels rather than option content, with sensitivity to option permutation versus option content as the discriminating evidence.

A per-head screening comparing AIE on bias-manifesting (G1) versus robustly correct (G2) questions for Qwen3-4B and Qwen3-8B takes the first

step in this direction, surfacing candidate heads near normalized depth 0.61 whose effects appear for G1 but not G2 and thereby narrowing the intervention target to concrete components (Appendix K). Beyond the main model range, the early-to-mid concentration extends to Qwen3-4B on all three datasets and to Qwen3-8B on HellaSwag, and base variants commit more shallowly than their post-trained counterparts (Appendix B).

## 7 Conclusion

Causal activation patching with a permutation-invariant AIE metric localizes the formation of positional bias in multiple-choice evaluation to the early-to-mid residual stream (normalized depth below 0.60) for the higher-capacity models tested; the conclusion holds across three output metrics ( $r > 0.89$ ), under multi-layer windowed patching, and at scale. Attribution within this range is MLP-dominated early ( $\approx 59\%$ ) and attention-dominated at mid-depth ( $\approx 76\%$ ) across the three tested families.

A factorization that permutes option labels and option order independently shows that the causal profile tracks the label-identity component of the bias rather than generic order sensitivity. Negative controls detect no early-layer separation between bias-manifesting and robustly correct questions, while robustly wrong questions carry no early signal, indicating that the early causal mass reflects the model’s commitment to a content-based answer, into which a label-associated prior is integrated, rather than a dedicated bias circuit. This shared pathway explains why output-level shuffling stabilizes benchmark scores without altering the model’s internal preference, and why late-layer patching suppresses rather than restores correct predictions.

Because the bias-carrying commitment shares its pathway with correct prediction, suppressing the mid-layer attention zone wholesale would trade accuracy for unbiasedness; the results instead motivate head-level analysis within the zone at normalized depth 0.33 to 0.66, where the per-head screening already identifies candidate heads. Establishing targeted interventions on these heads, and extending the scale and base-model checks reported here toward frontier models, are the natural next steps.

## Limitations

The main analysis covers instruction-tuned models in the 1.7B–4B parameter range across three architecture families, with preliminary checks on Qwen3-8B and on base variants of Qwen3-4B and Qwen3-8B; the 8B evidence is clear on HellaSwag, weak on MMLU, and inconclusive on ARC-Challenge, where only ten bias-manifesting questions were available, and frontier-scale models remain unexamined. The decomposition is reported at the sublayer level, and the head-level screening identifies candidate heads without demonstrating an intervention. The 1.7B versus 4B comparison documents a scale-dependent shift in commitment depth; a capacity reading of this shift is consistent with the data but is confounded with training-data, tokenizer, and architectural differences between the compared models. Because the residual stream is cumulative, patching effects are total effects through downstream computation, and finer attribution to circuits within layers requires additional analysis. The label/order factorization indicates that the cyclic-rotation profile is primarily aligned with label identity, but it does not exhaust all interactions between option semantics, labels, and prompt formatting.

## Acknowledgments

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00337250), by the Startup Growth Technology Development Program (Deep-Tech TIPS) funded by the Ministry of SMEs and Startups (MSS, Korea) (Project No. RS-2025-25461644), and by Institute of Information & communications Technology Planning & Evaluation (IITP, AI Computing Support Project for R&D) grant funded by the Korea government (MSIT) (RS-2026-25505492, High-Performance Research AI Computing Infrastructure Support at the 2 PFLOPS Scale).

## References

Hyeong Kyu Choi, Weijie Xu, Chi Xue, Stephanie Eckman, and Chandan K. Reddy. 2025. [Mitigating selection bias with node pruning and auxiliary options](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5190–5215, Vienna, Austria. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot,

Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try ARC, the AI2 reasoning challenge](#). *arXiv preprint arXiv:1803.05457*.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. [Towards automated circuit discovery for mechanistic interpretability](#). In *Advances in Neural Information Processing Systems*, volume 36.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. [A mathematical framework for transformer circuits](#). Transformer Circuits Thread.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. [Causal abstractions of neural networks](#). In *Advances in Neural Information Processing Systems*, volume 34.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. [Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. [Transformer feed-forward layers are key-value memories](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. [Found in the middle: Calibrating positional attention bias improves long context utilization](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995, Bangkok, Thailand. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranajpe, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.

- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.
- Pouya Pezeshkpour and Estevam Hruschka. 2024. [Large language models sensitivity to the order of options in multiple-choice questions](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.
- Joshua Robinson and David Wingate. 2023. [Leveraging large language models for multiple choice question answering](#). In *The Eleventh International Conference on Learning Representations*.
- Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. [Judging the judges: A systematic study of position bias in LLM-as-a-judge](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. [RoFormer: Enhanced transformer with rotary position embedding](#). *Neurocomputing*, 568:127063.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. [Investigating gender bias in language models using causal mediation analysis](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401.
- Elena Voita, David Talbot, Fedor Moiseev, Rico Senrich, and Ivan Titov. 2019. [Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.
- Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2025a. [LLMs may perform MCQA by selecting the least incorrect option](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5852–5862, Abu Dhabi, UAE. Association for Computational Linguistics.
- Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. [Interpretability in the wild: A circuit for indirect object identification in GPT-2 small](#). In *International Conference on Learning Representations*.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghui Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. 2024b. [Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think](#). In *First Conference on Language Modeling*.
- Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024c. [“my answer is C”: First-token probabilities do not match text answers in instruction-tuned language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.
- Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. 2025b. [Eliminating position bias of language models: A mechanistic approach](#). In *The Thirteenth International Conference on Learning Representations*.
- Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. [Unveiling selection biases: Exploring order and token sensitivity in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5598–5621, Bangkok, Thailand. Association for Computational Linguistics.
- Sarah Wiegrefe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. 2025. [Answer, assemble, ace: Understanding how LMs answer multiple choice questions](#). In *The Thirteenth International Conference on Learning Representations*.
- Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. 2024. [Strengthened symbol binding makes large language models reliable multiple-choice selectors](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4344, Bangkok, Thailand. Association for Computational Linguistics.
- Zhen Yang, Ping Jian, and Chengzhi Li. 2025. [Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1902–1917, Albuquerque, New Mexico. Association for Computational Linguistics.

- Soyoung Yoon, Dongha Ahn, Youngwon Lee, Minkyu Jung, HyungJoo Jang, and Seung-won Hwang. 2025. [RoToR: Towards more reliable responses for order-invariant inputs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18739–18760, Vienna, Austria. Association for Computational Linguistics.
- Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. 2025. [Mitigate position bias in LLMs via scaling a single hidden states channel](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6092–6111, Vienna, Austria. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- Fred Zhang and Neel Nanda. 2024. [Towards best practices of activation patching in language models: Metrics and methods](#). In *The Twelfth International Conference on Learning Representations*.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.
- Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. [Large language models are not robust multiple choice selectors](#). In *The Twelfth International Conference on Learning Representations*.

## A Extended Related Work

### A.1 Positional Bias in Multiple-Choice Evaluation

The phenomenon studied here has been documented under several closely related names. [Zheng et al. \(2024\)](#) showed that large language models exhibit a systematic preference for particular option positions on multiple-choice benchmarks, and [Pezeshkpour and Hruschka \(2024\)](#) reported substantial accuracy variation under permutation of option ordering on MMLU and other datasets. [Wei et al. \(2024\)](#) examined the phenomenon as a combination of order-related and token-related sensitivities.

A separate axis of this literature concerns how the answer probability is read off from the model. [Robinson and Wingate \(2023\)](#) compared first-token symbol-probability scoring against cloze-style probability of the option text and reported that the two procedures can yield different rankings. [Wang et al. \(2024c\)](#) found that, for instruction-tuned models, the first-token probability over option labels can disagree with the model’s free-form text answer; [Wang et al. \(2024b\)](#) reported that text-answer-based evaluation can give a different and in some cases more stable picture of multiple-choice performance than first-token scoring. [Wang et al. \(2025a\)](#) argued that language models may approach multiple-choice question answering in part by eliminating clearly wrong options rather than by directly identifying the correct one. Our use of three output metrics, restricted softmax (M1), full-vocabulary softmax (M2), and raw logit (M3), addresses the first-token-scoring axis of this literature, but our analysis is concerned with where in the network positional preference is formed rather than with which output-side probability to use.

### A.2 Mitigation Approaches

Most existing mitigation work operates at the output. [Zhao et al. \(2021\)](#) proposed contextual calibration to remove surface-form biases (majority-label, recency, and common-token effects) from few-shot in-context predictions. Within the multiple-choice setting specifically, [Xue et al. \(2024\)](#) introduced strengthened symbol binding for more reliable selection, [Choi et al. \(2025\)](#) combined node pruning with auxiliary options, [Yang et al. \(2025\)](#) investigated and mitigated option-symbol bias, and [Yoon et al. \(2025\)](#) addressed order sensitivity through order-invariant input encoding. These methods re-

duce the visible effect of positional bias on benchmark scores but do not directly characterize where in the model the bias is encoded.

A smaller body of work intervenes inside the model. [Yu et al. \(2025\)](#) identified a single hidden-state channel associated with positional preference and proposed scaling that channel to mitigate the bias. [Wang et al. \(2025b\)](#) proposed a mechanistic intervention to reduce position bias in language models. Our work is complementary to this internal-intervention line: rather than proposing an intervention, we provide a layer-resolved causal account of where the bias-relevant signal accumulates and how attention and MLP sublayers contribute, which we view as a prerequisite for principled head- or channel-level mitigation.

### A.3 Mechanistic Interpretability and Causal Patching

Our methodology builds on the causal mediation and activation-patching tradition. [Vig et al. \(2020\)](#) introduced causal mediation analysis to study bias in language models. [Meng et al. \(2022\)](#) used causal tracing to localize factual associations and identified mid-layer MLPs at the subject token’s last position as key sites for factual recall in GPT-style models. [Geiger et al. \(2021\)](#) formalized causal abstractions of neural networks. [Zhang and Nanda \(2024\)](#) and [Heimersheim and Nanda \(2024\)](#) provide best-practice guidance for activation patching, including the choice of metrics and the interpretation of negative effects, both of which inform our use of the AIE metric and our analysis of late-layer suppression.

The interpretation of attention versus MLP roles draws on prior work on transformer internals. [Geva et al. \(2021\)](#) characterized transformer feed-forward layers as key-value memories, and [Geva et al. \(2022\)](#) described feed-forward layers as building predictions by promoting concepts in the vocabulary space. [Elhage et al. \(2021\)](#) provided the mathematical framework for analyzing transformer circuits in terms of attention and MLP contributions. [Wiegrefe et al. \(2025\)](#) applied vocabulary projection and activation patching to formatted multiple-choice answering, localizing answer-symbol prediction to a few middle layers and their attention mechanisms; Section 2 discusses how our bias-conditioned design differs from that unconditioned analysis. The MLP-early, attention-mid staged organization we observe for positional commitment is consistent with this picture, with early

feed-forward layers carrying semantic-content signal and mid-layer attention integrating positional structure on top of it.

#### A.4 Other Forms of Positional Bias

Positional bias outside the multiple-choice setting has also been studied. In language-model-as-judge evaluation, Wang et al. (2024a) showed that models exhibit systematic position-dependent preferences when comparing two responses, and Shi et al. (2025) provided a systematic study of position bias in this setting. In long-context retrieval, Liu et al. (2024) documented the “lost in the middle” phenomenon, in which relevant information at the middle of a long context is utilized less than information at the ends, and Hsieh et al. (2024) addressed this by calibrating positional attention bias. These positional biases are mechanistically distinct from multiple-choice option-position bias, involving much longer effective sequence lengths and different task structures, but the present work suggests that, at least for option-position bias in short multiple-choice prompts, the locus lies in the early-to-mid residual stream rather than at the output softmax.

## B Model-Specific Behavior

### B.1 The Llama-3.2 Extreme Case

Llama-3.2-3B-Instruct produces  $\text{EarlyFrac} = 1.000 \pm 0.000$  across all three datasets and all three output metrics. This arises because the positive AIE profile is concentrated at a single normalized depth near  $0.48$  to  $0.50L$ , after which AIE becomes strongly negative in all subsequent layers. Since EarlyFrac counts only positive mass, and all positive mass falls before the  $0.6L$  threshold, the metric returns  $1.000$  by construction.

TotalPosAIE =  $0.118$  to  $0.163$  on MMLU and ARC-Challenge confirms the signal is non-trivial. The sharpness of the single positive peak at  $0.48L$  paired with the magnitude of negative suppression at deeper layers reflects a version of the staged commitment pattern in which the positional integration event is sharply localized and the subsequent late-layer divergence is severe. SmolLM3-3B ( $\text{EarlyFrac} \approx 0.68$  to  $0.78$ , more distributed positive mass) represents the other end of the same spectrum.

### B.2 The Qwen3 Scale Dependence

The Qwen3-1.7B model rejects H1 on MMLU ( $\text{EarlyFrac} = 0.143$ ,  $95\% \text{ CI} = [0.007, 0.468]$ ) and ARC-Challenge ( $\text{EarlyFrac} = 0.178$ ,  $95\% \text{ CI} = [0.018, 0.322]$ ) while supporting it on HellaSwag ( $\text{EarlyFrac} = 0.500$ ,  $95\% \text{ CI} = [0.333, 0.925]$ ).<sup>1</sup> The Qwen3-4B model, sharing the same architecture family with increased parameter count, produces  $\text{EarlyFrac} = 0.515$  on MMLU and  $0.615$  on HellaSwag, supporting H1 in both conditions. The increase from  $0.143$  (1.7B MMLU) to  $0.515$  (4B MMLU) corresponds to a difference of  $0.372$ .

The mean EarlyFrac across both datasets is  $0.369$  for Qwen3-1.7B and  $0.565$  for Qwen3-4B, a difference of  $+0.196$ . This pattern is compatible with a scale-related shift in commitment depth, with the larger model committing to semantic and label representations earlier in its residual stream. One reading is that the 1.7B model has insufficient capacity to form stable mid-layer commitments on the knowledge-intensive MMLU and ARC-Challenge tasks, while HellaSwag’s common-sense format admits earlier commitment at that scale; this reading is consistent with the data but is confounded with training-data, tokenizer, and architectural differences between the compared models, so the comparison does not isolate capacity as the cause.

The depth-decile profiles provide additional evidence. For Qwen3-1.7B averaged over MMLU and HellaSwag, decile D7 carries a positive mean AIE of  $+0.038$ , in contrast to the negative values at D7 to D10 seen in all other models. For Qwen3-4B, D7 carries  $-0.001$  on MMLU and the profile transitions to strongly negative at D8 ( $-0.092$ ), matching the suppression pattern. The 4B model shows an early peak followed by late suppression; the 1.7B model shows a more diffuse positive profile extending into D7 before suppression.

The attribution decomposition experiment was conducted on Qwen3-1.7B, Llama-3.2-3B-Instruct, and SmolLM3-3B. Despite Qwen3-1.7B rejecting H1 on MMLU and ARC-Challenge, its MLP-early, attention-mid attribution pattern holds, with early-zone Attn =  $0.399$  (MMLU) and mid-zone

<sup>1</sup>Values in this appendix come from the single-seed four-model sensitivity analysis ( $N = 200$  questions per dataset), which compares Qwen3-1.7B and Qwen3-4B under matched conditions. The Qwen3-1.7B/MMLU value of  $0.143$  therefore differs from the three-seed mean of  $0.194$  in Table 1; both reject H1 at the  $\tau = 0.30$  threshold and the bootstrap CIs of the two estimates overlap.

| Model    | Dataset | $N_{\text{bias}}$ | EF (M1) | 95% CI         | H1       |
|----------|---------|-------------------|---------|----------------|----------|
| Qwen3-4B | MMLU    | 50                | 0.701   | [0.353, 0.885] | Yes      |
| Qwen3-4B | ARC-C   | 27                | 0.607   | [0.292, 0.734] | Yes      |
| Qwen3-4B | HSwag   | 80                | 0.570   | [0.383, 0.697] | Yes      |
| Qwen3-8B | MMLU    | 44                | 0.342   | [0.073, 0.597] | Weak     |
| Qwen3-8B | ARC-C   | 10                | 0.161   | [0.069, 0.438] | Inconcl. |
| Qwen3-8B | HSwag   | 68                | 0.673   | [0.547, 0.759] | Yes      |

Table 8: Scale extension for instruction-tuned Qwen3-4B and Qwen3-8B (seed 42, 120 questions per dataset). The ARC-Challenge condition at 8B is inconclusive due to the small number of bias-manifesting questions.

Attn = 0.789 (MMLU). The staged architecture is present in Qwen3-1.7B; the scale limitation affects the timing of commitment (where positive causal mass accumulates) rather than the fundamental organization of sublayer processing.

Two aspects of the Qwen3-1.7B behavior warrant additional note. First, the HellaSwag condition for Qwen3-1.7B shows EarlyFrac = 0.500, supporting H1, while MMLU and ARC-Challenge do not. This asymmetry is consistent with the capacity reading above, since HellaSwag’s common-sense format demands less domain knowledge and is more amenable to early commitment at the 1.7B scale. Second, the rejection of H1 by Qwen3-1.7B on two of three datasets is a limitation of the universality claim.

### B.3 Larger Models and Base Variants

An extension covering Qwen3-4B, Qwen3-8B, and their base (non-instruction-tuned) counterparts probes behavior beyond the 1.7B–3B range (single seed, 120 questions per dataset, protocol otherwise identical). Table 8 reports the tuned-model results. Qwen3-4B supports early-to-mid causal concentration on all three datasets. Qwen3-8B supports it clearly on HellaSwag and weakly on MMLU; ARC-Challenge yields only ten bias-manifesting questions at this scale, too few for a stable estimate. Across all twelve model-dataset conditions in the extension, the layerwise M1–M2 Pearson correlation is at least 0.997, so the profiles are not normalization artifacts at these scales either.

Base variants show consistently shallower early fractions than their post-trained counterparts (Table 9): for Qwen3-4B, EarlyFrac drops from 0.701 to 0.334 on MMLU, from 0.607 to 0.150 on ARC-Challenge, and from 0.570 to 0.266 on HellaSwag. Post-training thus deepens early commitment, while the causal-localization framework itself applies to both regimes.

| Dataset | Post-trained EF | Base EF |
|---------|-----------------|---------|
| MMLU    | 0.701           | 0.334   |
| ARC-C   | 0.607           | 0.150   |
| HSwag   | 0.570           | 0.266   |

Table 9: Qwen3-4B versus its base variant. Base models show shallower early fractions across all three datasets.

## C Extended Metric Robustness Results

Table 10 reports the full pairwise profile correlation results for all three metric pairs across all nine model-dataset combinations. The M1-M2 correlation exceeds 0.998 in all cases, indicating that restricted and full-vocabulary softmax produce essentially identical AIE shapes. The M1-M3 Pearson correlation ranges from 0.899 to 0.996, with the lower bound occurring for ARC-Challenge on SmolLM3-3B, where the profile shape is similar but raw logit magnitudes differ more from normalized probabilities.

| Model      | Data  | $r(\text{M1}, \text{M2})$ | $\rho(\text{M1}, \text{M2})$ | $r(\text{M1}, \text{M3})$ | $\rho(\text{M1}, \text{M3})$ |
|------------|-------|---------------------------|------------------------------|---------------------------|------------------------------|
| Qwen3-1.7B | MMLU  | 1.000                     | 0.995                        | 0.939                     | 0.903                        |
| Qwen3-1.7B | ARC-C | 1.000                     | 0.996                        | 0.909                     | 0.704                        |
| Qwen3-1.7B | HSwag | 0.999                     | 0.992                        | 0.986                     | 0.841                        |
| SmolLM3-3B | MMLU  | 1.000                     | 0.992                        | 0.950                     | 0.939                        |
| SmolLM3-3B | ARC-C | 0.999                     | 0.989                        | 0.899                     | 0.945                        |
| SmolLM3-3B | HSwag | 0.998                     | 0.971                        | 0.950                     | 0.970                        |
| Llama-3.2  | MMLU  | 0.999                     | 0.996                        | 0.996                     | 0.950                        |
| Llama-3.2  | ARC-C | 1.000                     | 0.992                        | 0.988                     | 0.945                        |
| Llama-3.2  | HSwag | 1.000                     | 0.984                        | 0.991                     | 0.874                        |

Table 10: Full pairwise profile correlations (Pearson  $r$  and Spearman  $\rho$ ) between M1-M2 and M1-M3 AIE profiles for all nine model-dataset combinations, averaged over three seeds.

The Qwen3-1.7B M1-M3 correlation of 0.939 on MMLU and 0.909 on ARC-Challenge, despite the H1 rejection, indicates that the three metrics agree on the shape of the AIE profile even when the profile does not have its positive mass concentrated in early layers. The H1 rejection is a property of where the mass is located, not a disagreement among metrics about the profile shape.

## D Threshold Sensitivity

The choice of  $\tau = 0.30$  and  $c = 0.60$  for the H1 hypothesis could be questioned as post-hoc. We address this with a  $7 \times 7$  sensitivity grid over  $\tau \in \{0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55\}$  and  $c \in \{0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70\}$ .

For SmolLM3-3B on MMLU, H1 is supported in 23/49 grid cells. For Llama-3.2-3B-Instruct on MMLU, H1 is supported in 41/49 cells. The default setting ( $\tau = 0.30$ ,  $c = 0.60$ ) falls within the

support region for both SmolLM3-3B and Llama-3.2-3B-Instruct. Table 11 shows H1 support across  $\tau$  values at the fixed  $c = 0.60$  cutoff, indicating that H1 holds for SmolLM3-3B and Llama-3.2 at all tested  $\tau$  up to 0.55.

| Model      | Data  | $\tau$ |      |      |      |      |      |      |
|------------|-------|--------|------|------|------|------|------|------|
|            |       | 0.25   | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 | 0.55 |
| Qwen3-1.7B | MMLU  | N      | N    | N    | N    | N    | N    | N    |
| Qwen3-4B   | MMLU  | Y      | Y    | Y    | Y    | Y    | Y    | N    |
| SmolLM3-3B | MMLU  | Y      | Y    | Y    | Y    | Y    | Y    | Y    |
| Llama-3.2  | MMLU  | Y      | Y    | Y    | Y    | Y    | Y    | Y    |
| Qwen3-1.7B | HSwag | Y      | Y    | Y    | Y    | Y    | Y    | Y    |
| Qwen3-4B   | HSwag | Y      | Y    | Y    | Y    | Y    | Y    | Y    |
| SmolLM3-3B | HSwag | Y      | Y    | Y    | Y    | Y    | Y    | Y    |
| Llama-3.2  | HSwag | Y      | Y    | Y    | Y    | Y    | Y    | Y    |

Table 11: H1 support (Y/N) at fixed  $c = 0.60$  across seven  $\tau$  values on MMLU and HellaSwag.

## E Depth-Decile AIE Tables

Tables 12 and 13 report the mean AIE at each depth decile for Llama-3.2-3B-Instruct and SmolLM3-3B under the large-scale all-pairs protocol.

The positive peak at D5 for Llama (0.48L) and D6 for SmolLM3 (0.58L) both fall within the early zone. The consistently negative values from D6 to D10 (Llama) or D7 to D10 (SmolLM3) confirm the suppression phenomenon reported in Section 6.1. The transition from positive to negative AIE is sharp in both models, consistent with the interpretation that late-layer representations diverge rapidly across permutation contexts.

## F Zone Attribution by Dataset

Table 14 provides the full zone-level attribution breakdown for all model-dataset combinations from the attention-MLP decomposition experiment, conducted on Qwen3-1.7B, Llama-3.2-3B-Instruct, and SmolLM3-3B.

The cross-architecture consistency of the MLP-early, attention-mid pattern is visible across all rows. For Qwen3-1.7B on MMLU and ARC-C, mid-zone attention fraction (0.789 and 0.802) substantially exceeds early-zone attention fraction (0.399 and 0.312). For Llama-3.2, mid-zone attention reaches 0.793 to 0.855 while early-zone attention ranges from 0.329 to 0.477. For SmolLM3-3B, mid-zone attention is 0.746 to 0.798 while early-zone is 0.380 to 0.437.

The HellaSwag condition for Qwen3-1.7B shows a different pattern, with nearly balanced mid-zone attribution (0.499/0.501). This corresponds to the condition where EarlyFrac is border-

line (0.500) and TotalPosAIE is lower. This observation is consistent with the scale-dependent capacity interpretation, in which a 1.7B model may process HellaSwag questions under a different regime on this particular dataset where the transition from MLP- to attention-dominated attribution is less pronounced. Whether this reflects task-specific behavior or a general limitation of 1.7B-scale processing is not resolved by the current data and warrants further investigation.

## G Position Independence: Detailed Results

To verify that EarlyFrac does not depend on which absolute answer position the model prefers, bias-manifesting questions are stratified by preferred position (A, B, C, or D) and independence is tested using a Pearson chi-squared test of independence on a  $2 \times 4$  contingency table. Each bias-manifesting question is classified as H1-above (its individual  $\text{EarlyFrac} \geq 0.30$ ) or H1-below (its individual  $\text{EarlyFrac} < 0.30$ ); rows correspond to these two classes and columns correspond to the four position strata:

$$\chi^2 = \sum_{i \in \{\text{above, below}\}} \sum_{s \in \{A, B, C, D\}} \frac{(O_{i,s} - E_{i,s})^2}{E_{i,s}} \quad (16)$$

where  $O_{i,s}$  is the observed question count in cell  $(i, s)$  and  $E_{i,s} = N_i \cdot N_s / N$  is the expected count under independence between H1 status and preferred position. A non-significant result ( $p > 0.05$ ) provides no evidence against EarlyFrac being independent of which position the model prefers. Table 15 reports the complete results for all model-dataset combinations under both the all-pairs and first valid-pair methods.

SmolLM3-3B returns  $p = 0.865$  on MMLU and  $p = 0.799$  on ARC-Challenge under the all-pairs method, both well above the significance threshold. For Llama-3.2-3B-Instruct, the first valid-pair method returns  $p = 0.126$  on MMLU; the all-pairs method returns  $p = 0.046$ , which is marginally significant. The mean-profile EarlyFrac is 1.000 in every stratum (A, B, C, D) for Llama-3.2-3B-Instruct, meaning the position-averaged AIE profiles locate causal mass in identical early-layer regions. The marginally significant chi-squared result arises from imbalance in the H1-above question counts across positions ( $N_A = 455$ ,  $N_B = 98$ ,  $N_C = 327$ ,  $N_D = 262$ , summing the H1-above

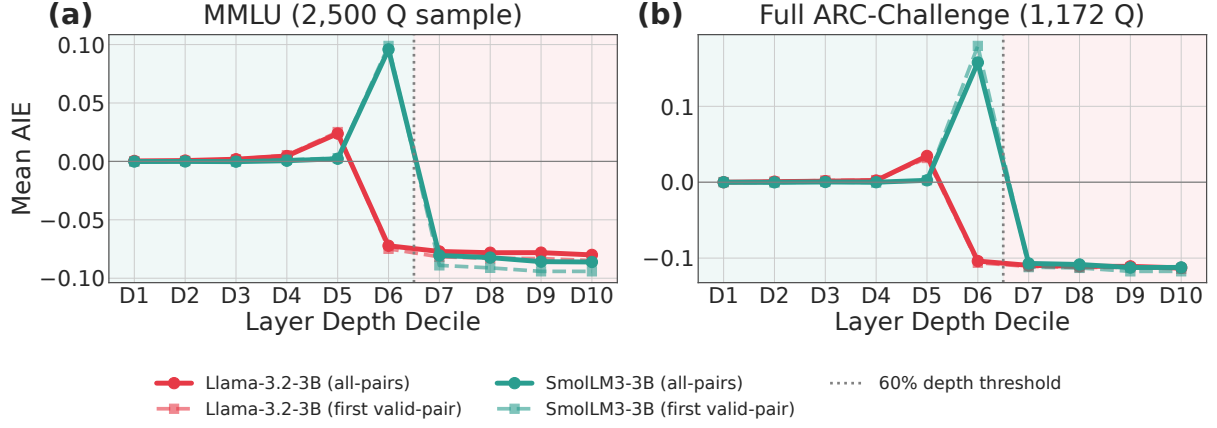


Figure 4: Depth-decile AIE profiles for Llama-3.2-3B-Instruct and SmolLM3-3B on MMLU and full ARC-Challenge. Solid lines show the all-pairs averaged method; dashed lines show the first valid-pair baseline. The vertical dotted line marks the 60% depth threshold.

| Data  | D1     | D2     | D3     | D4     | D5     | D6     | D7     | D8     | D9     | D10    |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| MMLU  | +0.000 | +0.001 | +0.002 | +0.005 | +0.024 | -0.072 | -0.077 | -0.078 | -0.078 | -0.080 |
| ARC-C | +0.000 | +0.001 | +0.001 | +0.002 | +0.035 | -0.104 | -0.110 | -0.111 | -0.111 | -0.113 |

Table 12: Mean AIE by depth decile for Llama-3.2-3B-Instruct using the all-pairs method on MMLU ( $N = 1,142$ ) and ARC-Challenge ( $N = 351$ ). Values truncated to three decimal places.

row of the contingency table) rather than any substantive difference in where causal mass lies.

For Llama-3.2-3B-Instruct on MMLU, the mean AIE profile within each position stratum places all positive causal mass before  $0.6L$ , yielding a stratum-mean EarlyFrac of 1.000 in every stratum. The contingency-table chi-squared test does not operate on these stratum means but on individual-question H1-above versus H1-below counts within each stratum; the marginal significance ( $p = 0.046$ ) reflects imbalance in those per-position counts rather than heterogeneity in where causal mass is located. The substantive conclusion is that causal commitment depth does not vary systematically with the model’s preferred answer position for Llama-3.2-3B-Instruct.

## H Late-Layer Probability Redistribution

When a clean late-layer activation is patched into a corrupted forward pass, the change in the model’s output probabilities falls into one of two patterns. A uniform diffusion would lower all option probabilities together, including both the correct option and the wrong options. A selective redistribution lowers the correct-option probability while raising the wrong-option probabilities. These two patterns have different mechanistic implications. Uniform diffusion is consistent with a generic loss of repre-

sentational structure, while selective redistribution indicates that the patched activation actively redirects probability mass toward incorrect answers.

Table 16 reports  $\Delta P(\text{correct})$  and the mean  $\Delta P(\text{wrong})$  for late-layer patching across all six tested model-dataset combinations. Every condition shows the selective pattern:  $\Delta P(\text{correct})$  is negative (ranging from  $-0.047$  to  $-0.107$ ) while  $\Delta P(\text{wrong})$  is positive (ranging from  $+0.016$  to  $+0.036$ ). This is consistent with the representational incompatibility account in Section 6.1: late-layer activations are specific enough to the option ordering they were produced under that, when transplanted into a different ordering, they direct probability toward the option that occupies the original correct slot in the new ordering, which is incorrect.

## I Negative Control Permutation Tests

Table 17 reports Bonferroni-corrected permutation test  $p$ -values comparing the bias-manifesting group (G1) to the robustly correct group (G2) and the robustly wrong group (G3) at two normalized depths. The tests were run with 10,000 permutations per comparison.

At early layers ( $l/L \leq 0.40$ ), all G1 versus G2 comparisons return  $p = 1.000$  after correction, providing no evidence for a difference between the

| Data  | D1     | D2     | D3     | D4     | D5     | D6     | D7     | D8     | D9     | D10    |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| MMLU  | −0.000 | −0.000 | −0.000 | +0.001 | +0.002 | +0.096 | −0.081 | −0.082 | −0.086 | −0.086 |
| ARC-C | −0.000 | −0.000 | +0.000 | +0.000 | +0.003 | +0.158 | −0.107 | −0.108 | −0.112 | −0.112 |

Table 13: Mean AIE by depth decile for SmoLLM3-3B using the all-pairs method on MMLU ( $N = 1,031$ ) and ARC-Challenge ( $N = 296$ ).

| Model                   | Data      | Zone  | AttnFrac          | MLPFrac           |
|-------------------------|-----------|-------|-------------------|-------------------|
| Qwen3-1.7B <sup>†</sup> | MMLU      | Early | $0.399 \pm 0.078$ | $0.601 \pm 0.078$ |
|                         |           | Mid   | $0.789 \pm 0.064$ | $0.211 \pm 0.064$ |
|                         |           | Late  | $0.430 \pm 0.061$ | $0.570 \pm 0.061$ |
|                         | HellaSwag | Early | $0.513 \pm 0.094$ | $0.487 \pm 0.094$ |
|                         |           | Mid   | $0.499 \pm 0.046$ | $0.501 \pm 0.046$ |
|                         |           | Late  | $0.461 \pm 0.047$ | $0.539 \pm 0.047$ |
|                         | ARC-C     | Early | $0.312 \pm 0.041$ | $0.688 \pm 0.041$ |
|                         |           | Mid   | $0.802 \pm 0.024$ | $0.198 \pm 0.024$ |
|                         |           | Late  | $0.369 \pm 0.012$ | $0.631 \pm 0.012$ |
| Llama-3.2               | MMLU      | Early | $0.477 \pm 0.065$ | $0.523 \pm 0.065$ |
|                         |           | Mid   | $0.793 \pm 0.046$ | $0.207 \pm 0.046$ |
|                         |           | Late  | $0.479 \pm 0.009$ | $0.521 \pm 0.009$ |
|                         | HellaSwag | Early | $0.329 \pm 0.023$ | $0.671 \pm 0.023$ |
|                         |           | Mid   | $0.795 \pm 0.043$ | $0.205 \pm 0.043$ |
|                         |           | Late  | $0.347 \pm 0.006$ | $0.653 \pm 0.006$ |
|                         | ARC-C     | Early | $0.455 \pm 0.034$ | $0.545 \pm 0.034$ |
|                         |           | Mid   | $0.855 \pm 0.010$ | $0.145 \pm 0.010$ |
|                         |           | Late  | $0.232 \pm 0.060$ | $0.768 \pm 0.060$ |
| SmoLLM3-3B              | MMLU      | Early | 0.380             | 0.620             |
|                         |           | Mid   | 0.746             | 0.254             |
|                         |           | Late  | 0.612             | 0.388             |
|                         | HellaSwag | Early | 0.437             | 0.563             |
|                         |           | Mid   | 0.771             | 0.229             |
|                         |           | Late  | 0.449             | 0.551             |
|                         | ARC-C     | Early | 0.413             | 0.587             |
|                         |           | Mid   | 0.798             | 0.202             |
|                         |           | Late  | 0.512             | 0.488             |

Table 14: Full zone-level attention and MLP attribution fractions for all model-dataset combinations in the attribution decomposition experiment. <sup>†</sup>Qwen3-1.7B rejects H1 on MMLU and ARC-Challenge. Early = 0 to 33% depth, Mid = 33 to 66%, Late = 66 to 100%.

| Model      | Data  | Method           | $\chi^2$ | $p$   | Independent? |
|------------|-------|------------------|----------|-------|--------------|
| Llama-3.2  | MMLU  | All-pairs        | 8.011    | 0.046 | Marginal     |
| Llama-3.2  | MMLU  | First valid-pair | 5.721    | 0.126 | Yes          |
| Llama-3.2  | ARC-C | All-pairs        | 5.495    | 0.139 | Yes          |
| Llama-3.2  | ARC-C | First valid-pair | 6.431    | 0.092 | Yes          |
| SmoLLM3-3B | MMLU  | All-pairs        | 0.734    | 0.865 | Yes          |
| SmoLLM3-3B | MMLU  | First valid-pair | 4.442    | 0.218 | Yes          |
| SmoLLM3-3B | ARC-C | All-pairs        | 1.009    | 0.799 | Yes          |
| SmoLLM3-3B | ARC-C | First valid-pair | 0.767    | 0.857 | Yes          |

Table 15: Full position independence chi-squared test results.

bias-manifesting and robustly correct AIE profiles in the early zone. At deeper layers ( $l/L = 0.5$  and  $0.8$ ), G1 and G2 begin to differ for SmoLLM3-3B on MMLU but show no detectable difference for Qwen3-4B at  $l/L = 0.5$ , suggesting that the divergence between bias-manifesting and robustly

| Model      | Data  | $\Delta P(\text{corr})$ | $\Delta P(\text{wrong})$ | Type      |
|------------|-------|-------------------------|--------------------------|-----------|
| Llama-3.2  | MMLU  | −0.082                  | +0.027                   | Selective |
| Llama-3.2  | HSwag | −0.095                  | +0.032                   | Selective |
| Qwen3-4B   | MMLU  | −0.107                  | +0.036                   | Selective |
| Qwen3-4B   | HSwag | −0.097                  | +0.032                   | Selective |
| SmoLLM3-3B | MMLU  | −0.095                  | +0.032                   | Selective |
| SmoLLM3-3B | HSwag | −0.047                  | +0.016                   | Selective |

Table 16: Softmax redistribution after late-layer patching.  $\Delta P(\text{corr})$  is the change in correct-option probability;  $\Delta P(\text{wrong})$  is the mean change in wrong-option probabilities.

correct processing emerges only near the output. The G1 versus G3 comparisons follow the same pattern: no detectable difference in the early zone, strongly different at  $l/L = 0.8$ . This is the basis

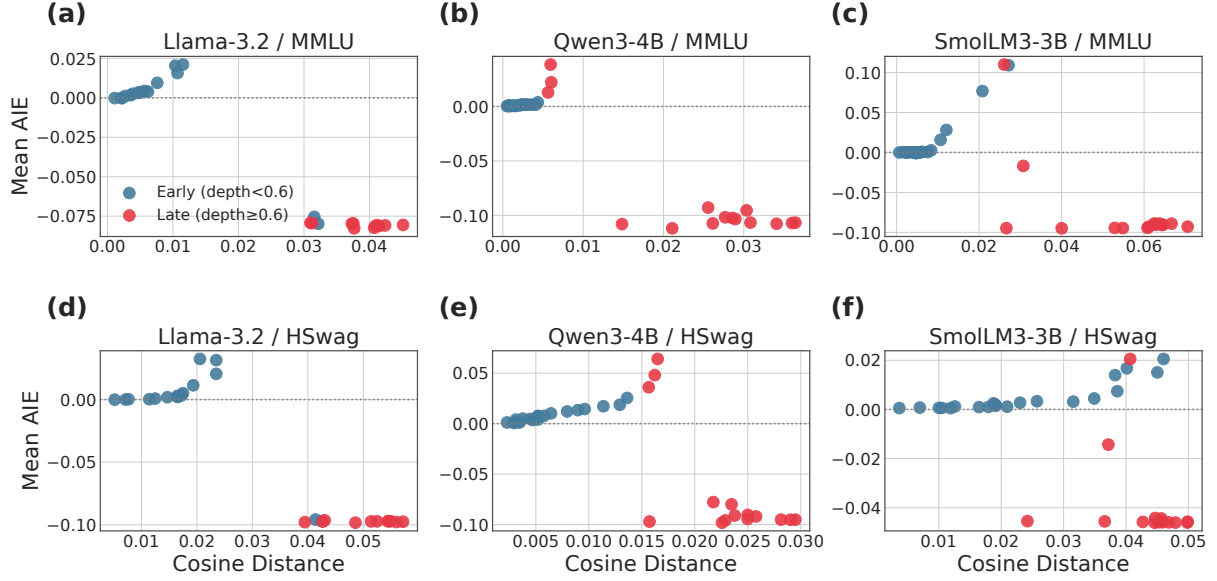


Figure 5: Scatter plots of mean AIE versus cosine distance between clean and corrupted residual streams, stratified by layer zone (early  $< 0.60$ : blue; late  $\geq 0.60$ : red) for Llama-3.2-3B-Instruct, Qwen3-4B, and SmolLM3-3B on MMLU and HellaSwag. Each point represents a single layer.

| Model      | Data | Comparison | $l/L = 0.5$ | $l/L = 0.8$ |
|------------|------|------------|-------------|-------------|
| SmolLM3-3B | MMLU | G1 vs G2   | 0.000**     | 0.000**     |
| SmolLM3-3B | MMLU | G1 vs G3   | 0.288       | 0.000**     |
| Qwen3-4B   | MMLU | G1 vs G2   | 1.000       | 0.000**     |
| Qwen3-4B   | MMLU | G1 vs G3   | 1.000       | 0.000**     |

Table 17: Bonferroni-corrected permutation test  $p$ -values comparing G1 (bias-manifesting) to G2 (robustly correct) and G3 (robustly wrong) at mid and late normalized depths. \*\*:  $p < 0.01$ .

for the claim in Section 6.2 that the early causal mass tracks general semantic commitment rather than positional preference specifically.

## J Sublayer Independence Analysis

Figure 6 shows the per-layer scatter of attention-AIE versus MLP-AIE for each model and dataset, alongside summary bar charts of zone-level attribution. If attention and MLP carried the positional bias signal as redundant copies, the per-layer values would be strongly correlated. If they carried independent signals, the correlation would be near zero.

Table 18 reports the Pearson correlation between the attention-AIE and MLP-AIE profiles for each model-dataset combination. The values range from  $-0.244$  to  $+0.179$ , all close to zero. The two sublayer types therefore do not act as linearly redundant copies of the same signal; their per-layer contributions vary largely independently across depth, consistent with the staged interpretation in Sec-

tion 5.5: MLP layers contribute primarily during the early content-representation phase, and attention layers contribute primarily during the mid-layer positional integration phase, with little overlap in the per-layer pattern.

| Model      | Dataset   | $r(\text{Attn, MLP})$ | Relationship          |
|------------|-----------|-----------------------|-----------------------|
| Qwen3-1.7B | MMLU      | $-0.020$              | Near-zero correlation |
| Qwen3-1.7B | HellaSwag | $-0.244$              | Near-zero correlation |
| Llama-3.2  | MMLU      | $0.179$               | Near-zero correlation |
| Llama-3.2  | HellaSwag | $0.068$               | Near-zero correlation |
| SmolLM3-3B | MMLU      | $-0.134$              | Near-zero correlation |
| SmolLM3-3B | HellaSwag | $0.070$               | Near-zero correlation |

Table 18: Pearson correlation between per-layer attention-AIE and MLP-AIE profiles for each model-dataset combination.

## K Head-Level Screening

To narrow the intervention target identified in Section 6.2, we compare per-head AIE on bias-manifesting (G1) versus robustly correct (G2) questions. For each attention head, the head’s output at the final token position is patched individually and the resulting AIE is computed separately for the two groups. Heads whose causal effect is present for G1 but absent or reversed for G2 are candidates for carrying the label-associated component rather than option content. Table 19 lists representative candidates; all lie near normalized depth 0.61, inside the mid-layer attention zone identified by the sublayer decomposition. The screening is a

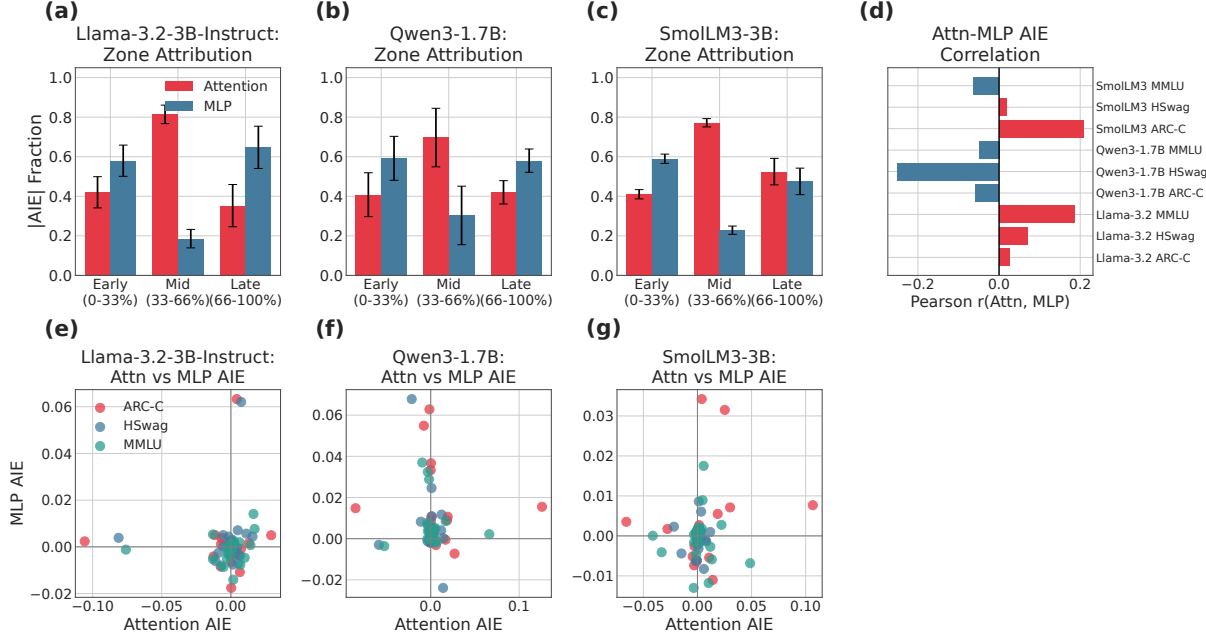


Figure 6: Mechanistic attribution analysis. Top row: bar charts of mean ( $\pm$ std) attention versus MLP |AIE| fraction per depth zone for each model. Top right: Pearson  $r$  between attention-AIE and MLP-AIE profiles. Bottom row: scatter plots of per-layer attention-AIE versus MLP-AIE for each model and dataset.

| Model    | Dataset | Layer (depth) | Head | G1 AIE | G2 AIE |
|----------|---------|---------------|------|--------|--------|
| Qwen3-8B | ARC-C   | 22 (0.61)     | 2    | 0.048  | 0.000  |
| Qwen3-4B | HSwag   | 22 (0.61)     | 10   | 0.007  | -0.018 |
| Qwen3-4B | MMLU    | 22 (0.61)     | 0    | 0.016  | -0.001 |

Table 19: Representative candidate heads from the per-head screening. G1 and G2 columns report per-head AIE on bias-manifesting and robustly correct questions respectively.

diagnostic that narrows the search space for future intervention; it does not establish that modifying these heads reduces positional bias without harming accuracy.

## L Implementation Details for Reproducibility

**Model loading.** All models are loaded from HuggingFace Hub in float16 precision with `device_map="auto"`. The attention implementation uses the native model attention (not Flash Attention 2) to ensure hook compatibility.

**Prompt format.** Each question is formatted as a single string with the question stem followed by options labeled A through D, using a standard prompt template. The final token position is the last token of the last option label and represents the model’s decision state.

**Hook registration.** For full residual stream patching, we register a `register_forward_hook` on the output of each transformer block’s final layer normalization. For attention-only patching, we hook the output of the `self_attn` module before the residual addition. For MLP-only patching, we hook the output of the `mlp` module before the residual addition. Clean activations are stored from a complete clean forward pass, then injected during the corrupted forward pass by setting the hooked activation to the stored clean value.

**Cyclic permutations.** Options are rotated cyclically: rotation  $k$  maps option  $i$  to position  $(i + k) \bmod 4$ . The correct answer index is adjusted accordingly. Token IDs for option labels (A, B, C, D) are extracted from the tokenizer’s vocabulary for each model.

**Hyperparameters.** Batch size per forward pass is 1 (sequential question processing). Bootstrap re-sampling uses  $B = 2,000$  samples for large-scale experiments and  $B = 1,000$  for smaller-scale ablations. All threshold values ( $\tau = 0.30$ ,  $c = 0.60$ ) were specified in the hypothesis before data collection.

**Hardware.** All experiments were conducted on a single server equipped with six GPUs.

**Random seeds.** Multi-seed experiments use  $\{42, 123, 456\}$ . Each seed controls (i) the random subsampling of questions from each dataset (so that the per-seed question subset differs while the dataset itself is fixed), (ii) Python and NumPy/PyTorch random number generators for any sampling-dependent operations, and (iii) the bootstrap resampling random number generator for confidence intervals. Model forward passes are deterministic given the question set, so across-seed variation reflects sampling variability of which questions enter the analysis rather than model stochasticity. Reported mean  $\pm$  std values aggregate over the three seeds.

**Negative control pair selection.** For G2 (robustly correct) and G3 (robustly wrong) question groups, the clean/corrupted pair is fixed as rotation 0 (clean) versus rotation 2 (corrupted). This fixed-pair design ensures that the negative control measurement is not influenced by pair selection, and that any observed early AIE in G2 reflects genuine model processing rather than an artifact of favorable pair choice.

## M Implications for Bias Mitigation

This appendix summarizes the implications of the causal-localization results for the design of bias-mitigation strategies.

**Location of the bias.** A common assumption treats positional bias as an output-level phenomenon, in which the model computes a probability distribution over the option tokens and a positional prior tilts that distribution at the final layer. Under this account, subtracting the prior at the output would be a principled correction. The findings here are inconsistent with it. The positional preference is formed several layers before the output and is integrated into the representation from which the output is read, so the output distribution reflects a commitment that has already been made.

**Scope of output-level shuffling.** Averaging predictions over shuffled option orders is a standard correction, and it is effective for estimating benchmark accuracy because position-specific errors cancel in aggregate. The intervention operates on the evaluation protocol rather than on the model. The positional preference is still formed on every individual forward pass, and the averaging cancels its

effect only at the level of the aggregate score; applications that depend on the reliability of a single forward pass therefore receive no benefit from it.

**Formation of the commitment.** The commitment forms in early-to-mid layers in two distinguishable phases. In roughly the first third of the network, MLP sublayers dominate the causal attribution and encode the semantic content of the question and options into the residual stream. In roughly the second third, attention sublayers become dominant and integrate option-label information into that content. The positional commitment is in place by the end of the mid-layers; late layers further specialize the representation to the specific permutation context without altering the commitment, which is why patching late layers across permutations interferes with the prediction rather than restoring it (Section 6.1).

**Interpretation of late-layer patching.** Late-layer patching produces the opposite of what a late-stage account would predict. If the bias arose near the output, replacing late-layer activations from an incorrect-ordering pass with activations from a correct-ordering pass should restore the correct prediction; instead, the correct-option probability decreases further. Late-layer representations are specialized to the option ordering under which they were produced, so an activation from a different ordering interferes rather than transfers. The decisive computation therefore occurs earlier, while the representation is still general enough to transfer across orderings.

**Shared commitment pathway.** The early-layer signal does not constitute a dedicated positional-bias circuit. Questions the model answers correctly under every ordering, for which no positional sensitivity is observed, exhibit the same early-layer profile as questions on which the bias determines the outcome. The signal reflects the model’s commitment to a content-based answer, and positional bias arises when that commitment incorporates label-associated priors acquired during pretraining. Because the bias and the model’s competence share this pathway, the pathway cannot be suppressed wholesale without degrading correct prediction.

**Zone-level suppression.** A plausible intervention would dampen mid-layer attention as a class, since the decomposition locates positional integration there. Such dampening, however, would reduce the bias only by broadly impairing informa-

tion integration. Robustly correct answering and bias-manifesting answering activate the same depth range with similar attention reliance, and the zone-level statistics do not distinguish which questions a given attenuation would affect most adversely.

**Head-level analysis.** The appropriate resolution is therefore the individual attention head. Within the mid-layer attention zone, some heads may carry option-label information as their primary signal while others carry semantic content. The per-head screening in Appendix K surfaces candidate heads near normalized depth 0.61 whose causal effects appear for bias-manifesting questions but not for robustly correct ones, and the sublayer decomposition supplies both the depth range in which attention dominates (normalized depth 0.33 to 0.66) and the discriminating evidence, sensitivity to option permutation versus sensitivity to option content. Whether targeted intervention on such heads reduces positional bias without harming correct prediction remains an open empirical question.