POSITION STILL MATTERS: LOCALIZING AND CONTROLLING LATENT POSITIONAL EFFECTS IN LLM COMPARISONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly used as evaluators, yet their judgments can depend on presentation order rather than only on relevant evidence. Evaluating both presentation orders can reveal inconsistent choices, but it does not show whether order still influences the decision when the final answer remains unchanged, where that influence enters the computation, or whether it can be reduced without weakening the comparison itself. We study these questions through a three-stage mechanistic framework. First, we decompose the model's decision margin into separate effects of comparison evidence and presentation position, revealing substantial positional shifts even in models whose choices are nearly invariant to swapping. Second, activation patching localizes this influence to late prompt states, while head-level interventions characterize how individual attention heads affect positional dependence and evidence sensitivity. Third, we construct interventions that reduce the positional effect under explicit constraints on comparison performance. On held-out numerical comparisons, the procedure removes 60–96% of the positional effect in five of six models, improving held-out balanced accuracy by up to 20 percentage points; in the remaining model, the same constraints permit only limited control. Reapplying the same procedure to country-population comparison and letter counting reduces the positional effect in all 12 model–task settings, by 21–100%, whereas the intervention selected on numerical comparison transfers inconsistently. These results show that behavioral order invariance can mask substantial latent positional influence, and that this influence can be causally localized and selectively controlled inside the model.

1 INTRODUCTION

Large language models are increasingly used to evaluate generated content, compare candidate answers, and support ranking and selection decisions (Zheng et al., 2023). These applications require judgments that reflect the relevant qualities of the alternatives rather than incidental features of their presentation. Position is one such influence: swapping two responses can change an LLM judge's preference (Wang et al., 2023), and rearranging answer options can alter its prediction on a multiple-choice question (Pezeshkpour & Hruschka, 2024). Understanding how presentation order influences model judgments is therefore important for their use as evaluators.

A visible sign of positional dependence is a change in the selected candidate after the alternatives are reordered. But the final choice reveals only the net result of multiple influences on the decision. Relevant evidence may favor one candidate while presentation order shifts the model toward the other. When the evidence is strong, even a substantial positional effect may leave the selected candidate unchanged; when the evidence is weak, a smaller effect may reverse the choice. Accuracy and agreement across presentation orders therefore do not reveal how strongly position contributes to the decision. Measuring this contribution requires separating the effect of presentation order from the model's response to relevant comparison evidence.

This distinction matters not only for measuring positional dependence, but also for understanding and controlling it. An order effect observed at the output does not reveal where it enters the computation, which components carry it, or whether those components also support the comparison itself.

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

Activation patching and related causal interventions provide tools for tracing which internal states and components influence an output (Geiger et al., 2022; Zhang & Nanda, 2024; Goldowsky-Dill et al., 2023). We use these tools to ask whether positional influence can be localized and selectively reduced without weakening the model's use of comparison evidence.

We study this problem in binary comparison, a basic operation underlying evaluation, ranking, and selection among larger sets of alternatives. We begin with numerically scored items, where the correct choice is unambiguous and the strength of the relevant evidence can be varied directly. We decompose the model's decision margin into separate effects of comparison evidence and presentation position, causally localize the positional influence in the model's computation, and characterize attention heads by how interventions on them affect both quantities. We then construct head-level interventions that reduce positional dependence under explicit constraints on evidence sensitivity and decision margin.

Across six models, presentation order significantly shifts the decision margin, including by several logits in models whose choices are essentially invariant to swapping: Gemma-2-9B and Qwen2.5-7B achieve 99.8–100% accuracy in both orders while retaining positional shifts of -2.57 and -2.73 logits. Strong comparison evidence can therefore mask positional influence without eliminating it. Activation patching localizes this influence to late prompt and decision states, with the same late-position structure appearing across all six models. On held-out numerical comparisons, the selected interventions remove 60–96% of the positional effect in five of six models while largely preserving or increasing evidence sensitivity. The resulting behavioral gains reach 20.3 percentage points in balanced accuracy; in the remaining model, the same performance constraints permit only limited control.

We then test whether this control extends beyond comparisons of explicit numerical scores. Country-population comparison requires retrieving evidence from model knowledge, whereas letter counting requires computing it from the prompt. A head set and gain selected on numerical comparison transfer inconsistently and can even increase the positional effect when its direction changes across tasks. In contrast, rerunning the same control procedure on each target task reduces the positional-effect magnitude in all 12 model–task settings, by 21–100%, while generally preserving or increasing evidence sensitivity. The task-specific procedure also yields larger positional-effect reduction and disfavored-order accuracy-gain point estimates than the frozen numerical recipe in all 12 settings. These results separate the generality of the control procedure from the transferability of a particular intervention: positional influence can be measured relative to useful evidence, causally localized within the model, and selectively reduced, but the components required for control depend on the comparison task.

2 Related work

Positional effects in evaluation. Presentation order can affect LLM judgments in pairwise evaluation (Zheng et al., 2023; Wang et al., 2023) and multiple-choice question answering (Pezeshkpour & Hruschka, 2024). Systematic studies further show that behavioral order consistency depends on the model, task, and quality difference between candidates (Shi et al., 2025). These findings motivate examining positional effects as a function of comparison evidence. We study a controlled setting in which numerical score gaps can be varied directly, and measure whether presentation order continues to shift the decision margin after generated answers become accurate and invariant to swapping.

Prior estimation and decision readouts. Probability-based approaches already provide ways to measure and correct preferences beyond the final answer. Liusie et al. (2024) calibrate pairwise preference probabilities to mitigate positional bias. PriDe (Zheng et al., 2024) separates the prediction distribution into an option-ID prior and a content-dependent component, using permutations to estimate a prior that can capture both token and position preferences. Our margin analysis builds on this distinction by estimating evidence sensitivity and positional shifts along a controlled numerical comparison axis. Readout choice also matters: first-token probabilities can disagree with generated answers in instruction-tuned models (Wang et al., 2024a). We therefore measure logits at the answer-token decision position and separately validate intervention effects under free generation.

Comparison circuits and causal analysis. Mechanistic studies have identified components supporting answer selection and numerical comparison. Lieberum et al. (2023) characterize attention heads involved in mapping answer content to multiple-choice labels, while Hanna et al. (2023) identify a circuit implementing greater-than behavior in GPT-2. Our localization uses residual-stream and attention-head activation patching (Zhang & Nanda, 2024). We then characterize each attention head through its intervention-induced changes in both the positional effect and evidence sensitivity. Measuring attenuation and amplification separately, and evaluating selected heads jointly, allows us to test which interventions reduce position dependence while preserving the comparison signal.

Internal interventions for positional control. Several studies directly modify internal computations to mitigate positional bias. Anchored Answers (Li & Gao, 2025) identifies and intervenes on MLP components and attention heads contributing to first-option bias in GPT-2. PINE (Wang et al., 2024b) changes inter-document attention and positional assignments to obtain order-invariant inference, including in pairwise judging. Yu et al. (2024) scale a positional hidden-state dimension to attenuate, amplify, or reverse positional preference in long-context tasks. Building on these precedents, we examine the selectivity and limits of internal control in pairwise comparison, measuring positional reduction alongside evidence sensitivity, decision margins, and accuracy. Our transfer experiments distinguish applying a selected head set and gain to another task from rerunning the selection procedure on that task, testing how the effectiveness and cost of control depend on the evidence required for the comparison.

3 METHOD

3.1 MEASURE: PAIRED COMPARISON AND MARGIN DECOMPOSITION

We study the minimal ranking problem of comparing two scored items. Each trial presents two opaque, single-token identifiers, each paired with a numerical score, and asks the model to return the identifier associated with the higher score. We denote the higher- and lower-scoring items by x_{hi} and x_{lo}, with scores $s_{\mathrm{hi}} > s_{\mathrm{lo}}$. Each pair is presented in both orders: once with x_{hi} in the first slot and once with it in the second. The identifiers and scores remain unchanged between the paired trials, isolating presentation order as the experimental difference. At the behavioral level, we measure accuracy separately for the two presentation orders.

Behavioral accuracy only reveals positional effects large enough to change the generated answer. To measure smaller shifts in the model's preference, we examine its logits at the decision position, immediately before it generates the selected identifier. Let $\tau(x)$ denote the token corresponding to item x, and let $\ell(v)$ denote the pre-softmax logit for token v at this position. We define the signed decision margin as

$$m = \ell\left(\tau(x_{\mathrm{hi}})\right) - \ell\left(\tau(x_{\mathrm{lo}})\right). \tag{1}$$

A positive margin favors the higher-scoring item, while its magnitude measures the strength of that preference.

We separate the effects of score evidence and presentation position by fitting

$$m = \beta_0 + \beta_{\mathrm{e}}\left(s_{\mathrm{hi}} - s_{\mathrm{lo}}\right) + \beta_{\mathrm{p}} p + \varepsilon, \qquad p = \mathbf{1}[x_{\mathrm{hi}} \text{ is in slot 2}]. \tag{2}$$

The evidence coefficient β_{e} measures the change in margin per unit increase in score gap. The positional effect β_{p} measures the change in margin when x_{hi} moves from the first slot to the second. Thus, $\beta_{\mathrm{p}} > 0$ indicates a preference for the second slot, whereas $\beta_{\mathrm{p}} < 0$ indicates a preference for the first. This margin-level measure can detect a positional shift even when that shift is not large enough to reverse the generated answer.

3.2 LOCALIZE: TRACING AND ATTRIBUTING POSITIONAL INFLUENCE

We localize the positional effect in three steps that progressively narrow the search. Activation patching identifies residual-stream states through which presentation order affects the decision margin. Head-level activation patching tests which attention heads contribute to those states. Per-head interventions then measure how each head affects both the order coefficient β_{p} and the evidence coefficient β_{e}.

3.2.1 ACTIVATION PATCHING

We use paired-swap activation patching (Zhang & Nanda, 2024; Geiger et al., 2022) to identify internal states through which presentation order affects the decision margin. For each paired comparison, the run in which x_{hi} appears second serves as the source, and the run in which it appears first serves as the target. At each layer L and token position P, we replace the target run's residual-stream state with the corresponding state from the source run and recompute the decision margin.

We quantify how much of the average full swap effect is reproduced by each patch:

$$\text{transmission}(L, P) = \frac{\mathbb{E}_i\left[m^{(i)}_{\text{patched}}(L, P) - m^{(i)}_{\text{first}}\right]}{\mathbb{E}_i\left[m^{(i)}_{\text{second}} - m^{(i)}_{\text{first}}\right]}, \tag{3}$$

where i indexes paired comparisons, and $m^{(i)}_{\text{first}}$ and $m^{(i)}_{\text{second}}$ are the decision margins when x_{hi} appears first and second, respectively. A transmission of 0 means that the patch produces no average change in the target margin, whereas a transmission of 1 means that it reproduces the average full swap effect. Negative values indicate movement opposite to the swap effect, and values greater than 1 indicate overshooting; transmission is therefore not restricted to $[0, 1]$.

3.2.2 HEAD-LEVEL ACTIVATION PATCHING

After identifying residual-stream states that transmit the swap effect, we patch individual attention-head outputs at the corresponding token positions. For each head, we replace its output in the target run with the output from the source run and recompute the decision margin. The resulting margin change measures how strongly that head can transmit the effect of presentation order at the patched position.

3.2.3 PER-HEAD INTERVENTION EFFECTS

Patching identifies components that transmit the swap effect, but it cannot tell whether a component carries positional information specifically or also contributes useful comparison evidence. We therefore measure how every attention head affects the positional effect β_{p}, the evidence coefficient β_{e}, and the mean decision margin.

Let o_h denote head h's output contribution to the residual stream. We apply the mean-centered intervention

$$\widetilde{o}_h = \overline{o}_h + \alpha\,(o_h - \overline{o}_h)\,, \tag{4}$$

where $\bar{o}_h$ is the baseline output of the head averaged across the prompts to which the intervention is applied, separately at each token position. During selection these means are estimated on development prompts; for held-out evaluation they are recomputed on the evaluation prompts (Appendix A.4).

We intervene on each head individually at $\alpha = 0$ and $\alpha = 2$ and refit Eq. 2 after every intervention. Relative to the unmodified model, we define

$$\Delta\beta_{\text{p}}(h, \alpha) = \widehat{\beta}_{\text{p}}(h, \alpha) - \widehat{\beta}^{0}_{\text{p}}, \tag{5}$$

$$\Delta\beta_{\text{e}}(h, \alpha) = \widehat{\beta}_{\text{e}}(h, \alpha) - \widehat{\beta}^{0}_{\text{e}}, \tag{6}$$

where the superscript 0 denotes the baseline model. We likewise record the change in mean decision margin. Together, these measurements give each head a functional profile describing how its attenuation or amplification affects positional dependence and comparison evidence.

Head selection uses the census measurement corresponding to the direction in which the head will be manipulated: attenuation candidates are evaluated using their effects at $\alpha = 0$, whereas amplification candidates are evaluated using their effects at $\alpha = 2$. We do not assume that attenuation and amplification have symmetric effects.

3.3 CONTROL: CONSTRUCTING AND CALIBRATING CONTROL INTERVENTIONS

We use the direction-specific census measurements to construct candidate sets of heads. For a model with baseline positional effect $\widehat{\beta}^{0}_{\text{p}}$, an intervention is bias-reducing when its change in β_{p} opposes

the sign of the baseline effect. We retain heads whose bias-reducing effect lies at least five robust standard deviations below the median single-head effect. Candidate sets are constructed separately for attenuation and amplification.

We consider two rules for forming candidate sets. The first retains bias-reducing heads whose individual intervention also increases the mean decision margin by at least 10%. The second retains bias-reducing heads whose effects on β_{e} fall within a robust range around the median single-head effect. The first rule favors heads that simultaneously reduce position dependence and strengthen the correct decision, whereas the second more directly limits disruption to evidence sensitivity.

Single-head effects need not combine additively. We therefore do not predict the effect of a candidate set by summing its members' individual effects. Instead, we intervene on all heads in the set jointly using a shared gain α in Eq. 4 and directly re-estimate β_{p}, β_{e}, and the mean decision margin.

For each candidate set, we sweep a fixed grid of gains on the development set. When β_{p} changes sign between adjacent gains, we use linear interpolation to estimate a zero-bias operating point. We first consider sets constructed using the margin-improvement rule and then those constructed using the evidence-preservation rule. A zero crossing is accepted as low-cost when its evidence coefficient is at least 90% of the baseline value.

If no candidate set produces an acceptable zero crossing, we apply a budgeted partial-control fallback. Among the evaluated interventions that retain at least 90% of both the baseline evidence coefficient and mean decision margin, we select the one with the smallest $|\beta_{\mathrm{p}}|$. The unchanged model is included as a candidate, so the procedure returns no intervention when no admissible intervention improves on the baseline. After selection, the head set and gain are frozen and evaluated on held-out data. Full pseudocode and implementation details are provided in Appendix A.

4 Experimental setup

We evaluate six instruction-tuned models from four families: Gemma-2-2B and Gemma-2-9B (Gemma Team, 2024), Qwen2.5-3B and Qwen2.5-7B (Qwen Team, 2025), Phi-3-mini (Abdin et al., 2024), and Mistral-7B (Jiang et al., 2023). Localization is shown primarily for Gemma-2-2B, with corresponding results for all other models in the appendix; the head census and control experiments cover all six models.

Numerical comparisons use score gaps $s_{\mathrm{hi}} - s_{\mathrm{lo}} \in \{1, 2, 3, 6, 12, 24, 32\}$ and paired presentation orders with identical identifiers and scores. Head selection and gain calibration use development data, after which the intervention is frozen for held-out evaluation. Exact checkpoints, prompts, data construction, readout positions, decoding settings, numerical precision, and statistical procedures are given in Appendix A.

5 Results

5.1 Behavioral consistency can hide positional effects

Choice-point predictions show substantial differences in order sensitivity across models. Qwen2.5-3B reaches 95.7% margin-sign accuracy when the higher-scoring item appears first, its favored position, but only 81.0% when it appears second. Phi-3 exhibits the opposite preference: its accuracy rises from 50.0% with the higher-scoring item first to 91.9% with it second. By contrast, Gemma-2-9B and Qwen2.5-7B achieve 99.8–100% accuracy in both orders. Judged only by the sign of the decision margin, these two models appear nearly insensitive to presentation order.

The decision margins reveal a different picture. As shown in Table 1, all six models have a significant positional effect, ranging in magnitude from 0.83 to 6.46 logits. In particular, Gemma-2-9B and Qwen2.5-7B retain positional shifts of -2.57 and -2.73 logits despite their near-perfect margin-sign accuracy. Behavioral consistency therefore does not imply that presentation order has ceased to influence the model's decision.

Whether this latent shift changes the answer depends on its size relative to the comparison margin. Qwen2.5-7B has a mean margin of 21.1 logits, so its 2.73-logit positional shift rarely approaches the decision boundary. Phi-3, by contrast, has a mean margin of only 1.22 logits and a positional effect of 2.36 logits, making presentation order large enough to reverse many choices. Its evidence

Model	β_p	β_e	Margin-sign acc.	Acc. (high first / second)
Gemma-2-2B	$\mathbf{+0.83}$	$\mathbf{0.05}$	91.9%	90.5 / 93.3%
Gemma-2-9B	$\mathbf{-2.57}$	$\mathbf{0.04}$	99.9%	100 / 99.8%
Qwen2.5-3B	$\mathbf{-4.31}$	$\mathbf{0.12}$	88.3%	95.7 / 81.0%
Qwen2.5-7B	$\mathbf{-2.73}$	$\mathbf{0.10}$	100%	100 / 100%
Phi-3	$\mathbf{+2.36}$	0.001	71.0%	50.0 / 91.9%
Mistral-7B	$\mathbf{-6.46}$	$\mathbf{0.06}$	94.4%	99.8 / 89.0%

Table 1: Behavioral performance and fitted positional and evidence effects. "High first / second" indicates whether the higher-scoring item is presented first or second. Bold coefficients are significant at $p < 0.01$.

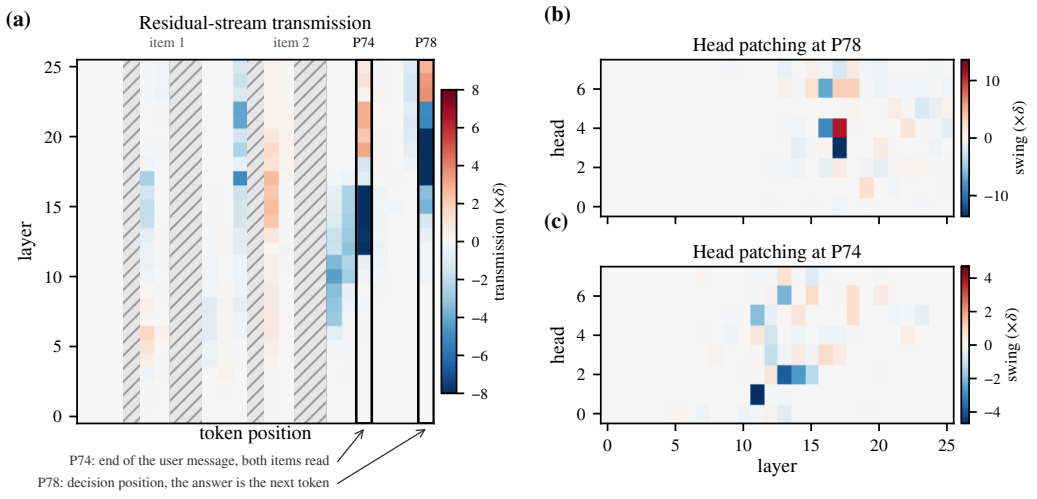

Figure 1: **Localizing the positional effect in Gemma-2-2B**. **(a)** Residual-stream transmission by layer and token position, in units of the mean full-swap effect δ (clipped at ± 8). Hatched columns are the item-content slots, masked because patching them replaces the items rather than the order information; framed columns are the two positions that transmit the swap effect, P74 (end of the user message) and P78 (decision position). **(b,c)** Margin change from patching each head's output at the two positions, in units of δ; the strongest effects are sparse and confined to layers 11–18.

coefficient is also near zero, so increasing the score gap does little to counteract this effect. Strong comparison evidence can therefore mask the behavioral consequences of a positional effect without eliminating the effect itself.

5.2 Positional effects localize to late decision states

Using the paired-swap activation-patching procedure described in Section 3.2, we localize where presentation order influences the decision margin. Figure 1 shows that, in Gemma-2-2B, transmission is concentrated at a small number of late prompt positions. The effect first appears near the `<end_of_turn>` token around layers 12–16 and then shifts to the decision position immediately before answer generation, with the strongest late-layer transmission around layer 18. The same late-stage pattern—with transmission at the end of the user message in earlier layers and at the decision position in later layers—appears across the other five models, although the extent to which the effect is concentrated in individual heads varies substantially (Appendix B).

Transmission changes substantially across these late layers. At intermediate layers, patching the swapped activation moves the decision margin opposite to the full-swap effect, producing transmission values approaching -8. At later layers, the effect reverses and approaches a transmission of 1, indicating that the patch increasingly reproduces the swapped-run margin. The positional influence is therefore not simply propagated unchanged through the residual stream; its effect on the decision is transformed across late layers.

Model	Heads (α)	$\beta_{\mathrm{p}}^{0} \rightarrow \beta_{\mathrm{p}}^{\mathrm{int}}$	β_{p} removed	$\Delta\beta_{\mathrm{e}}$	$\Delta\bar{m}$
Gemma-2-2B	1 (2.35)	$+1.14 \rightarrow +0.46$	60%	+19%	+42%
Gemma-2-9B	22 (0.47)	$-2.58 \rightarrow +0.09^{\ddagger}$	96%	-6%	+16%
Qwen2.5-3B	15 (1.73)	$-3.98 \rightarrow -0.47^{\ddagger}$	88%	+9%	0%
Qwen2.5-7B	28 (1.31)	$-2.41 \rightarrow +0.28^{\ddagger}$	89%	-1%	+1%
Phi-3†	3 (3.00)	$+2.33 \rightarrow +0.57$	76%	–	+150%
Mistral-7B†	31 (0.50)	$-6.44 \rightarrow -5.16$	20%	-3%	-6%

Table 2: Held-out effects of the selected head interventions. β_{p} removal is $1 - |\beta_{\mathrm{p}}^{\mathrm{int}}|/|\beta_{\mathrm{p}}^{0}|$; changes in β_{e} and mean margin $\bar{m}$ are relative to baseline. ‡ denotes a residual β_{p} not significantly different from zero; † denotes use of the partial-control fallback.

Model	ΔBalanced acc.	ΔDisfavored acc.	ΔGen. disfavored	$\Delta\mathcal{L}_{\mathrm{Wiki}}$
Gemma-2-2B	+3.8	$+3.8$ $[+2.1, +5.7]$	$+4.2$ $[+2.3, +6.5]$	+0.51%
Gemma-2-9B	-0.1	ceiling	ceiling	+0.30%
Qwen2.5-3B	+0.6	$+3.8$ $[+1.9, +5.7]$	$+3.6$ $[+1.4, +6.2]$	+1.34%
Qwen2.5-7B	-0.2	ceiling	ceiling	+0.25%
Phi-3†	+20.3	$+38.8$ $[+34.5, +43.3]$	$+42.9$ $[+37.6, +48.2]$	+2.56%
Mistral-7B†	+0.6	$+1.4$ $[0.0, +3.1]$	$+1.1$ $[-0.3, +2.8]$	+0.69%

Table 3: Behavioral and off-task effects on held-out data. Accuracy changes are percentage points; brackets give 95% confidence intervals. "Disfavored" denotes the presentation order with lower baseline accuracy. Wikipedia loss changes are relative to baseline. † denotes the partial-control fallback.

Head-output patching further shows that a small number of late-layer heads can produce large changes in the decision margin in Gemma-2-2B (Fig. 1b,c). However, transmission magnitude alone does not identify a useful control target: a head that strongly transmits the swap effect may also contribute useful comparison evidence, and strongly transmitting heads can have opposing effects. We therefore use activation patching to localize where positional influence reaches the decision, and the functional head census in Section 3.2.3 to determine which components can be manipulated selectively.

5.3 Selected head interventions provide positional control

We next test whether interventions selected on development data can reduce the positional effect without substantially weakening the underlying comparison. For each model, we jointly vary the selected heads using a shared gain and choose the operating point using the procedure in Section 3.3. The procedure finds a qualifying zero crossing for Gemma-2-2B, Gemma-2-9B, Qwen2.5-3B, and Qwen2.5-7B. For Phi-3 and Mistral-7B, no qualifying zero crossing is available, so the budgeted fallback selects the admissible intervention with the smallest remaining positional effect. The selected head set and gain are then frozen and evaluated on a fresh held-out draw (Table 2).

On held-out comparisons, the four zero-crossing interventions remove 60–96% of the positional effect. For Gemma-2-9B, Qwen2.5-3B, and Qwen2.5-7B, the residual positional coefficient is no longer significantly different from zero. This control does not come from uniformly suppressing the comparison signal: β_{e} increases by 19% for Gemma-2-2B and 9% for Qwen2.5-3B, while decreasing by only 6% and 1% for Gemma-2-9B and Qwen2.5-7B. Mean decision margin is preserved or increased in all four models.

The two fallback cases reveal different limits of control. For Phi-3, the selected three-head intervention removes 76% of the positional effect, reducing β_{p} from $+2.33$ to $+0.57$, while increasing the mean decision margin by 150%. Its evidence coefficient remains near zero and is therefore not meaningfully identified. Mistral-7B is substantially less controllable under the same performance budget: the intervention removes only 20% of the positional effect while decreasing β_{e} and mean margin by 3% and 6%, respectively. Thus, failure to find a qualifying zero crossing can correspond to very different degrees of remaining controllability.

We next test whether reducing the latent positional effect improves observed behavior and whether the intervention produces broader off-task changes (Table 3).

		Recipe transfer			Procedure transfer		
Model	Task	β_p removed	$\Delta\beta_e$	Δ disfav. acc.	β_p removed	$\Delta\beta_e$	Δ disfav. acc.
Gemma-2-2B	Pop.	-4%	$+30\%$	$+2.1$	97%	$+33\%$	$+23.0$
	Count	7%	$+50\%$	$+13.0$	$57\%^{\dagger}$	$+237\%$	$+48.7$
Gemma-2-9B	Pop.	-28%	$+8\%$	-7.0	$61\%^{\dagger}$	-7%	$+16.1$
	Count	30%	$+3\%$	$+2.3$	89%	-4%	$+10.3$
Qwen2.5-3B	Pop.	18%	-19%	0.0	$21\%^{\dagger}$	-15%	$+4.6$
	Count	8%	$0.14 \rightarrow 0.26$	$+3.3$	$54\%^{\dagger}$	$0.14 \rightarrow 0.51$	$+22.0$
Qwen2.5-7B	Pop.	12%	$+9\%$	$+3.2$	78%	-1%	$+6.0$
	Count	19%	$+11\%$	$+4.3$	59%	-1%	$+5.0$
Phi-3	Pop.	78%	$+59\%$	$+16.5$	92%	$+65\%$	$+18.7$
	Count	-300%	$+92\%$	-11.0	100%	$+114\%$	$+17.0$
Mistral-7B	Pop.	13%	-13%	$+2.8$	$38\%^{\dagger}$	-5%	$+11.3$
	Count	8%	$0.11 \rightarrow 0.09$	-1.3	$32\%^{\dagger}$	$0.11 \rightarrow 0.42$	$+37.3$

Table 4: Same-draw transfer results on previously unused held-out pairs. Negative β_p removal indicates an increased positional effect. For nonsignificant baseline β_e, absolute coefficients are shown instead of relative changes. † denotes the partial-control fallback.

The behavioral changes are consistent with the latent-control results. For Gemma-2-2B and Qwen2.5-3B, margin-sign accuracy in the presentation order disfavored at baseline increases by 3.8 points for both models, while greedy-generation accuracy in that order increases by 4.2 and 3.6 points, respectively. Gemma-2-9B and Qwen2.5-7B remain effectively at ceiling, with changes of -0.1 and -0.2 points in balanced margin-sign accuracy.

Phi-3 shows the largest behavioral response. Its balanced margin-sign accuracy rises from 71.0% to 91.3%, a gain of 20.3 points, driven primarily by a 38.8-point improvement in the presentation order disfavored at baseline. Under greedy generation, balanced accuracy rises from 73.8% to 96.1%, with a 42.9-point gain in the disfavored order. Mistral-7B changes much more modestly: balanced accuracy increases by 0.6 points, while its 1.1-point generated disfavored-order gain has a confidence interval that includes zero.

The interventions produce modest but nonzero broader model changes. On 200 held-out Wikipedia passages, relative per-token loss increases by less than 0.7% for Gemma-2-2B, Gemma-2-9B, Qwen2.5-7B, and Mistral-7B, by 1.34% for Qwen2.5-3B, and by 2.56% for Phi-3. For Qwen2.5-3B, generated-answer accuracy is conditional on parseable outputs because a nontrivial fraction of generations do not emit either identifier.

6 Transfer across comparison tasks

The numerical task provides a controlled setting for identifying and manipulating positional effects, but it supplies the comparison evidence directly as scores. We next ask whether the same control framework extends to tasks in which the model must obtain the relevant evidence itself.

We consider two such tasks. In *country-population* comparison, the model is given two countries and asked which has the larger population, without population values in the prompt. In *letter counting*, it is given two strings and must determine which contains more occurrences of a specified target letter. The former requires retrieving evidence from model knowledge, while the latter requires computing it from the prompt. Both preserve the two-slot comparison format while changing how the evidence is obtained.

We evaluate generalization at two levels. In *recipe transfer*, we retain the head set and gain selected on numerical comparison, recomputing only the task-specific centering means. This tests whether a particular intervention remains effective when the comparison task changes. In *procedure transfer*, we rerun the same selection procedure of Section3.3 on each target task, using the same selection rules, thresholds, and gain calibration. This tests whether the procedure can identify an effective task-specific intervention. For each model–task setting, the frozen recipe and the task-specific procedure are evaluated on the same previously unused held-out pairs, against the same baseline.

Recipe transfer. The frozen numerical-task interventions transfer inconsistently (Table 4). Positional-effect reduction ranges from -300% to 78%, exceeds 30% in only one of twelve settings, and is negative in three. The limitation is particularly apparent when the direction of the positional effect changes across tasks. In the four such settings, the frozen recipe removes -4%, 7%, -28%, and -300% of the effect, respectively, providing little control or increasing the positional shift.

Behavior can nevertheless improve even when positional control transfers poorly. For example, on Gemma-2-2B letter counting, the frozen intervention increases evidence sensitivity by 50% and improves accuracy in the baseline-disfavored order by 13.0 percentage points, while removing only 7% of the positional effect. Thus, an intervention can preserve or strengthen comparison-relevant effects across tasks without preserving its control of positional dependence.

Procedure transfer. Rerunning the same selection procedure of Section 3.3 on each target task yields a qualifying zero crossing in six of the twelve settings and invokes the partial-control fallback in the remaining six. On held-out comparisons, the zero-crossing interventions remove 59–100% of the positional effect, while the fallback interventions remove 21–61%. Thus the task-specific procedure reduces the positional-effect magnitude in all twelve settings, by 21–100%.

These reductions do not generally come from weakening comparison evidence. Among the six zero-crossing interventions, held-out β_e either increases or decreases by at most 4%. Among the four fallback settings with a significantly positive baseline β_e, three increase it or reduce it by at most 7%; Qwen2.5-3B on population comparison is the exception, with a 15% decrease. In the two remaining fallback settings, where baseline β_e is not significantly different from zero, the absolute coefficient increases from 0.14 to 0.51 for Qwen2.5-3B counting and from 0.11 to 0.42 for Mistral-7B counting.

The task-specific procedure also recovers control when the positional effect changes direction across tasks. Across the four sign-reversal settings, it removes 57–100% of the effect. The contrast is especially clear for Gemma-2-9B on population comparison, where the frozen numerical recipe increases the effect by 28% and reduces accuracy in the disfavored order by 7.0 points, whereas the task-specific intervention removes 61% of the effect and improves that order by 16.1 points. On Phi-3 letter counting, the frozen recipe increases the positional-effect magnitude fourfold and reduces disfavored-order accuracy by 11.0 points; rerunning the procedure instead reduces the estimated positional effect to approximately zero and improves that order by 17.0 points.

Because both approaches are evaluated on identical held-out pairs, they can be compared directly. The task-specific procedure has a larger positional-effect reduction point estimate than the frozen recipe in all 12 settings, and a larger disfavored-order accuracy-gain point estimate in all 12. Its balanced-accuracy point estimate increases relative to baseline in every setting. These results distinguish the generality of the control procedure from the transferability of any particular intervention: the same control procedure applies across tasks, but the heads and gains needed to control positional dependence are task-dependent.

7 Discussion and limitations

Our results separate behavioral order invariance from the absence of positional influence: strong comparison evidence can make a several-logit positional shift behaviorally invisible. They further show that this influence is partially separable from useful comparison evidence, since selected head-level interventions can substantially reduce β_p while largely preserving β_e and decision margins.

The required heads and gains nevertheless vary across models and tasks, and frozen intervention recipes transfer inconsistently; what generalizes more reliably is the procedure for identifying a task-specific intervention. Our experiments are limited to controlled two-item comparisons, require access to internal activations and target-task development data, and do not establish what the localized states represent. Extending the analysis to naturalistic preference judgments and more task-independent control remains future work.

Reproducibility Statement

We provide the prompt construction, data splits, readout definitions, activation-patching procedure, head-intervention method, selection thresholds, gain calibration, held-out evaluation protocol, and statistical procedures in Appendix A. The appendix also gives the constants and pseudocode used to select control interventions, together with additional localization results across models. Code and supporting materials will be released with the final version.

AI Use Statement

Generative AI tools were used to assist with language editing, organization, and critical feedback on the presentation and interpretation of the methodology and results. The authors reviewed and verified all AI-assisted content and take responsibility for the final experiments, analyses, results, and claims.

References

Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pp. 7324–7338. PMLR, 2022.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36:76033–76060, 2023.

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

Ruizhe Li and Yanjun Gao. Anchored answers: Unravelling positional bias in GPT-2's multiple-choice questions. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. arXiv:2405.03205.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? Evidence from multiple choice capabilities in Chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

Adian Liusie, Potsawee Manakul, and Mark Gales. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 139–151, 2024.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of AACL-IJCNLP*, 2025. arXiv:2406.07791.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *Conference on Language Modeling*, 2024a. arXiv:2404.08382.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. *arXiv preprint arXiv:2407.01100*, 2024b.

Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. Mitigate position bias in large language models via scaling a single dimension. *arXiv preprint arXiv:2406.02536*, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024. arXiv:2309.03882.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

A Implementation and evaluation details

This appendix provides the implementation details for Sections 3.1–3.3. Unless stated otherwise, all experiments run in float32 with eager attention. Interventions are applied through TransformerLens hooks to each attention head's output (`hook_z`) before the output projection.

A.1 Models, prompts, and data

Identifiers. For each model, we construct a tokenizer-specific pool of candidate identifiers. An identifier is a string of 2–6 capital letters that tokenizes to a single token both in isolation and after a space, and that the model does not continue when prompted with it alone. The resulting pools contain 196–1329 identifiers, depending on the tokenizer. The two identifiers within a comparison are sampled without replacement.

Scores and gaps. Scores are four-character decimals of the form `XX.X`. For score gap g, the lower score is drawn uniformly from $[10.0, 99.9 - g]$, and the higher score is obtained by adding g. Development and evaluation use

$$g \in \{1, 2, 3, 6, 12, 24, 32\}. \tag{7}$$

Prompt and readout. The prompt instructs the model to rank all items from highest to lowest score and to output only the ordered identifiers, one per line. It then presents the two lines `ID: score` and is wrapped in the model's chat template. The decision position is the position immediately before the first identifier is generated. Mistral-7B emits a fixed preamble (`1. `) before the first identifier; we teacher-force this preamble and read the logits for the space-prefixed identifier tokens at the resulting decision position. Phi-3 also uses the space-prefixed identifier tokens. The remaining models use the bare identifier tokens.

Alignment and splits. A comparison is retained only if its two presentation orders have equal token length and differ only within the two item lines, affecting at most ten token positions. This ensures position-wise alignment for activation patching. The development set contains 40 pairs per gap, or 280 pairs and 560 ordered prompts in total. Evaluation draws use a separate random seed and contain 60 pairs per gap, or 420 pairs. No evaluation pair has the same identifier pair and scores as a development pair. Individual identifiers may recur across splits, but complete comparisons do not.

Margin and regression. We read the margin in Eq. 1 from the pre-softmax logits at the decision position under teacher forcing. Equation 2 is fit by ordinary least squares over both orders of every pair, with standard errors clustered by pair. Margin-sign accuracy is the fraction of prompts with $m > 0$, and balanced accuracy averages accuracy across the two presentation orders. The *disfavored order* places the higher-scoring item in slot 1 when $\beta_{\mathrm{p}} > 0$ and in slot 2 when $\beta_{\mathrm{p}} < 0$.

A.2 Localization experiments

Residual-stream patching replaces the pre-block residual state (`resid_pre`) at one layer and token position in the target run with the corresponding state from the source run, then recomputes the decision margin. Transmission in Eq. 3 is computed as the ratio of the mean patched effect to the mean full-swap effect, rather than as the mean of per-pair ratios.

After identifying late residual-stream states with high transmission, we replace each attention head's output at P74 or P78 with its value from the source run. The main localization figure and the corresponding appendix figure use 14 pairs per gap, for 98 development pairs in total. Because transmission is normalized by the mean swap effect of the evaluated draw, it can become large when that denominator is small. The Gemma-2-2B residual map in Figure 1 is clipped to $[-8, 8]$ for visualization; the other-model maps in Figure 2 state their scales separately.

A.3 Head intervention and control selection

Single-head census. Every attention head is intervened on individually using Eq. 4 at $\alpha = 0$ and $\alpha = 2$. The position-wise mean $\overline{o}_h$ is estimated over both presentation orders of all development prompts. After each intervention, we refit Eq. 2 and record the changes in the positional coefficient, evidence coefficient, and mean margin.

For head h and intervention gain α, define the sign-adjusted positional effect

$$d_{\mathrm{B}}(h, \alpha) = \operatorname{sign}\left(\widehat{\beta}_{\mathrm{p}}^{0}\right)\left[\widehat{\beta}_{\mathrm{p}}(h, \alpha) - \widehat{\beta}_{\mathrm{p}}^{0}\right], \tag{8}$$

so that negative values move the baseline positional effect toward zero. We also define

$$d_{\mathrm{E}}(h, \alpha) = \widehat{\beta}_{\mathrm{e}}(h, \alpha) - \widehat{\beta}_{\mathrm{e}}^{0}. \tag{9}$$

For either effect distribution d, we estimate a robust null scale as

$$\widehat{\sigma}(d) = 1.4826 \ \operatorname{median}_h \left| d(h) - \operatorname{median}_{h'} d(h') \right|. \tag{10}$$

Thus, the thresholds below are measured in robust standard deviations estimated from the median absolute deviation; they are not raw multiples of the MAD. Attenuation candidates are screened using their measurements at $\alpha = 0$, and amplification candidates using their measurements at $\alpha = 2$.

Candidate sets. For each intervention direction, we retain heads whose d_{B} lies at least k_{B} robust standard deviations below the median. From these bias-reducing heads, the *margin-gain* rule retains those whose individual intervention increases the mean decision margin by at least γ. The *evidence-preservation* rule instead retains heads whose d_{E} lies within k_{E} robust standard deviations of its median. Applying both rules separately to attenuation and amplification produces at most four candidate sets per model and task.

Table 5 lists the constants shared by every model and task. A wider range was explored during method development; these values were fixed before held-out evaluation.

Quantity	Value	Role
k_{B}	5 robust SDs	Bias-reduction threshold
k_{E}	5 robust SDs	Evidence-effect gate
γ	10%	Minimum individual margin gain
$\mathcal{A}$	$\{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 2, 2.5, 3\}$	Joint gain grid
ρ_{E}	0.90	Minimum retained evidence coefficient
ρ_{M}	0.90	Minimum retained mean margin in fallback

Table 5: Constants used for candidate construction, gain calibration, and budgeted partial control.

Joint sweep and zero crossing. All heads in a candidate set are intervened on jointly with a shared gain $\alpha \in \mathcal{A}$. At every grid point, we refit Eq. 2 and record $\widehat{\beta}_{\mathrm{p}}$, $\widehat{\beta}_{\mathrm{e}}$, mean margin, and balanced accuracy. If $\widehat{\beta}_{\mathrm{p}}$ changes sign between adjacent grid points, we linearly interpolate the gain and recorded quantities on that segment to estimate the zero crossing. On held-out data, the model is evaluated directly at the resulting interpolated gain.

A crossing is low-cost when its interpolated evidence coefficient is at least ρ_{E} times the baseline value. If the baseline evidence coefficient is not significantly different from zero at $p < 0.05$, as for Phi-3 on the numerical task, the relative evidence change is undefined and the same threshold is applied to mean margin instead. Among low-cost crossings, the margin-gain family is preferred to the evidence-preservation family. If both attenuation and amplification within the preferred family cross, we select the crossing with higher development balanced accuracy.

Budgeted partial-control fallback. When no low-cost crossing exists, we consider every already-evaluated candidate (S, α) together with the unchanged model. A candidate is admissible when it retains at least ρ_{E} of the baseline evidence coefficient and ρ_{M} of the baseline mean margin. If the baseline evidence coefficient is not identified, only the margin budget is applied. Among admissible candidates, we choose the one with the smallest $|\widehat{\beta}_{\mathrm{p}}|$. Including the unchanged model ensures that no intervention is returned when none improves the positional effect within the budgets.

The fallback rule was formulated after the initial analyses. To avoid selecting on data that had already been inspected, all results reported through it were evaluated only after the head set and gain were frozen, on draws not used in any earlier analysis. For the numerical task, this applies to the Mistral-7B intervention (Table 2). For the transfer tasks, both the frozen numerical recipe and the task-specific intervention were evaluated on a single fresh held-out draw per model–task setting, so every row of Table 4, zero-crossing and fallback alike, is compared against the same baseline on the same pairs.

Selection algorithm. Algorithm 1 summarizes candidate construction, calibration, and fallback selection for one model and task.

A.4 Held-out evaluation

After selection, the head set and gain are frozen and applied to a held-out evaluation draw. The position-wise means $\overline{o}_h$ are recomputed over the evaluation prompts; this is part of applying the intervention and does not use scores, labels, or evaluation outcomes. We report $\widehat{\beta}_{\mathrm{p}}$, $\widehat{\beta}_{\mathrm{e}}$, mean margin, and margin-sign accuracy overall and separately by presentation order.

Accuracy changes are accompanied by 95% percentile intervals from 1000 bootstrap resamples over pairs. Both orders of a pair are resampled together, and baseline and intervened predictions are paired within each resample.

Generated answers. For models below ceiling, we additionally generate greedily for eight tokens with the intervention active. We take the first generated token matching either identifier in either its bare or space-prefixed encoding. Generated accuracy pools score gaps $g \geq 2$ and is conditional on emitting a parseable identifier. Unparsed trials are excluded from accuracy.

Algorithm 1 Selecting a positional-control intervention

Require: Development pairs $\mathcal{D}$, heads $\mathcal{H}$, gain grid $\mathcal{A}$, and constants in Table 5
Fit Eq. 2 on $\mathcal{D}$ to obtain baseline $\widehat{\beta}_{\mathrm{p}}^{0}$, $\widehat{\beta}_{\mathrm{e}}^{0}$, and $\bar{m}^{\,0}$.
for all $h \in \mathcal{H}$ and $\alpha \in \{0, 2\}$ **do**
    Intervene on h alone; refit Eq. 2; record d_{B}, d_{E}, and the fractional margin change.
end for
for all directions $a \in \{0, 2\}$ **do**
    Estimate the direction-specific medians and robust scales of $d_{\mathrm{B}}(\cdot, a)$ and $d_{\mathrm{E}}(\cdot, a)$.
    Let B_a contain heads whose d_{B} is at least k_{B} robust SDs below its median.
    Form $S_{\mathrm{mg},a}$ from heads in B_a with margin gain $\geq \gamma$.
    Form $S_{\mathrm{ev},a}$ from heads in B_a whose d_{E} lies within k_{E} robust SDs of its median.
end for
for all candidate sets S and gains $\alpha \in \mathcal{A}$ **do**
    Intervene on all heads in S at gain α; refit and store $\widehat{\beta}_{\mathrm{p}}$, $\widehat{\beta}_{\mathrm{e}}$, $\bar{m}$, and balanced accuracy.
end for
for all candidate sets S **do**
    Interpolate the first adjacent gain interval on which $\widehat{\beta}_{\mathrm{p}}$ changes sign.
end for
Let $\mathcal{C}$ be the low-cost crossings from margin-gain sets; if empty, use the low-cost crossings from evidence-preservation sets.
if $\mathcal{C} \neq \emptyset$ **then**
    return the crossing in $\mathcal{C}$ with highest development balanced accuracy.
end if
Let $\mathcal{F}$ contain the unchanged model and all swept candidates that satisfy the evidence and margin budgets.
return $\arg\min_{c \in \mathcal{F}} |\widehat{\beta}_{\mathrm{p}}(c)|$.

Off-task cost. We measure mean per-token cross-entropy on 200 held-out Wikipedia passages, each containing at least 400 characters and truncated to its first 512 tokens. The intervention hooks remain active. The position-wise mean $\overline{o}_h$ is taken from the development prompts and extended beyond their length by repeating its final position. We report the relative loss change from the intact model, with a 95% bootstrap interval over passages.

B LOCALIZATION ACROSS MODELS

Section 5.2 localizes the positional effect in Gemma-2-2B (Figure 1). This appendix reports the same measurements for the other five models and states what they contribute to, and what they do not contribute to, the selection of control interventions.

Procedure. For each model we use its own development set (98 paired comparisons, 14 at each of the seven gaps; float32), so the stimuli, the identifier pool and the readout position match those used for the census and for calibration. For every layer and token position we replace the target run's residual-stream state with the corresponding state of its swap-mate and recompute the decision margin, reporting transmission as the ratio of mean effects (Eq. 3). We then repeat the measurement for individual attention heads at the positions that transmit the effect. The scale δ, the mean full-swap effect, differs by an order of magnitude across models (0.30 for Gemma-2-2B on this draw, 6.42 for Mistral-7B), so each panel is scaled independently and colour intensity is not comparable across models.

The register structure replicates. In every model the swap effect is transmitted at two token positions and nowhere else: the token that closes the user message and the decision position itself. The first carries the effect in earlier layers and the second in later ones. The positions differ only as the chat template dictates: P74 and P78 for Gemma-2-2B, P75 and P78 for Gemma-2-9B, P88 and P93 for both Qwen models, P74 and P76 for Phi-3, and P80 and P86 for Mistral-7B, whose decision position follows the teacher-forced preamble (Appendix A). Transmission is negative in the mid-stack of every model, meaning that a patched state moves the receiving run's margin toward the item that occupies the donor's winning slot rather than toward the donor's winning item.

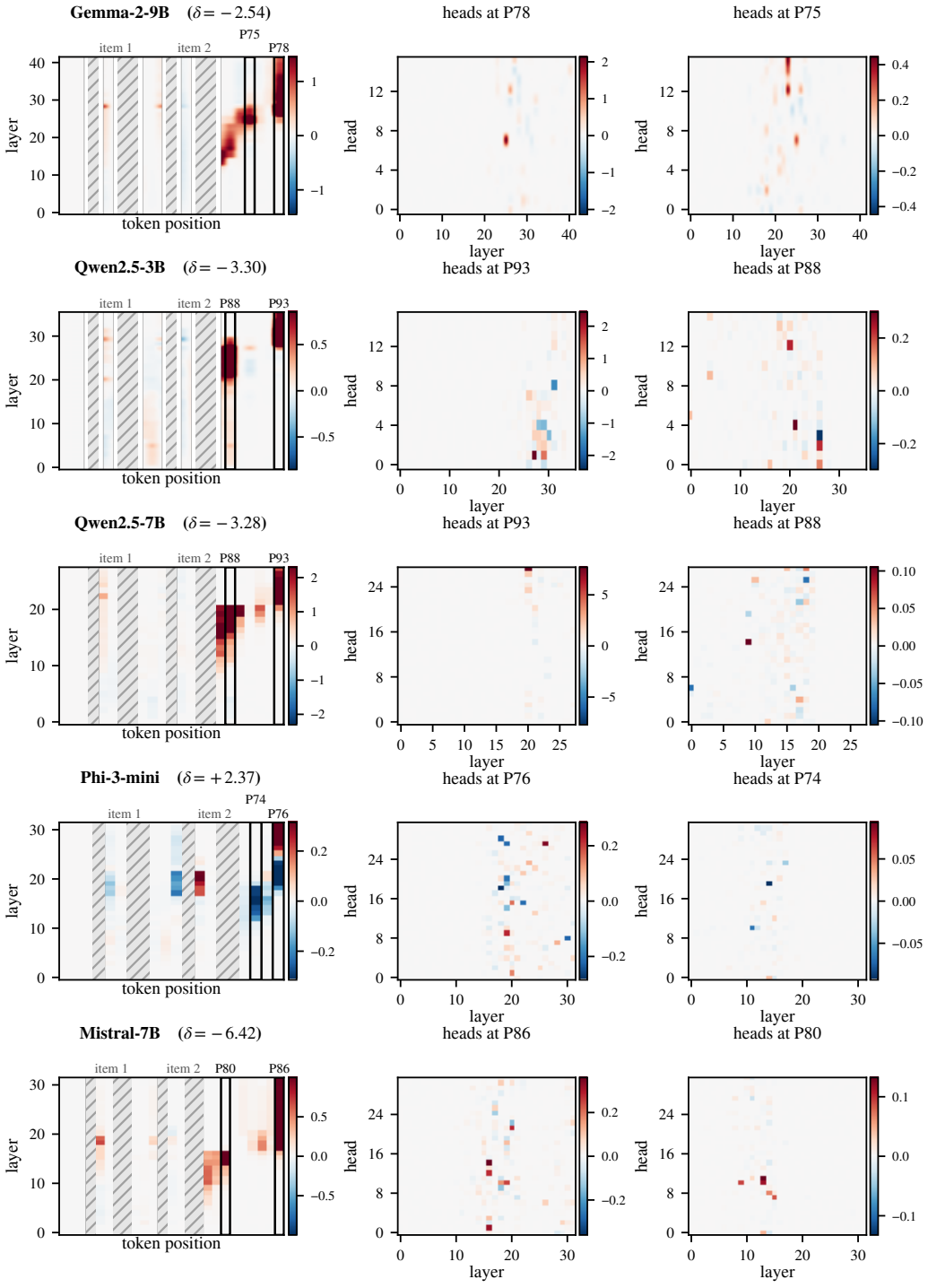

Figure 2: **Localization in the five remaining models** (development sets, 98 pairs each; float32). Rows: Gemma-2-9B, Qwen2.5-3B, Qwen2.5-7B, Phi-3-mini, Mistral-7B, with each model's mean full-swap effect δ in the row title. Left: residual-stream transmission by layer and token position, with the item-content slots masked and the two transmitting positions framed, as in Figure 1. Middle and right: change in the decision margin from patching each head's output at those two positions, in units of that model's own δ. Each panel is scaled independently, so the location and sparsity of the bands are comparable across rows but colour intensity is not.

Concentration does not replicate. How much of the effect any single head carries varies by more than an order of magnitude. At its decision position Gemma-2-2B has one head worth $13.6\,\delta$ and Qwen2.5-7B one worth $7.7\,\delta$, whereas the largest single head in Phi-3-mini reaches $0.28\,\delta$ and in Mistral-7B $0.36\,\delta$; Qwen2.5-3B and Gemma-2-9B lie between, at 2.4 and $2.1\,\delta$. This ordering matches the census: the two models with no concentrated writer are the two whose positional effect the census finds distributed over the largest number of heads, and they are also the two models for which no low-cost zero crossing exists (Section 5).

Localization narrows the search; it does not make the selection. Transmission measures how much of the swap effect passes through a component. It does not distinguish a head that carries the positional prior from one that carries the comparison evidence, and the two are not separable by magnitude. In Gemma-2-2B the two largest-transmission heads at the decision position form an opposing pair, $-13.6\,\delta$ and $+10.9\,\delta$, whose net contribution is small; the head the method selects is the third largest ($-8.5\,\delta$). Across the six models, of the ten highest-transmission heads only zero to three appear in the selected intervention, and the selected heads occupy a wide range of transmission ranks (median rank 3 of 208 in Gemma-2-2B, 74 of 576 in Qwen2.5-3B, 63 of 784 in Qwen2.5-7B). We therefore use localization to establish where in the network order information reaches the decision, and the functional census (Section 3.2.3) to decide which components to move.