

LEVEE: Layer-wise Evaluation of Vector Encoding Effects Bias Localization and Functional Separation across 11 LLMs

Dongmin Kim

Incheon National University
kdmin0611@inu.ac.kr

Youhyun Shin[†]

Korea University
syh809@korea.ac.kr

Abstract

Large language models encode social biases that instruction fine-tuning aims to suppress. Instruction fine-tuning applies its weight changes through the alignment task vector ΔW , yet prior work treats it as a monolithic object, leaving its internal layer structure opaque: which layers carry bias suppression, which carry general capability, and whether these are separable, remains unknown. We map this structure by injecting each layer’s ΔW_i independently into the base model across 11 LLMs from Qwen2.5, Gemma-2/3, and Llama-3 families (0.5B–14B). Testing via per-layer sufficiency sweeps, block removal, and a BBQ–MMLU dissociation, we find three results. First, BBQ-measured bias suppression localizes to the mid-third of transformer layers in 9 of 11 models, with mid-third removal raising bias across all 11 while first- and last-third removals produce smaller L-BBQ increases in all nine structured-signal models. Second, mid- and first-third removals dissociate BBQ and MMLU: mid-third removal is BBQ-critical across all tested models while first-third removal can sharply reduce MMLU in Gemma architectures, supporting functional separation of bias suppression and capability in weight space. Third, mid-third removal raises bias *above* the unaligned base in 3 of 11 cases: partial ΔW transfer without the mid-third can make a model *more* biased than the unaligned base model. These findings provide an empirical roadmap for targeted debiasing and safe task vector composition.¹

1 Introduction

Weight merging has emerged as an increasingly adopted paradigm for adapting large language models without retraining. Task arithmetic (Ilharco et al., 2023) formalizes this via the alignment task vector $\Delta W = W_{\text{instruct}} - W_{\text{base}}$: add it to a base model and the encoded behavioral changes transfer

with it. Subsequent methods extend this to interference resolution (Yadav et al., 2023) and sparsification (Yu et al., 2024), all treating ΔW as a monolithic object without examining its internal layer structure.

This monolithic view leaves a mechanistic question unanswered. Alignment fine-tuning encodes multiple behavioral changes simultaneously in ΔW (improved instruction-following, bias suppression, and refined refusal), but the internal layer structure of these encodings is entirely opaque. Critically, the growing literature on task vector composition (Yadav et al., 2023; Yu et al., 2024) focuses on capability-level transfers and interference resolution. None of these works has examined which layer ranges of ΔW carry bias suppression versus general capability, or whether these are separable. Prior work identifies behaviorally relevant directions in activation space, including representation engineering (Zou et al., 2023) and refusal-mediating directions (Arditi et al., 2024); the layer-wise organization of the paired alignment update ΔW itself has not been systematically studied. Without knowing this structure, task vector composition risks deploying models that are more biased than the unaligned base model.

We address this with a layer-wise interventional analysis of ΔW across 11 instruction-tuned LLMs (0.5B–14B; Qwen2.5, Gemma 2/3, Llama 3.x). We inject each layer’s ΔW_i independently into the base model, measure bias on BBQ (Parrish et al., 2022) and capability on MMLU (Hendrycks et al., 2021), and run five experimental designs spanning sufficiency (per-layer sweep, cumulative injection), necessity (block removal), directional specificity (sign-flip, norm-matched controls).

Our findings are consistent across model families. **(1) Layer-wise localization.** All 9 structured-signal models (§3) exhibit a mid-third minimum in the per-layer BBQ profile ($p \approx 5.1 \times 10^{-5}$, exact binomial); mid-third block removal reverses debi-

[†] Corresponding author.

¹ github.com/GDM1nu/EMNLP-2026-LEVEE

Family	Model	N
Qwen2.5	0.5B	24
	1.5B	28
	3B	36
	14B	48
Gemma	Gemma-2-2B	26
	Gemma-2-9B	42
	Gemma-3-12B [†]	48
Llama	Llama-3.2-1B	16
	Llama-3.2-3B	28
	Llama-3-8B	32
	Llama-3.1-8B [†]	32

Table 1: All 11 model pairs (Yang et al., 2024; Gemma Team, 2024, 2025; Meta AI, 2024a,b). [†]Direction-specificity: zero-injection baseline only.

asing in all 11 tested models ($\Delta\text{L-BBQ} + 0.007$ to $+0.215$; 10 of 11 most critical, $p \approx 1.3 \times 10^{-4}$), while first- and last-third removals produce smaller L-BBQ increases in all nine structured-signal models. **(2) Functional separation.** In Gemma-2 models, removing the first third reduces MMLU by up to 35.0 pp with $|\Delta\text{L-BBQ}| \leq 0.006$; mid-third removal causes larger BBQ changes but smaller MMLU losses. To our knowledge this is the first layer-wise interventional decomposition of an alignment task vector that separates stereotype suppression from general capability in weight space, across 11 base–instruction-tuned model pairs. Prior work established functional layer specialization within single models (Zhao et al., 2024), but not for the paired update. **(3) Active suppression.** By *active suppression* we mean that removing mid-third ΔW produces a model *more biased* than the unaligned base, indicating that mid-third layers do not merely deposit bias-neutral information but counteract a bias-amplifying contribution from the outer thirds, contrary to a passive-deposit view. This occurs in 3 of 11 models (up to +46% above base), so partial ΔW transfer that omits the mid-third makes a model more biased than the unaligned base model.

2 Related Work

Task vectors and model merging. Ilharco et al. (2023) introduced *task vectors*, showing ΔW encodes task-specific capabilities as an additive object in weight space; subsequent work extends this to interference resolution (Yadav et al., 2023; Yu et al., 2024). All treat ΔW as a monolithic object; none

examines which layer ranges carry which behavioral components.

Mechanistic interpretability and layer specialization. ROME (Meng et al., 2022) and its batch extension MEMIT (Meng et al., 2023) localize factual associations to mid-layer MLP weights via causal tracing; Dai et al. (2022) similarly identify knowledge neurons in specific FFN positions via gradient attribution. Transformer FFN layers act as key-value memories (Geva et al., 2021) that promote concepts in the vocabulary space (Geva et al., 2022). Sharma et al. (2024) show that reducing the rank of weight matrices at selected layers can improve reasoning performance, establishing that layer-selective weight modification changes behavior; it does not decompose a paired alignment update or separate distinct alignment-induced functions. Todd et al. (2024) show via causal mediation that *function vectors*—compact representations averaged over in-context demonstrations—are transported by a small set of mid-layer attention heads, establishing activation-level functional specialization. Zhao et al. (2024) identify three functional sections in instruction-tuned Llama-2 using representational comparison (MOSSA and CKA) between pretrained and instruction-tuned models, finding early/mid/late specialization. Concurrent work (Li et al., 2025) identifies contiguous “safety layers” in the middle portion of aligned LLMs, combining input-vector analysis with parameter-scaling experiments and fixing those layers’ gradients during fine-tuning. Critically, representation-level analyses measure *correlation* between representations and labels rather than causal transmission (Belinkov, 2022). These approaches differ in what they analyze—some the activations of a single model, others the layer organization or parameter sensitivity of an aligned model—but none decomposes the paired difference ΔW between a base and its instruction-tuned counterpart. The layer encoding a stereotype in the base model need not coincide with the layer whose ΔW_i encodes the alignment modification.

Bias measurement and weight-level intervention. Social bias in NLP is systematic and pervasive across models and tasks (Blodgett et al., 2020; Zhao et al., 2018); LLM-specific measurements use BBQ (Parrish et al., 2022), CrowS-Pairs (Nangia et al., 2020), and StereoSet (Nadeem et al., 2021). Targeted debiasing methods remove protected-attribute information directly, for

example by subspace nullification (Ravfogel et al., 2020; Liang et al., 2021). Instruction fine-tuning (Ouyang et al., 2022) and preference optimization (Rafailov et al., 2023) are broader alignment procedures that shift model behavior without targeting social bias specifically, and improvements on a bias metric need not indicate that the underlying structure has been removed (Gonen and Goldberg, 2019). Where in the alignment update such suppression is stored has not been established. Vig et al. (2020) apply causal mediation analysis to attention heads, and Zou et al. (2023) propose representation engineering (RepE), steering behaviors via activation-space interventions on a fixed aligned model. These approaches operate at the *activation* level on an already-aligned model. We ask instead which layers of ΔW *interventionally transmit* alignment-induced bias suppression by measuring the weight-level structure of the alignment modification itself, not the hidden-state representations it produces. Prior work thus studies task-vector merging, activation- or parameter-level localization within individual models, and generic layer importance largely as separate problems. LEVEE instead treats the paired alignment update itself as a layer-structured object, and intervenes on it directly rather than reading structure off a single model’s activations or weights.

3 Method

3.1 Task Vector Construction

For each model pair, the task vector is:

$$\Delta W_i = W_{\text{instruct},i} - W_{\text{base},i} \quad (1)$$

for $i \in \{0, \dots, N-1\}$, covering shared tensors within each transformer block (attention projections $q/k/v/o$, MLP weights, and block layer norms; 1-D bias tensors excluded); token embeddings, the output head, and the final model norm remain at their base values, including in the condition labeled *full*. Injection substitutes selected ΔW_i tensors into the base model before evaluation (lm-evaluation-harness); non-selected layers retain base weights. Weight-space statistics (Frobenius norm, sign coherence, participation ratio (Wegner, 1980)) require fully-decomposed ΔW tensors and are therefore computed on 9 of the 11 model pairs; Gemma-3-12B and Llama-3.1-8B are included in the per-layer sweep and block-removal experiments only (Table 1).

3.2 Bias Score

Our primary behavioral measure is a BBQ-derived ambiguous-context score, which we denote **L-BBQ** to distinguish it from the official BBQ bias score of Parrish et al. (2022). For each ambiguous-context response we assign a score

$$b_i = \begin{cases} +1 & \text{response classified as biased,} \\ -1 & \text{response classified as anti-biased,} \\ 0 & \text{UNKNOWN response,} \end{cases}$$

and define

$$S_{\text{L-BBQ}} = (1 - \text{Acc}_{\text{amb}}) \cdot \frac{1}{N} \sum_{i=1}^N b_i, \quad (2)$$

where N is the number of ambiguous-context responses that receive one of these three labels, and Acc_{amb} is accuracy on ambiguous items. The first factor scales the signed stereotyping direction by how often the model commits to a non-neutral choice.

How responses are classified. A response is labelled biased or anti-biased by matching the selected answer’s group against the `stereotyped_groups` metadata of the item; responses that the matching rule does not resolve to one of the three labels are excluded from N . Unlike the official BBQ scorer, L-BBQ does not reverse the biased/anti-biased assignment according to question polarity, and it normalizes the signed response score over all labelled responses, including UNKNOWN predictions. L-BBQ and the official BBQ bias score are therefore not the same metric; Appendix Q re-scores the three models shown in Figure 1 under the official metadata- and polarity-aware definition.

Answer selection uses log-probability multiple-choice scoring (lm-evaluation-harness v0.4.x; greedy decoding). The score is computed jointly over all 11 BBQ social-group categories (Appendix B); lower (toward zero) indicates less bias. Figures and tables report L-BBQ unless explicitly marked as “Official”; “BBQ category” refers to the benchmark’s social-group categories.

3.3 Layer Group Definitions

For a model with N layers, we partition layers into three equal contiguous blocks—first (L_0 – $L_{\lfloor N/3 \rfloor - 1}$), mid ($L_{\lfloor N/3 \rfloor}$ – $L_{\lfloor 2N/3 \rfloor - 1}$), and last ($L_{\lfloor 2N/3 \rfloor}$ – L_{N-1}). These thirds serve as aggregation units for thirds-level summary statistics and

as block-level conditions in the block-removal experiment. Per-layer best-layer attributions are computed from the layer sweep independently of this grouping, so the third boundaries do not influence which layer is reported as best.

We classify a model as having a *structured signal* if the per-layer BBQ range exceeds 15% of the base-model L-BBQ score. This threshold falls within a natural gap in our data: Qwen2.5-0.5B (range/base = 10%) and Qwen2.5-1.5B (2.6%) cluster well below it, while the remaining 9 models all exceed 23%—a >10 pp gap that makes the classification insensitive to the exact threshold value. We refer to the first and last thirds jointly as the *outer thirds*.

3.4 Experimental Designs

Five experimental designs address the three questions in §1. The per-layer sweep and block removal both cover all 11 models, spanning all three architecture families.

Experiment 1: Per-layer Sufficiency Sweep.

For each layer i , we construct $\hat{W}_i = W_{\text{base}} + \Delta W_i$, leaving all other layers at base weights. If a single mid-third layer produces a lower bias score than any other layer, that third is *sufficient* to carry the bias signal.

Experiment 2: Block Removal. We apply the full task vector but omit one contiguous third, $\hat{W} = W_{\text{base}} + \sum_{i \notin \text{block}} \Delta W_i$. Conditions: full, $-$ first, $-$ mid, $-$ last, and mid-only (mid-third alone). We operationally define a block as *most critical* if its removal produces the largest positive BBQ delta among the three thirds; we report the degree of disruption rather than claiming binary necessity, as the extent of reversal varies across models (range and per-model values in §4.3). The mid-only condition additionally tests whether mid-third alone reproduces the full alignment effect. Each condition is also evaluated on MMLU to address the functional-separation question (§1).

Experiment 3: Cumulative Injection. Layers are injected cumulatively $\hat{W}_k = W_{\text{base}} + \sum_{i=0}^k \Delta W_i$, tracing bias formation with depth; results in §4.5 (Figure 2; full per-model curves in Appendix F).

Experiments 4–5: Direction-Specificity. Sign-flip ($-\Delta W_i$ injection) and norm-matched controls (random Gaussian, same $\|\Delta W_i\|_F$; shuffled ΔW_i)

jointly test whether the mid-third effect is directionally encoded rather than OOD-driven; full results in Appendix G.

3.5 Benchmarks

BBQ (Parrish et al., 2022) is our primary bias benchmark, using the score defined in §3. We additionally evaluate on CrowS-Pairs (Nangia et al., 2020) (stereotyped-pair preference fraction) and StereoSet (Nadeem et al., 2021) (iCAT score) to assess whether the mid-third tendency generalizes across benchmark formats of varying sensitivity (§4.2). For capability, we use MMLU 5-shot (Hendrycks et al., 2021) on a 14-subject domain-stratified subset (4 STEM, 3 Humanities, 4 Social Sciences, 3 Professional; Appendix A) that we selected prior to finalizing the protocol, without reference to injection results. As in prior intervention work (Li et al., 2024; Kolbeinsson et al., 2025), MMLU serves here as a general-utility control. We report ΔMMLU (pp relative to base) as the capability cost of each block-removal condition.

4 Results

4.1 Per-layer Sufficiency Sweep

Figure 1 shows the per-layer L-BBQ score under single-layer ΔW injection for three representative models; Table 2 summarizes the sweep across all 11 instruction-tuned models, of which 9 show a structured per-layer signal with a clear mid-third minimum. In every model with a structured signal, the minimum bias score is achieved by a mid-third layer.

Qwen2.5-0.5B and Qwen2.5-1.5B exhibit a diffuse bias profile rather than clear localization: per-layer bias scores vary minimally across layers (intra-third range ≤ 0.025), with no layer achieving substantial bias reduction. All other 9 models, including Llama-3.2-1B at 1B, show clear mid-third concentration, indicating a soft, architecture-modulated capacity threshold rather than a universal parameter-count effect. Under the null hypothesis that the globally best layer is equally likely to fall in any of the three equal thirds ($p_0 = 1/3$), observing all 9 structured-signal models with a mid-third best layer occurs with probability $p \approx 5.1 \times 10^{-5}$ (exact binomial test), ruling out chance localization.²

²Gemma-2-2B is classified as structured by the BBQ-range criterion (range/base $> 15\%$) despite its norm-mediated suppression profile (§4.5); excluding it gives 8/8, $p = (1/3)^8 \approx 1.5 \times 10^{-4}$, preserving significance.

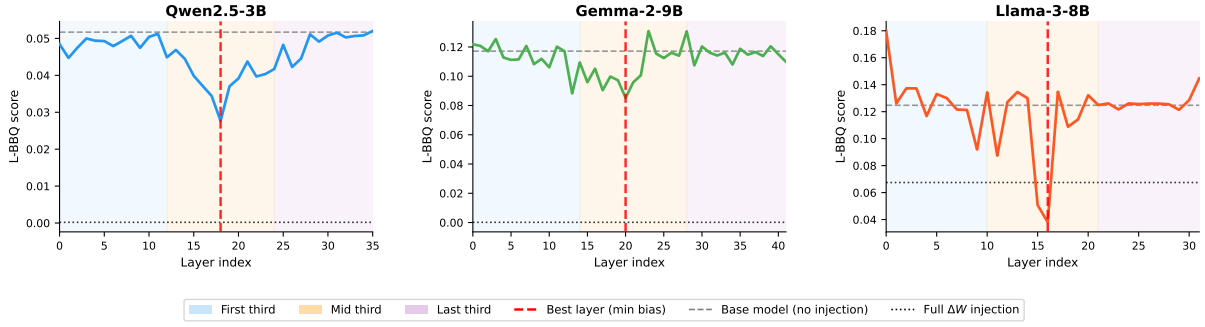


Figure 1: Per-layer L-BBQ score under single-layer ΔW_i injection (three representative models, one per architecture family). Gray dashed = base; black dotted = full ΔW ; red dashed = best layer (single-layer min bias); shaded regions = layer thirds (light blue = first, orange = mid, light purple = last). In all three, the minimum falls within the mid-third. Same pattern holds for all 9 structured-signal models (Appendix C; heatmap in Appendix D).

Model	Best layer	Third	L-BBQ
Qwen2.5-0.5B	L10	mid	0.236 [‡]
Qwen2.5-1.5B	L23	last	0.268 [‡]
Qwen2.5-3B	L18	mid	0.028
Qwen2.5-14B	L21	mid	0.004
Gemma-2-2B	L11	mid	0.125
Gemma-2-9B	L20	mid	0.085
Gemma-3-12B [†]	L18	mid	0.076
Llama-3.2-1B	L9	mid	0.201
Llama-3.2-3B	L10	mid	0.093
Llama-3-8B	L16	mid	0.038
Llama-3.1-8B [†]	L11	mid	0.076

Table 2: Per-layer sufficiency sweep (L-BBQ; lower = less biased). [‡]Flat profile (intra-third range ≤ 0.025); third not bolded; block-removal third dominance may differ (§4.3). [†]9 non-dagger models; base BBQ: Gemma-3-12B 0.095, Llama-3.1-8B 0.118.

Direction specificity. Norm-matched and shuffled controls (Experiment 5) confirm direction-specificity: 4 of 5 evaluated models show Best/NM ratios of $0.54\text{--}0.83\times$ (Table 11); Qwen2.5-0.5B’s near-unity ratio ($0.944\times$) and Gemma-2-2B’s $\Delta_{\text{mid}} = -0.002$ both cluster with diffuse models, consistent with their respective profiles (§4.5).

Category-level robustness. Mid-third dominance holds across 10 of 11 BBQ categories (11/11 near-unanimous for Age, Disability, Physical appearance; 7/11 simple majority for Race/ethnicity and Sexual orientation). The sole exception is SES (first-third dominance in 7 of 11 models); we do not attribute this to a specific mechanism and leave it for further investigation (Appendix H). All subsequent analyses average across all 11 categories.

4.2 Mid-third Tendency Across Bias Benchmarks

The three-benchmark design tests whether the mid-third finding extends beyond BBQ (Table 3). BBQ shows the strongest localization (10/11 at both thirds-level and global-min); StereoSet shows a moderate mid-third plurality (6/11 global min; 6/9 structured-signal models at thirds level); CrowS-Pairs shows the weakest signal (3/11 thirds-level), with per-layer absolute differences below 0.5 bias-score points—too flat for reliable attribution.

	BBQ	CrowS	StereoSet
Thirds-level mid-dominant	10/11	3/11	6/11
Global min, mid-third	10/11	1/11	6/11
Spearman (r) vs. BBQ	—	+0.40 ns	+0.36 ns

Table 3: Mid-third tendency across benchmarks. Thirds-level = fraction of models whose mid-third average is the lowest among three thirds. Global min = fraction with mid-third global best layer. Spearman r vs. BBQ best-layer ranks; ns = $p > 0.05$. CrowS-Pairs per-layer absolute differences < 0.5 points; global-min attribution unreliable.

Best-layer ranks correlate positively with BBQ across both secondary benchmarks ($r = +0.40$ and $+0.36$; non-significant at $n = 11$; Table 3), consistent with a shared localization tendency of varying detectability. We hypothesize that CrowS-Pairs’ flat profile reflects its stereotyped-pair format; intrinsic metrics of this kind are known to track downstream behavior unreliably (Goldfarb-Tarrant et al., 2021), but we do not test the format account directly. These results defend against a BBQ-only interpretation: the mid-third tendency scales with benchmark sensitivity across format-

diverse benchmarks.

4.3 Block Removal: Necessity and Sufficiency

Table 4 reports block removal results for all 11 models (per-model bar charts in Figure 12, Appendix K). In all eleven models, removing the mid-third raises BBQ bias ($\Delta +0.007$ to $+0.215$); the effect is largest in the three active-suppression models ($+0.115$ – $+0.215$). Across all 11 models, Δ_{mid} is the largest among the three thirds in 10 of 11; the sole exception is Qwen2.5-0.5B, whose diffuse profile yields first-third dominance ($\Delta_{\text{first}} = +0.119$ vs. $\Delta_{\text{mid}} = +0.078$), consistent with its diffuse per-layer sweep (§4). Under the null that any third is equally critical ($p_0 = 1/3$), $p \approx 1.3 \times 10^{-4}$ (exact binomial; Llama-3.1-8B borderline, Table 4), establishing mid-third ΔW as necessary for bias suppression.

Active suppression. In 3 of 11 models, post-removal bias exceeds the unaligned base by +42%, +46%, and +15% respectively (Llama-3.2-3B, Llama-3-8B, Gemma-2-2B; computed from Table 5). This empirically satisfies the definition of active suppression (§1): outer-only injection yields higher BBQ than the unaligned base (Table 5), establishing that mid-third ΔW counteracts a latent bias-amplifying contribution from the outer thirds. As a result, transferring only the outer thirds yields a model more biased than the unaligned base, a risk that monolithic ΔW transfer conceals entirely because the outer thirds’ amplifying contribution is masked whenever the mid-third suppressor is also present. Active suppression is not a floor effect: Llama-3.2-1B (base = 0.251) passes the audit despite its high base bias level (Table 5), confirming that the phenomenon is architecture- and alignment-recipe-specific rather than driven by measurement-ceiling constraints.

Mid-third alone is sufficient. Mid-third-only injection (four of the five models evaluated for this comparison, shown in Table 6; Llama-3-8B excluded due to its anomalous MMLU profile, §4.4) achieves BBQ = 0.001–0.020 (89–99% reduction from base) with MMLU +1.7 to +6.2 pp, matching the full task vector from a single third of ΔW . The –mid condition uses twice as many layers as mid-only yet produces the worst bias outcome: what matters is the *presence* of mid-third ΔW , not the degree of partiality.

Model	base	–mid	–first	–last
Qwen2.5-0.5B [‡]	0.252	+0.078	+0.119	–0.003
Qwen2.5-1.5B [‡]	0.273	+0.118	+0.045	+0.004
Qwen2.5-3B	0.050	+0.019	<0.001	<0.001
Qwen2.5-14B	0.013	+0.007	<0.001	<0.001
Gemma-2-2B	0.160	+0.169*	–0.006	<0.001
Gemma-2-9B	0.116	+0.057	+0.001	<0.001
Gemma-3-12B	0.095	+0.069	+0.020	<0.001
Llama-3.2-1B	0.251	+0.111	+0.014	+0.013
Llama-3.2-3B	0.176	+0.215*	–0.011	–0.002
Llama-3-8B	0.125	+0.115*	–0.011	–0.005
Llama-3.1-8B	0.118	+0.030 [§]	+0.028	–0.001

Table 4: Block removal results, all 11 models. base: un-injected BBQ; –mid/–first/–last: ΔL -BBQ from full ΔW (positive = more biased after removing that third). *post-removal BBQ exceeds base. [‡]diffuse-signal. [§]borderline; robust without it ($p \approx 3.6 \times 10^{-4}$, 9/10).

Layer-selective audit. The *layer-selective auditing protocol* evaluates (a) full ΔW , (b) outer-only (–mid), and (c) mid-only injection; condition (b) bias above the unaligned base flags a latent amplifying component concealed by monolithic evaluation (Table 5).

4.4 BBQ–MMLU Dissociation

Block removal shows that mid-third ΔW is necessary for bias suppression; we next examine whether this comes at a capability cost. Table 6 (scatter in Appendix I) confirms that bias and capability occupy functionally distinct layer ranges within ΔW . The BBQ-side dissociation holds across all four models: ΔL -BBQ from –mid is large and positive ($+0.019$ – $+0.215$), whereas ΔL -BBQ from –first is near-zero or negative ($|\Delta| \leq 0.011$). MMLU-side separation (first-third removal incurring larger MMLU cost than mid-third) holds in Gemma models (Gemma-2-9B: $8.97\times$; Gemma-2-2B: $1.85\times$) but not in Llama-3.2-3B or Qwen2.5-3B (first-to-mid ratios $0.21\times/0.36\times$). Last-third removal has negligible effect on both axes, showing that the last third of ΔW carries neither bias-suppression nor capability-critical information.

Bidirectional replication. The dissociation holds in the *injection direction*: mid-third-only injection achieves the BBQ reduction in §4.3 with minimal MMLU cost; first-third-only leaves bias unchanged while recovering up to +31.8 pp

Model	base	Full ΔW	Outer-only	Audit
Qwen2.5-0.5B	0.252	0.093	0.171	Pass
Qwen2.5-1.5B	0.273	0.134	0.252	Pass
Qwen2.5-3B	0.050	<0.001	0.019	Pass
Qwen2.5-14B	0.013	<0.001	0.007	Pass
Gemma-2-2B	0.160	0.016	0.184	Fail
Gemma-2-9B	0.116	<0.001	0.058	Pass
Gemma-3-12B	0.095	0.003	0.072	Pass
Llama-3.2-1B	0.251	0.093	0.204	Pass
Llama-3.2-3B	0.176	0.035	0.249	Fail
Llama-3-8B	0.125	0.067	0.182	Fail
Llama-3.1-8B	0.118	0.037	0.068	Pass

Table 5: Layer-selective audit, all 11 models. **base**: un-injected BBQ. **Full ΔW** : all-layer injection. **Outer-only**: first+last thirds (–mid). **Audit**: Fail if outer-only BBQ > base (latent amplifier present); Pass otherwise. 3 of 11 fail.

Model	Cond.	ΔL -BBQ	$\Delta MMLU$
Qwen2.5-3B	–mid	+0.019 $\uparrow$	–0.008
	–first	+0.000	+0.003
Gemma-2-2B	–mid	+0.169 $\uparrow\uparrow$	–0.067
	–first	–0.006	–0.124
Gemma-2-9B	–mid	+0.057 $\uparrow$	–0.039
	–first	+0.001	–0.350
Llama-3.2-3B	–mid	+0.215 $\uparrow\uparrow$	–0.048
	–first	–0.011	–0.010

Table 6: BBQ–MMLU dissociation (Llama-3-8B excluded; §4.4). ΔL -BBQ/ $\Delta MMLU$: change from full ΔW . $\uparrow/\uparrow\uparrow$: ΔL -BBQ +0.01–0.10/> +0.10. –mid is bias-critical across all four; –first is capability-critical in Gemma.

MMLU (Gemma-2-9B: –35.0 pp lost under –first, +31.8 pp recovered under first-only; within 3.2 pp of full recovery), replicating the separation.

Llama-3-8B anomaly. Llama-3-8B’s full ΔW injection reduces MMLU by –35.8 pp (vs. ≤ 5 pp for all others) and removing its first-third recovers MMLU (+5.0 pp), the inverse of every other model. We attribute this to an unusually large capability gap between the Llama-3-8B base and instruct variants: the first-third ΔW encodes a net-negative MMLU transfer for this pair, likely reflecting idiosyncrasies in its instruction-tuning data mix, not a departure from the general pattern. Its BBQ block removal (Table 4) follows the same mid-third pattern, confirming the anomaly is MMLU-specific; a parallel anomaly appears in its sign-flip response (§4.5).

4.5 Convergent Evidence and Scale Analysis

Scale and architecture jointly determine structured localization (Figure 11, Appendix J): among the 9 structured-signal models, the best layer achieves $0.7\text{--}3.0\times$ lower bias than the next-best third’s average. The two exceptions—Qwen2.5-0.5B (range: 0.025) and Qwen2.5-1.5B (range: 0.007)—show minimal per-layer variation with no mid-third minimum. Llama-3.2-1B (1B) exhibits clear mid-third signal (range: 0.058; best layer 15.6% below mean). We observe, in our data, a soft and architecture-dependent scale threshold for localization: Llama localizes below 1B, while Qwen requires between 1.5B and 3B. This is reminiscent of scale-dependent behavior reported more broadly (Wei et al., 2022), though apparent thresholds can also arise from the choice of metric (Schaeffer et al., 2023). Gemma-2-2B (2B) achieves positional mid-third localization yet remains direction-ambiguous in the specificity control (ratio $0.779\times$; Table 11), suggesting the threshold governs *directional* encoding specifically (see §5).

Convergent evidence: cumulative injection. As an independent check, we inject layers cumulatively ($L_0 \rightarrow L_k$) and measure BBQ bias at each prefix (Figure 2). The Spearman correlation between each layer’s marginal gain and its per-layer sweep score is negative in all 8 models with available data (Gemma-3-12B, Llama-3.1-8B, and Llama-3.2-1B excluded: cumulative sweep incomplete at submission); 6 of 8 show moderate-to-strong correlations ($\rho \leq -0.40$, all 6 significant at $p < 0.05$; Table 10, Appendix F); the all-negative sign pattern has probability $(1/2)^8 \approx 0.004$ (one-sided sign test), providing independent methodological triangulation of the localization result. The dominant marginal-gain third is mid in 5 of 6 structured-signal models, consistent with a localized suppression hub. The exception (Gemma-2-9B, first-dominant) reflects its early amplification phase before mid-third suppression takes effect (Figure 6); its sufficiency and necessity profiles both still identify the mid-third as primary.

Weight-Space Dissociation. We ask whether static weight-space statistics predict the mid-third behavioral minimum. Across nine models (excluding Gemma-3-12B and Llama-3.1-8B which lack fully-decomposed ΔW tensors), three statistics (Frobenius norm $\|\Delta W_i\|_F$, sign coherence (Yadav et al., 2023), and participation ratio) all predict

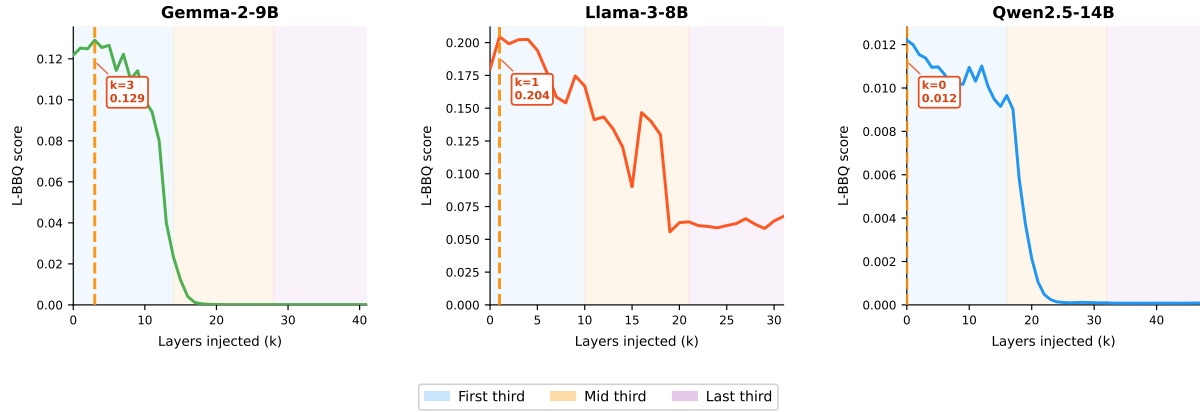


Figure 2: Cumulative ΔW injection ($\hat{W}_k = W_{\text{base}} + \sum_{i=0}^k \Delta W_i$) for three models. Shaded: blue = first third, orange = mid, purple = last; annotation = peak bias layer. Early layers raise bias *above* the base before mid-third suppresses it; Qwen2.5-14B shows suppression without the amplification phase.

last-third dominance: norm and PR peak in the last third, sign coherence is uniformly near-zero (median ≈ 0.0001).

The mid-third behavioral minimum is therefore not explainable by any static weight statistic, confirming that behavioral localization requires interventional analysis. One exception: Gemma-2-2B’s Frobenius norm correlates *negatively* with behavioral suppression ($\rho = -0.54$, $p < 0.01$); consistent with its norm-mediated (direction-ambiguous) profile; larger norms incidentally reduce bias without direction-specific encoding (see §5).

Sign-flip: Directional Encoding Confirmed. Across five models in the paired-bar comparison (Figure 13, Appendix L; Llama-3-8B excluded due to its sign-flip anomaly, consistent with its MMLU anomaly in §4.4), injecting $+\Delta W_i$ suppresses bias below baseline in all five; $-\Delta W_i$ raises bias above baseline in four of five (Llama-3.2-1B: near-neutral, $\Delta = -0.007$). The direction predominantly determines the behavioral sign, confirming the mid-third effect is directionally encoded, not an OOD-disruption artifact. Per-layer reversal rates, computed across all eight evaluated models (the five above plus Qwen2.5-14B, Gemma-2-9B, and Llama-3-8B for which per-layer data are available), range from 38% (Gemma-2-2B) to 100% (Qwen2.5), with the mid-third as the most symmetric region in 5 of 8 models (Table 9, Appendix E).

Convergent summary. Five designs support mid-third localization (10 of 11 models); block removal also reveals functional separation across thirds and active suppression in 3 of 11 models.

5 Discussion

5.1 Mid-third Layers as the Primary Interventional Locus

Convergent evidence in §4 establishes that ΔW is spatially organized along a functional axis: mid-third layers are the primary interventional locus of bias suppression, first-third layers carry capability-critical structure in Gemma, and the outer thirds are potentially bias-amplifying when the mid-third suppressor is absent (§4.3)—an organization invisible to static weight statistics.

Anomalous models as boundary probes. The two diffuse-signal models (Qwen2.5-0.5B, Qwen2.5-1.5B) probe capacity limits rather than contradicting mid-third localization: diffuse per-layer profiles signal insufficient ΔW capacity for structured alignment encoding (§4.5). Gemma-2-2B is a structured-signal model whose behavioral minimum is norm-mediated rather than direction-specific (§4.5): positional localization can emerge at 2B, but directional encoding cannot—suggesting the capacity threshold marks the onset of *directional* encoding specifically, not positional localization alone. Together these three models demarcate the capacity-and-geometry boundary within which mid-third localization holds. Sufficiency and necessity converge on the mid-third in all 9 structured-signal models; divergence in diffuse-signal models reflects noise-level variation in the absence of a localized signal, not a dissociation between the two measures.

Why the mid-third? Our data establish *where* bias suppression is encoded but not *why*. Mid-

third layers may occupy a representational inflection point (Tenney et al., 2019): transformer FFN layers in this range act as key-value memories that promote semantic associations (Geva et al., 2021), and causal editing work localizes factual knowledge to the same depth range (Meng et al., 2022). A plausible hypothesis is that alignment fine-tuning modifies this mid-layer semantic substrate to suppress stereotyped associations; circuit-level validation (Elhage et al., 2021) remains open.

5.2 Implications of Active Suppression

The active-suppression result (§4.3) carries a direct deployment risk: outer-only ΔW transfer can yield a model more biased than the unaligned base. As Table 5 shows, 3 of 11 models formally fail the layer-selective audit, confirming that the mid-third suppressor is the only component preventing bias from exceeding the unaligned base. This risk is invisible under monolithic ΔW evaluation but can surface in model merging when two ΔW s interfere in the mid-third range.

5.3 Broader Implications

Beyond auditing—which cannot be substituted by full- ΔW evaluation alone—these results carry implications for layer-selection methods more broadly. Structured pruning methods score layer importance by a single generic criterion; ShortGPT, for instance, ranks layers by Block Influence, computed as one minus the average cosine similarity between a block’s input and output representations (Men et al., 2025). Our BBQ–MMLU dissociation shows an orthogonal axis invisible to such criteria. Across the four models with full dissociation data (Table 6), first-third removal yields $|\Delta L\text{-BBQ}| \leq 0.011$ while causing an MMLU loss of up to 35.0 pp (Gemma-2-9B); mid-third removal reverses debiasing with at most 6.7 pp of MMLU loss. A generic capability- or redundancy-oriented score would protect the first-third for capability but silently amplify bias by discarding the mid-third. Conversely, mid-third-only injection enables targeted debiasing without modifying capability-critical layers. More broadly, methods for task vector composition resolve *parameter* interference (Yadav et al., 2023); our results indicate that doing so does not by itself preserve alignment-induced behavior, so behavior-aware layer auditing is additionally needed for safe composition: the auditing protocol (§5.2) extends naturally to other alignment-induced components such as refusal sup-

pression (Arditi et al., 2024), safety steering, or factual calibration.

6 Conclusion

We have shown that ΔW ’s behavioral components are spatially organized: the mid-third of transformer layers is the consistent interventional hub of BBQ-measured bias suppression (block removal: 10 of 11 models), functionally separable from the first-third, which carries capability-critical structure. Weight-space statistics predict last-third dominance, so behavioral geometry is invisible without interventional analysis. Our auditing protocol flags 3 of 11 models as bias-amplifying: in these, mid-third layers actively counteract a bias-amplifying outer-third contribution, so partial ΔW transfer can yield a model more biased than the unaligned base—a deployment risk invisible to monolithic evaluation. These findings provide an empirical roadmap for layer-selective ΔW editing and safe task vector composition.

Limitations

Generalization scope. The main controlled results are restricted to 0.5B–14B models from the Qwen, Gemma, and Llama families. Appendix M reports a separate extension to four additional same-release pairs, including Mistral and 24–32B models; these were not part of the primary 11-model controlled set, and generalization to other families remains limited. Whether analogous localization holds for other alignment-induced *behaviors* (refusal, toxicity, factual calibration) remains untested. **Scale boundary.** Our scaling analysis (§4.5) identifies a capacity threshold below which structured localization fails (Qwen2.5 ≤ 1.5 B), but does not characterize an upper boundary. At frontier scale (≥ 70 B), two competing predictions are consistent with our data: (i) further sharpening of mid-third localization, or (ii) a shift toward distributed or norm-mediated encoding analogous to the Gemma-2-2B profile (§4.5). Our capacity-threshold framing predicts that the directional-encoding regime may itself have an upper scale boundary, but this is not empirically tested. **Benchmark scope.** BBQ is a QA-format benchmark; the mid-third tendency extends to StereoSet (6/11) and weakly to CrowS-Pairs (3/11 thirds-level, flat signal), but generalization to alternative bias *metrics* (toxicity scores, factual error rates) or formats beyond stereotyping benchmarks remains untested. **Interventional**

scope. Single-layer injection assumes cross-layer interactions are small. The Qwen sign-flip symmetry result partially supports this; cumulative injection (§4.5) partially captures multi-layer effects. **Compute.** All experiments were conducted on NVIDIA A100 GPUs; runtime scales with model size (0.5B–14B) and is detailed in the accompanying repository. **Scoring.** The main 11-model analysis uses L-BBQ (§3.2) with the original string-based response mapper, which left a condition-dependent subset of predictions unassigned and excluded from the score normalization. Metadata-aware, polarity-aware rescoring preserved the block-level necessity and sufficiency conclusions for three representative model families, although the absolute score magnitudes changed. This robustness check was limited to block-level conditions and does not establish preservation of the complete per-layer profiles across all 11 models (Appendix Q). **Mechanistic scope.** Our protocol identifies the interventional locus of bias-encoding ΔW but does not constitute a circuit-level mechanistic account (Elhage et al., 2021). Probing geometry (Belinkov, 2022) of ΔW across training checkpoints remains unexplored, and the speculative “Why the mid-third?” account in §5 is untested.

Ethical Considerations

This study uses only publicly available models and benchmarks; no new datasets were collected and no human subjects were involved. We acknowledge a dual-use risk: the same layer-wise injection protocol could in principle be used to remove bias-suppressing components rather than preserve them; our intent is strictly defensive: to identify where alignment’s bias-suppression effect resides so it can be monitored and preserved under model editing or merging.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2026S1A5A8011475).

References

Andy Arditi, Oscar Obeso, Aaqib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of ACL*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of ACL*.

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*.

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

Gemma Team. 2025. [Gemma 3 technical report](#). *arXiv preprint arXiv:2503.19786*.

Gemma Team. 2026. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of EMNLP*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of EMNLP*.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. [Intrinsic bias metrics do not correlate with application bias](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940.

Hila Gonen and Yoav Goldberg. 2019. [Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of ICLR*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hananeh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *Proceedings of ICLR*.

- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Arinbj  rn Kolbeinsson, Kyle O’Brien, Tianjin Huang, Shanghua Gao, Shiwei Liu, Jonathan Richard Schwarz, Anurag Jayant Vaidya, Faisal Mahmood, Marinka Zitnik, Tianlong Chen, and Thomas Hartvigsen. 2025. [Composable interventions for language models](#). In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, and 27 others. 2024. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pages 28525–28550. PMLR.
- Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. 2025. [Safety layers in aligned large language models: The key to LLM security](#). In *Proceedings of ICLR*.
- Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of ICML*.
- Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2025. [ShortGPT: Layers in large language models are more redundant than you expect](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20192–20204.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Proceedings of NeurIPS*.
- Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *Proceedings of ICLR*.
- Meta AI. 2024a. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Meta AI. 2024b. Llama 3.2 model cards and prompt formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/. Official model card for the Llama 3.2 1B and 3B models.
- Mistral AI. 2025. Mistral-small-24b-instruct-2501 model card. <https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501>. Instruction-tuned from Mistral-Small-24B-Base-2501.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of ACL-IJCNLP*.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of EMNLP*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of NeurIPS*.
- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. [BBQ: A hand-built bias benchmark for question answering](#). In *Findings of ACL*, pages 2086–2105.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of ACL*.
- Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Pratyusha Sharma, Jordan T Ash, and Dipendra Misra. 2024. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. In *International Conference on Learning Representations*.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL*.
- Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. [Function vectors in large language models](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- 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 *Proceedings of NeurIPS*.

- Franz Wegner. 1980. [Inverse participation ratio in \$2+\epsilon\$ dimensions](#). *Zeitschrift für Physik B Condensed Matter*, 36(3):209–214.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-merging: Resolving interference when merging models. In *Proceedings of NeurIPS*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL*.
- Zheng Zhao, Yftah Ziser, and Shay B. Cohen. 2024. [Layer by layer: Uncovering where multi-task learning happens in instruction-tuned large language models](#). In *Proceedings of EMNLP*, pages 15195–15214.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A MMLU 14-Subject Evaluation Subset

Table 7 lists the 14 subjects used in all MMLU capability evaluations. The subset was fixed before observing any injection results and was selected to achieve domain-stratified coverage. Prior intervention work likewise uses MMLU as a general-utility control (Li et al., 2024; Kolbeinsson et al., 2025), though with different subject selections.

Domain	Subject
STEM (4)	Abstract Algebra
	High School Mathematics
	High School Physics
	High School Computer Science
Humanities (3)	Philosophy
	High School World History
	Moral Scenarios
Social Sciences (4)	Sociology
	High School Psychology
	High School Government and Politics
	High School Statistics
Professional (3)	Professional Law
	Medical Genetics
	Global Facts

Table 7: The 14 MMLU subjects used for capability evaluation in all block-removal and block-only injection experiments.

B BBQ Social-Group Categories

BBQ (Parrish et al., 2022) covers 11 social-group categories. All bias scores in this paper are computed jointly over all 11 categories unless noted otherwise; category-level results are discussed in §4.

#	Category
1	Age
2	Disability status
3	Gender identity
4	Nationality
5	Physical appearance
6	Race/ethnicity
7	Race $\times$ SES
8	Race $\times$ gender
9	Religion
10	Socioeconomic Status (SES)
11	Sexual orientation

Table 8: The 11 social-group categories in BBQ (Parrish et al., 2022). Mid-third localization holds across 10 of 11 categories; SES (category 10) shows first-third dominance in 7 of 11 models.

C Per-Model Sufficiency Curves

Per-layer L-BBQ score under single-layer ΔW_i injection for all nine structured-signal models; complements the three-model summary in Figure 1.

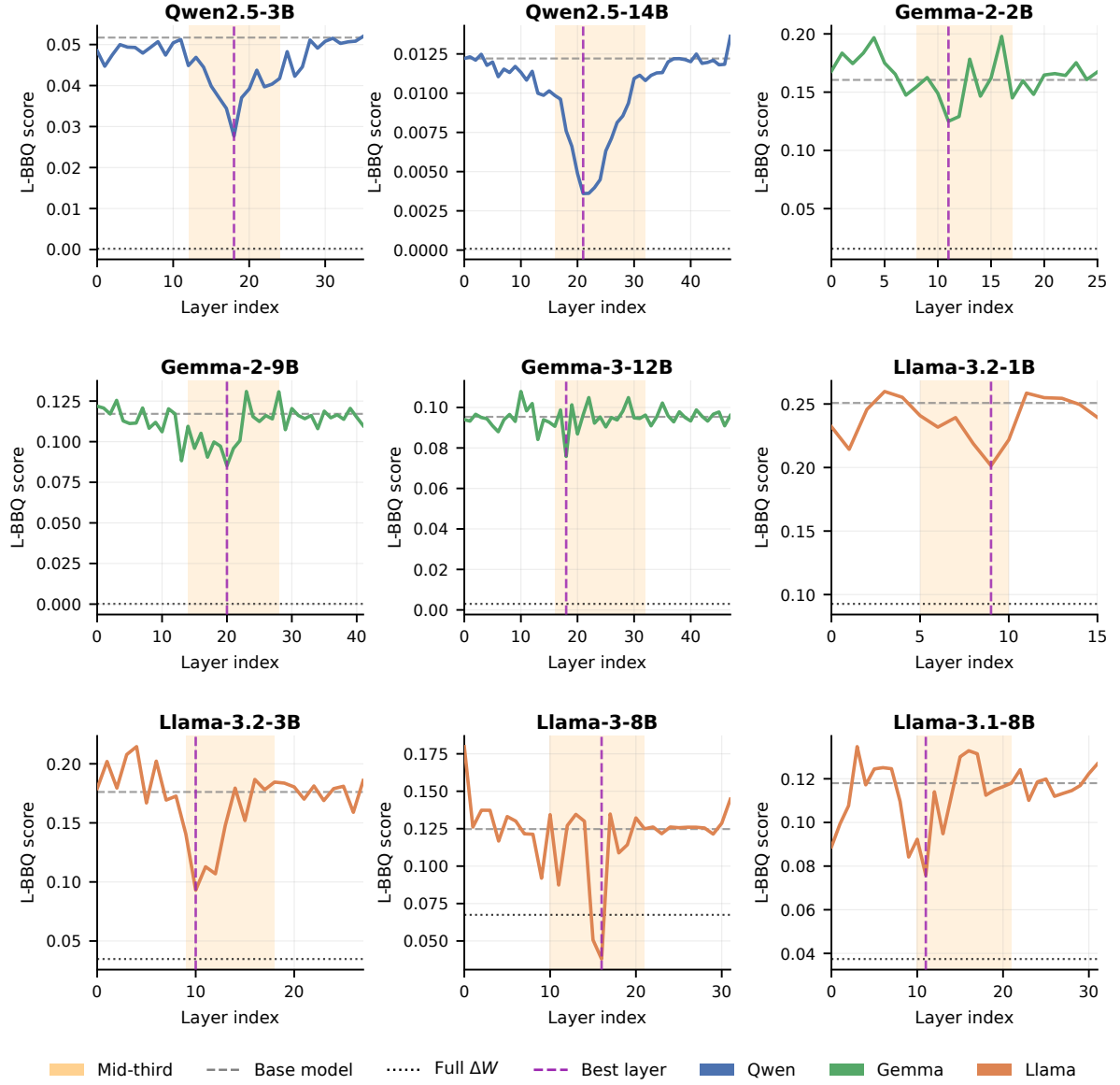
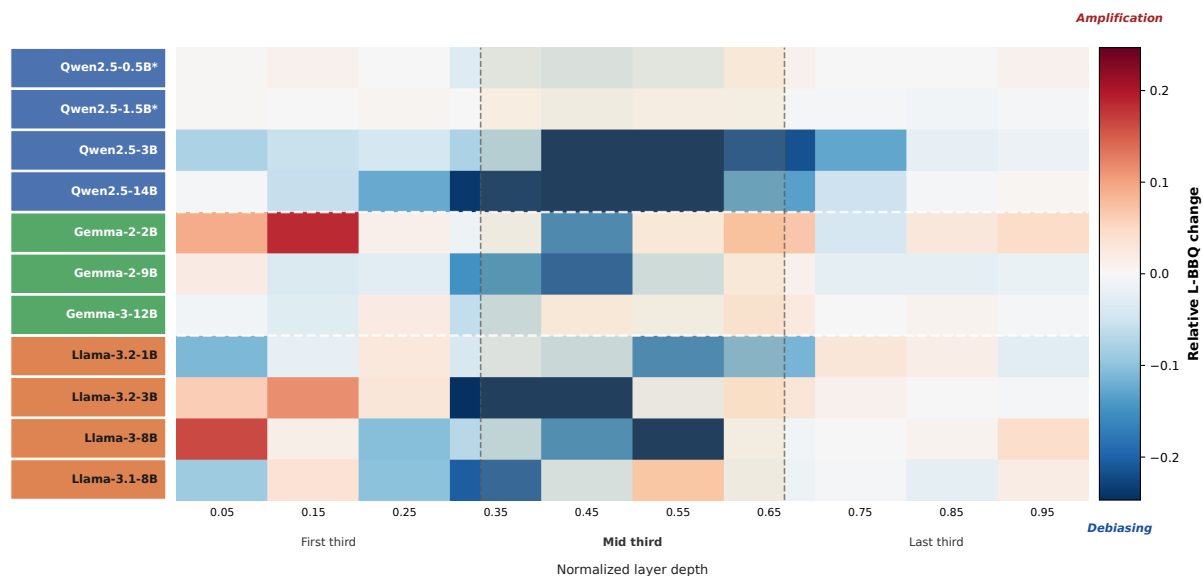


Figure 3: Per-layer L-BBQ score under single-layer ΔW_i injection for all nine structured-signal models. Each panel shows the bias score as layer index varies; mid-third is shaded (orange). Dashed gray = base-model bias; dotted black = full- ΔW instruct model; dashed purple vertical line = best layer (minimum bias). Family color: Qwen (blue), Gemma (green), Llama (orange).

D Per-Layer Relative Bias Change Heatmap

Heatmap view of per-layer bias scores across all 11 models; rows = models, columns = layers.



* diffuse signal (no structured localization)

Figure 4: Relative bias change ($\text{bias}_\ell / \text{bias}_{\text{base}} - 1$) across all 11 models, binned into ten normalized-depth intervals. Blue = debiasing; red = amplification; dashed lines = mid-third boundaries (0.33, 0.67). (*) diffuse-signal models (Qwen2.5-0.5B, Qwen2.5-1.5B).

E Sign-flip Symmetry

Model	Mid rev.	Most-sym.
Qwen2.5-3B	12/12 (100%)	mid
Qwen2.5-14B	16/16 (100%)	mid
Gemma-2-9B	12/14 (86%)	mid
Gemma-2-2B	3/8 (38%)	first
Llama-3.2-3B	7/9 (78%)	mid
Llama-3.1-8B	8/11 (73%)	mid
Llama-3-8B	6/10 (60%)	last
Llama-3.2-1B	2/5 (40%)	last

Table 9: Sign-flip symmetry across 8 of 11 models with sufficient per-layer signal. **Mid rev.:** fraction of mid-third layers where $-\Delta W_i$ raises bias above baseline when $+\Delta W_i$ suppresses it (i.e., most symmetric sign-flip response). **Most-sym.:** third with the highest reversal rate. **Bold** = 100%. Higher rates indicate stronger directional encoding of bias suppression.

Per-layer profiles are shown below.

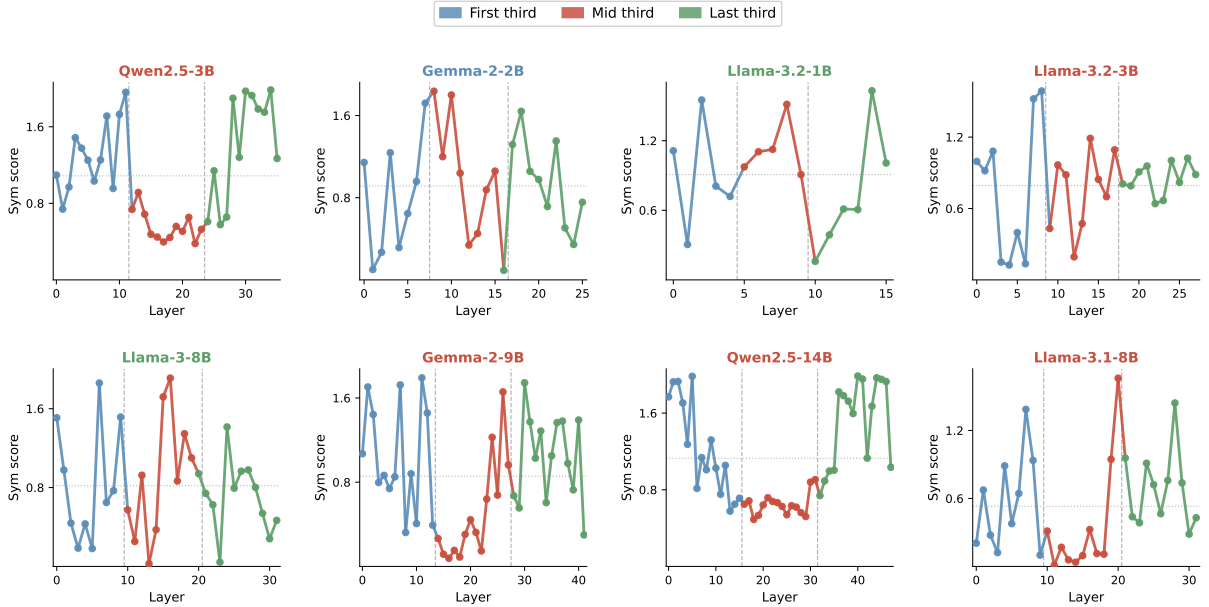


Figure 5: Per-layer sign-flip symmetry score across eight representative models. Lower = more symmetric ($+\Delta W_i$ and $-\Delta W_i$ produce mirror-image effects). Mid-third layers (shaded) show lowest symmetry scores in Qwen models, confirming directional encoding. Dashed line = per-model mean.

F Cumulative Injection Curves

Statistical analysis and Figure 2 are discussed in the main text (§4.5). Figure 6 shows the full cumulative injection curves for all eight models with valid data. Table 10 reports the per-model Spearman correlations.

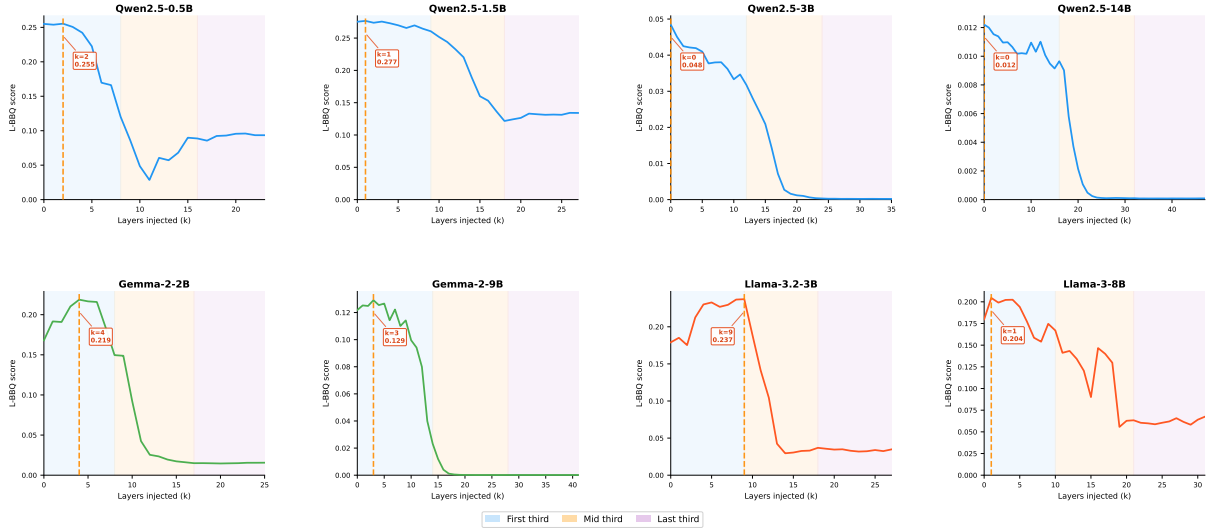


Figure 6: Cumulative ΔW injection curves for all eight evaluated models ($\hat{W}_k = W_{\text{base}} + \sum_{i=0}^k \Delta W_i$). Orange dashed vertical line = peak bias layer (maximum across prefix lengths). Shaded regions: blue = first third, orange = mid-third, purple = last third. The three large-scale models shown in the main text Figure 2 (Gemma-2-9B, Llama-3-8B, Qwen2.5-14B) are included here for completeness.

Model	ρ	p	$p < 0.05$	Dom. third
Qwen2.5-0.5B	-0.304	0.158		first
Qwen2.5-1.5B	-0.412	0.033	✓	mid
Qwen2.5-3B	-0.664	<0.001	✓	mid
Qwen2.5-14B	-0.399	0.006	✓	mid
Gemma-2-2B	-0.634	<0.001	✓	mid
Gemma-2-9B	-0.502	<0.001	✓	first
Llama-3.2-3B	-0.513	0.006	✓	mid
Llama-3-8B	-0.096	0.609		mid

Table 10: Spearman ρ between per-layer marginal cumulative gain and per-layer sweep score (L-BBQ) for each model. **Bold:** $\rho \leq -0.40$, $p < 0.05$ (moderate-to-strong structured signal; 6 of 8). Gemma-3-12B, Llama-3.1-8B, and Llama-3.2-1B lacked sufficient cumulative data and are excluded.

G Direction-Specificity Control Results

Full results for Experiment 5 (norm-matched and shuffled controls); summarized in §4.5.

Model	Best	NM	Shuffled	Ratio
Qwen2.5-0.5B	0.236	0.250	0.252	0.944×
Qwen2.5-3B	0.028	0.051	0.051	0.548 ×
Gemma-2-2B	0.125	0.161	0.161	0.779×
Llama-3.2-1B	0.201	0.244	0.239	0.827×
Llama-3.2-3B	0.093	0.174	0.176	0.535 ×

Table 11: Direction-specificity controls. **NM** = norm-matched Gaussian (30 seeds); **Ratio** = Best/NM ($\ll 1$ = direction-specific). **Bold**: ratio $\leq 0.55\times$.

H Category-Level Localization Heatmap

Mid-third dominance across all 11 BBQ social-group categories; referenced in §4.1.

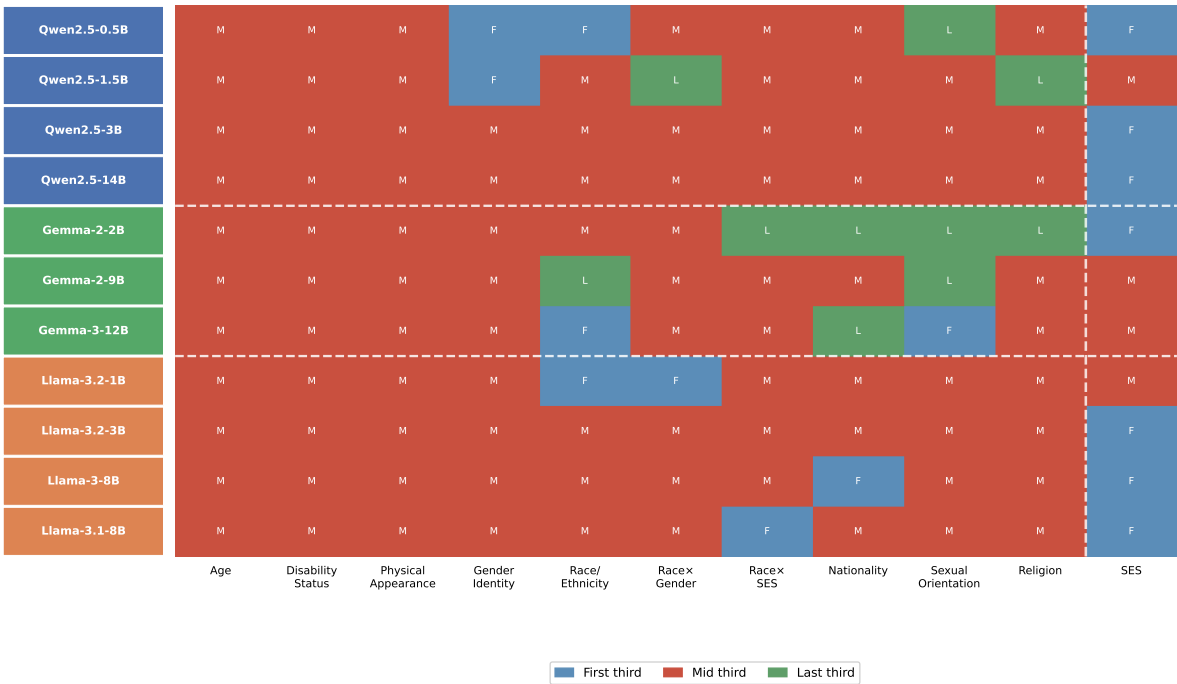


Figure 7: BBQ category-level localization heatmap. Each cell shows which third (first/mid/last) achieves the lowest bias score for a given model–category pair. Mid-third dominance (red) holds across 10 of 11 BBQ categories in 9 of 11 models; Socioeconomic Status (SES) is the exception, showing first-third dominance (7 of 11 models). We report SES as a documented exception without attributing it to a specific mechanism (§4).

I BBQ–MMLU Dissociation: Visualizations

Three complementary visualizations of the BBQ–MMLU dissociation result (§4.4).

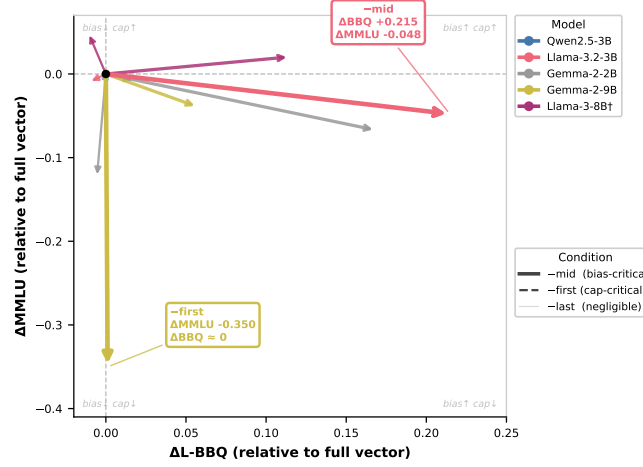


Figure 8: Bias-capability dissociation arrow plot. Each arrow originates from the full-vector baseline (0, 0) and ends at the ($\Delta L\text{-BBQ}$, $\Delta MMLU$) coordinate of the ablation condition. Mid-third removal arrows (rightward, $->$) shift primarily along the bias axis with negligible MMLU cost; first-third removal arrows (downward, $-|>$) show the largest MMLU losses in the Gemma-2 models with near-zero bias change; this capability-side pattern is not universal across the displayed models. Two hero trajectories anchor the dissociation: Llama-3.2-3B $-mid$ ($\Delta L\text{-BBQ} = +0.215$, $\Delta MMLU = -0.048$) and Gemma-2-9B $-first$ ($\Delta MMLU = -0.350$, $\Delta L\text{-BBQ} \approx 0$). Last-third removal (faded) is negligible in both dimensions. [†]Llama-3-8B included; MMLU anomaly noted in §4.4.

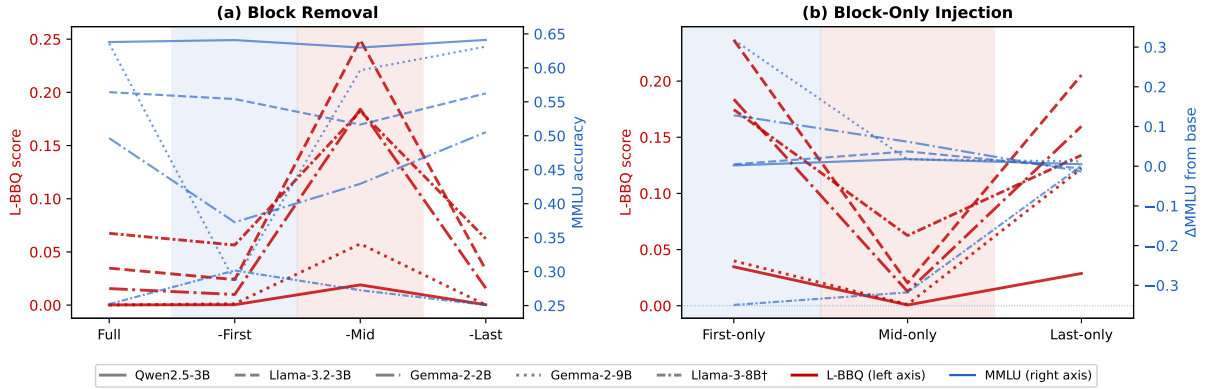


Figure 9: BBQ–MMLU double dissociation shown as line curves across block conditions. **(a) Block Removal:** red lines (left axis) = L-BBQ score; blue lines (right axis) = MMLU accuracy. Red shading marks $-Mid$ (BBQ spikes across all models); blue shading marks $-First$ (largest MMLU losses in the Gemma-2 models; not universal across the displayed models). **(b) Block-Only Injection:** red shading marks Mid-only (lowest BBQ); blue shading marks First-only (highest MMLU recovery). The two shaded regions fall on different conditions in both panels, confirming functional separation of bias-suppression and capability-encoding layer ranges.

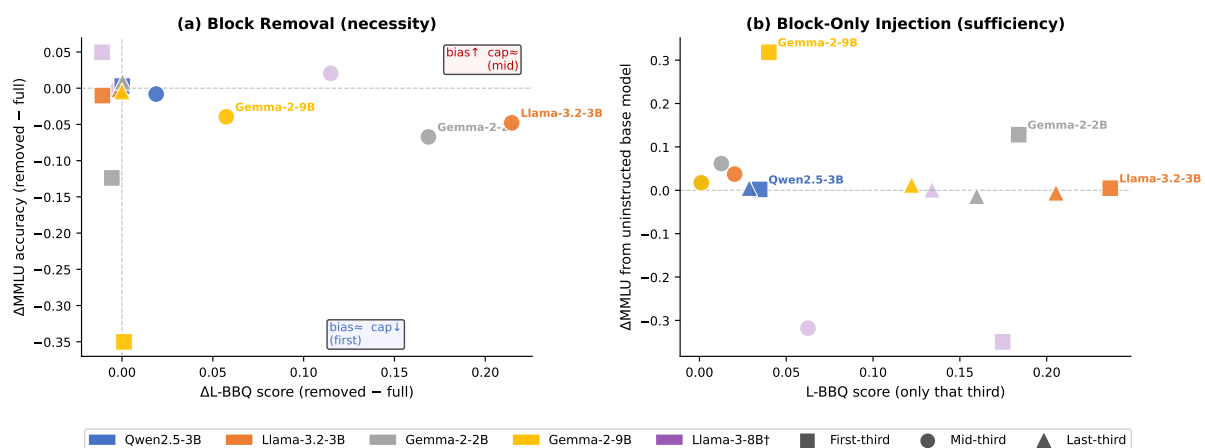


Figure 10: BBQ-MMLU double dissociation (scatter). Each point is a block-removal or block-only condition. Mid-third removal moves right (more bias) with minimal MMLU cost; first-third removal causes the largest MMLU losses in the Gemma-2 models with minimal BBQ change.

J Scale Effect on Best-Layer Localization

Per-layer L-BBQ heatmap for the Qwen model family, illustrating the capacity-threshold analysis in §4.5.

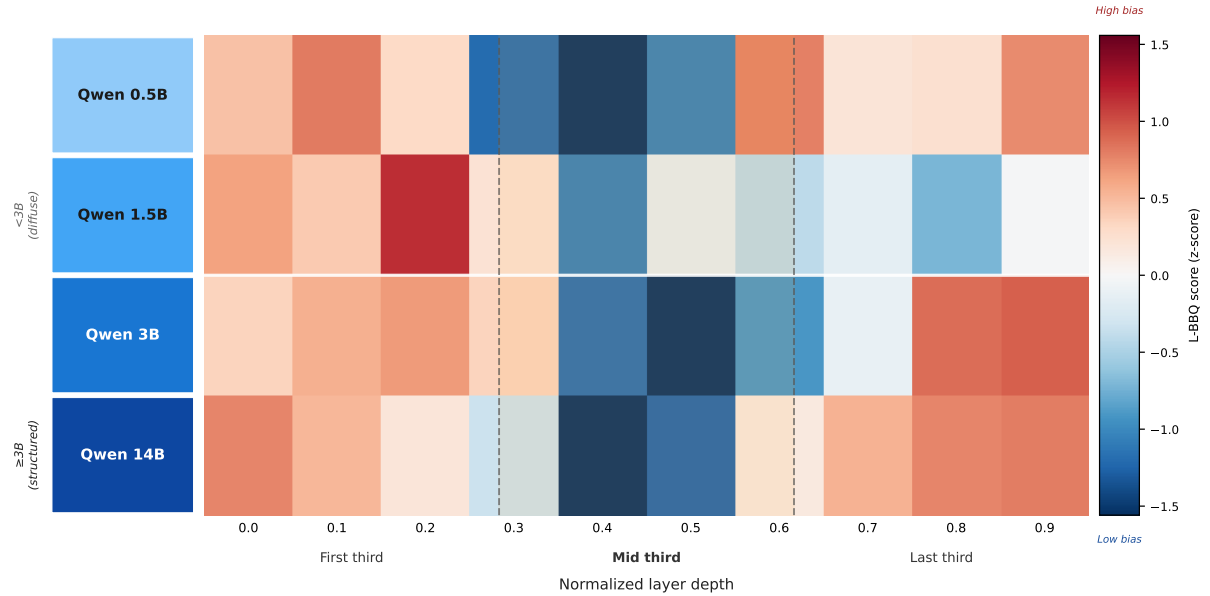


Figure 11: Per-layer L-BBQ z-score heatmap for the four Qwen model sizes. Each row is one model; columns are ten equal-width bins of normalized layer depth. Color encodes the z-scored per-layer bias under ΔW_i injection (blue = below-average bias, debiasing; red = above-average bias, amplifying). Diffuse-signal models (Qwen2.5-0.5B, Qwen2.5-1.5B; top) show no consistent mid-third localization; structured-signal models (Qwen2.5-3B, Qwen2.5-14B; bottom) show a clear debiasing band in the mid-third (0.33–0.67), consistent with a soft, architecture-modulated capacity threshold between 1.5B and 3B (§4.5).

K Necessity Analysis: Block Removal

Per-model block removal results (BBQ and MMLU) for five representative models with full BBQ+MMLU data; full 11-model BBQ results in Table 4.

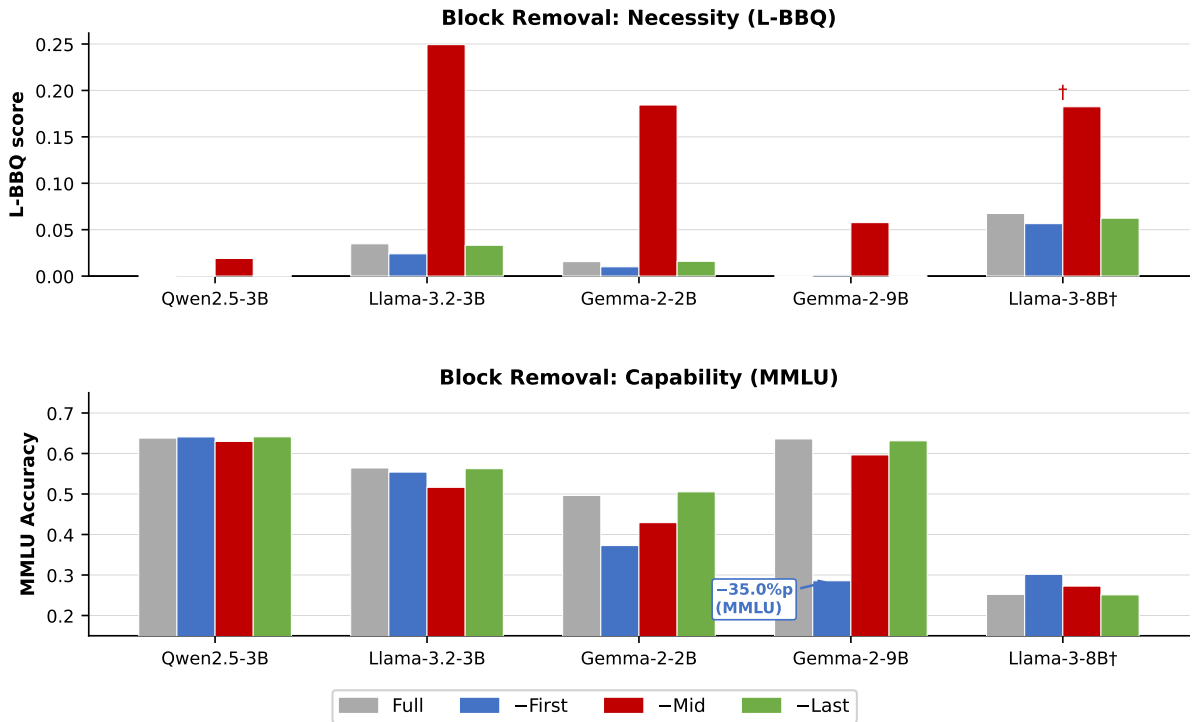


Figure 12: Block removal necessity experiment across five representative models. **Top:** L-BBQ score after omitting each third’s ΔW contribution (base weights retained, $\hat{W} = W_{\text{base}} + \sum_{i \notin \text{block}} \Delta W_i$; §3). Removing the mid-third consistently produces the largest increase in bias score, confirming that mid-third layers are *necessary* for bias suppression. **Bottom:** Corresponding MMLU accuracy under the same removal conditions. General-knowledge performance degrades modestly under mid-third removal but remains well above chance, indicating that the mid-third bias role is not merely a proxy for overall model capability.

L Sign-flip Symmetry: Paired Bar Comparison

Paired bar view of $+\Delta W_i$ vs. $-\Delta W_i$ injection across five models (Llama-3-8B excluded; its sign-flip result inverts unexpectedly, consistent with the MMLU anomaly in §4.4); complements the per-layer scatter in Appendix E.

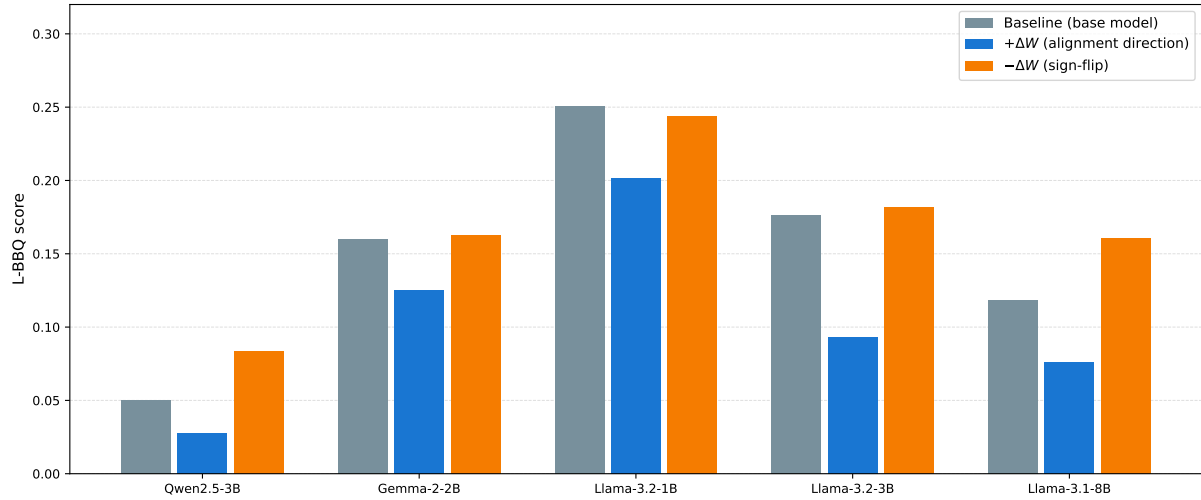


Figure 13: Sign-flip symmetry: paired bars across five models evaluated at each model’s best injection layer (L-BBQ score averaged across all 11 social-group categories). **Gray** = Baseline (base model); **Blue** = $+\Delta W$ (alignment direction, suppresses bias); **Orange** = $-\Delta W$ (sign-flip). $+\Delta W$ reduces bias below baseline in all five models. $-\Delta W$ raises bias above baseline in four of five; Llama-3.2-1B shows a near-neutral sign-flip response ($\Delta = -0.007$). Llama-3-8B excluded: its sign-flip anomaly (sign-flip < alignment < base) is consistent with its MMLU-anomaly profile (§4.4).

Model	Removal				Injection		
	full	—first	—mid	—last	only-first	only-mid	only-last
Mistral-7B	0.0869	0.0426	0.1779	0.0848	0.1688	0.0385	0.0714
Mistral-Small-24B	0.0054	0.0508	0.0569	0.0055	0.0549	0.0533	0.1264
Qwen2.5-32B	0.0000	0.0001	0.0033	0.0001	0.0085	0.0001	0.0040
Gemma-4-31B	0.0006	0.0079	0.0944	0.0002	0.0824	0.0008	0.0800

Table 12: L-BBQ score (lower is less biased) for the four additional models, full 58,492-item BBQ. **Bold** marks the largest-bias removal (necessity) and the lowest-bias single-third injection (sufficiency) within each row.

Model	full	—first	—mid	—last
Mistral-7B	0.499	0.542	0.523	0.498
Mistral-Small-24B	0.773	0.762	0.766	0.773
Qwen2.5-32B	0.822	0.823	0.824	0.821
Gemma-4-31B	0.679	0.725	0.664	0.685

Table 13: MMLU accuracy for the four additional models with MMLU data (lm-eval-harness fixed evaluation set). The BBQ–MMLU dissociation is one-directional here: mid-third removal, which is BBQ-critical (Table 12), moves MMLU by at most 2.4 pp in any of the four, and where it lowers MMLU at all the loss is at most 1.5 pp.

M Generalization to New Families and Larger Scales

Status of this appendix. The main text reports the *primary controlled set*: 11 same-release base–instruction pairs, so that ΔW isolates the alignment step. The four models introduced here are also same-release pairs, but were added after the primary set was finalized; they test whether the mid-third result *reproduces* outside that set, on a family the main text never touches (Mistral) and at a scale it does not reach (24–32B). We therefore report them as a separate extension and do not merge them into any main-text table. Appendix N uses the same four models and raises the resolution from block level to per-layer analysis rather than adding coverage.

Setup. Four additional models: Mistral-7B (Jiang et al., 2023), Mistral-Small-24B (Mistral AI, 2025), Qwen2.5-32B (Yang et al., 2024), and Gemma-4-31B (Gemma Team, 2026). Each is a same-release base–instruct pair: Mistral-7B-v0.1 / Mistral-7B-Instruct-v0.1, Mistral-Small-24B-Base-2501 / -Instruct-2501, Qwen2.5-32B / -Instruct, and gemma-4-31B / -it. Protocol is identical to the main text (§3): unscaled ΔW injection, bf16, no quantization, and the *full* BBQ set of 58,492 items — no subsampling — for every model and every condition. Layer thirds use the same $t_1 = \lfloor n/3 \rfloor$, $t_2 = \lfloor 2n/3 \rfloor$ split.

Result. Necessity replicates in all four models, including the new Mistral family and both ≥ 31 B models. Sufficiency replicates in all four as well. The two large models behave as expected but with different magnitudes: Gemma-4-31B shows a large necessity effect ($0.0006 \rightarrow 0.0944$), whereas Qwen2.5-32B is *directional only* — its base model is already close to unbiased on BBQ (0.00005), so all its effects are small in absolute terms even though their ordering is correct. Mistral-Small-24B likewise orders correctly with a narrow margin (0.0533 only-mid vs. 0.0549 only-first). We state these calibration caveats rather than reporting only the ordering.

Model	Layers	Thirds (first/mid/last)	Best layer	Best L-BBQ	Third means (first/mid/last)	Lowest
Mistral-7B	32	0–9 / 10–20 / 21–31	L18 (mid)	0.0364	0.0766 / 0.0533 / 0.0622	mid
Mistral-Small-24B	40	0–12 / 13–25 / 26–39	L15 (mid)	0.1072	0.1279 / 0.1254 / 0.1305	mid
Qwen2.5-32B	64	0–20 / 21–41 / 42–63	L37 (mid)	0.0045	0.0089 / 0.0067 / 0.0085	mid
Gemma-4-31B	60	0–19 / 20–39 / 40–59	L57 (last)	0.0231	0.0458 / 0.0377 / 0.0523	mid

Table 14: Per-layer single-layer injection sweeps for the four additional models (L-BBQ, full 58,492-item BBQ; lower is less biased). **Bold** marks the lowest third mean. The mid-third has the lowest mean L-BBQ in all four models. The best single layer is mid-third in three of four; in Gemma-4-31B the global minimum is L57 (last third, 0.0231), numerically near-tied with the best mid-third layers (L26 at 0.0233, L27 at 0.0235), while the last-third *mean* is inflated by the final two layers (L58 0.106, L59 0.132).

Model	mid – first	mid – last
Mistral-7B	–0.0234 [–0.0252, –0.0215]	–0.0089 [–0.0106, –0.0072]
Mistral-Small-24B	–0.0024 [–0.0051, +0.0002]	–0.0050 [–0.0077, –0.0023]
Qwen2.5-32B	–0.0021 [–0.0024, –0.0018]	–0.0017 [–0.0020, –0.0014]
Gemma-4-31B	–0.0081 [–0.0093, –0.0069]	–0.0146 [–0.0158, –0.0132]

Table 15: Bootstrap 95% CIs for the per-layer third-mean contrasts (L-BBQ), $B = 10,000$, stratified by category. **Bold** = interval excludes zero. Negative values mean the mid-third mean is lower (less biased).

N Per-layer Resolution for the Additional Models

Appendix M establishes the block-level result for the four additional models. Here we raise the resolution to individual layers: for each model we inject each layer’s ΔW_i alone and score the full 58,492-item BBQ set under L-BBQ (Eq. 2), exactly as in the main-text sufficiency sweep (§3); protocol is otherwise unchanged (unscaled ΔW injection, bf16, no subsampling). This is $32 + 40 + 64 + 60 = 196$ single-layer conditions. Table 14 summarizes the sweeps and Figure 14 shows the full curves.

Confidence intervals. Table 15 reports 95% bootstrap CIs for the two third-mean contrasts, mid–first and mid–last, using the same stratified multinomial bootstrap as Appendix P ($B = 10,000$, stratified by BBQ category; independent draws across layers, hence conservative). All eight point estimates are negative, and seven of the eight confidence intervals exclude zero. The single exception is the mid–first contrast of Mistral-Small-24B ([–0.0051, +0.0002]), the same model already flagged in Appendix M for its narrow block-level margin; its mid–last contrast does exclude zero.

These sweeps provide higher-resolution evidence consistent with the block-level result. These four same-release pairs were added after the primary 11-model set was finalized and are therefore reported as a separate extension (Appendix M).

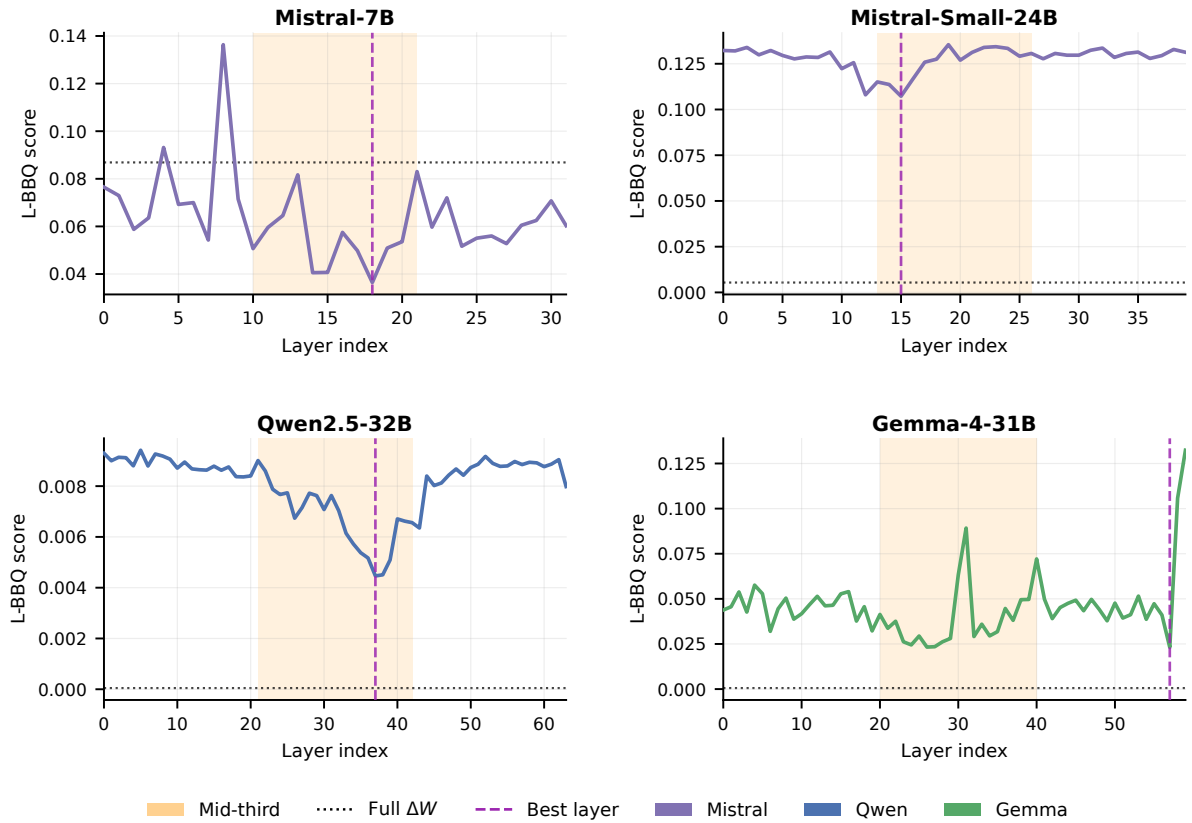


Figure 14: Per-layer L-BBQ under single-layer ΔW_i injection for the four additional models. Shading marks the mid-third; the dotted line is the full- ΔW score; the dashed line marks the best (lowest) single layer. Qwen2.5-32B shows the mid-third minimum observed in the main-text sweeps; Gemma-4-31B shows a contiguous mid-third low region (L23–L29) with a near-tied isolated minimum at L57 and a sharp rise over the final two layers.

Model	Family	full	−first	−mid	−last
Qwen2.5-0.5B	Qwen	43.7	+2.3	+1.1	+0.4
Qwen2.5-1.5B	Qwen	56.0	−0.0	+0.5	+0.4
Qwen2.5-3B	Qwen	63.8	+0.3	−0.8	+0.3
Qwen2.5-14B	Qwen	78.9	−0.4	−1.0	−0.2
Gemma-2-2B	Gemma	49.6	−12.4	−6.7	+0.9
Gemma-2-9B	Gemma	63.6	−35.0	−3.9	−0.5
Gemma-3-12B	Gemma	57.0	−7.5	−6.8	−0.4
Llama-3.2-1B	Llama	39.6	+2.6	−13.5	+2.7
Llama-3.2-3B	Llama	56.4	−1.0	−4.8	−0.2
Llama-3.1-8B	Llama	64.5	−1.5	−0.5	+0.1
Llama-3-8B	Llama	25.2	+5.0	+2.0	−0.1
Mistral-7B	Mistral	49.9	+4.3	+2.4	−0.1
Mistral-Small-24B	Mistral	77.3	−1.2	−0.7	−0.1
Qwen2.5-32B	Qwen	82.2	+0.1	+0.2	−0.1
Gemma-4-31B	Gemma	67.9	+4.6	−1.5	+0.6

Table 16: MMLU under block removal across 15 models. *full* is absolute accuracy (%); the removal columns are differences from *full* in percentage points. The large first-third capability cost is confined to Gemma-2/3. Mid-third removal is not uniformly cheap either — Llama-3.2-1B loses 13.5 pp — so the dissociation is a claim about the four models of Table 6, where both benchmarks were measured under matched conditions, and not a claim about all 15.

O Family Dependence of the First-Third MMLU Effect

The main text reports that removing the first-third can collapse MMLU while leaving BBQ nearly unchanged. Table 16 shows that this is **not a property of the first-third in general but of specific architectures**. Within the controlled set the separation is complete: the three Gemma models are ranks 1–3 of MMLU loss (−35.0, −12.4, −7.5 pp; mean −18.3), while the eight non-Gemma models average +0.9 pp — that is, first-third removal slightly *improves* MMLU outside Gemma. No non-Gemma model in the controlled set loses more than 1.5 pp.

Scale is a confound at 31B. Gemma-4-31B does *not* show the effect (+4.6 pp), but it is also far larger than the Gemma models in the controlled set (2–9B), so this single point cannot separate a scale effect from an architecture effect. The Mistral lineage provides evidence consistent with scale-associated weakening: the mid-minus-first localization gap drops by an order of magnitude from Mistral-7B to Mistral-Small-24B while its direction is preserved, although model and checkpoint differences between the two releases prevent attributing this to scale alone. We therefore do not read the Gemma-4-31B null result as evidence against the dissociation itself, which is stated only for the version-matched controlled set (Table 6) and does not depend on why the effect is absent at 31B. We accordingly scope the claim to the architectures and scales in which it is observed and do not present first-third removal as a general capability lesion. This is also why the abstract and introduction state the MMLU collapse conditionally.

Model	necessity (−mid − full)	sufficiency (only-mid − base)	vs. base (−mid − base)
Qwen2.5-0.5B	+0.0783 [+0.0678, +0.0891]	−0.0466 [−0.0612, −0.0319]	−0.0814 [−0.0945, −0.0678]
Qwen2.5-1.5B	+0.1179 [+0.1051, +0.1306]	−0.0865 [−0.1002, −0.0730]	−0.0199 [−0.0354, −0.0047]
Qwen2.5-3B	+0.0188 [+0.0165, +0.0212]	−0.0510 [−0.0557, −0.0465]	−0.0327 [−0.0379, −0.0276]
Qwen2.5-14B	+0.0066 [+0.0054, +0.0077]	−0.0121 [−0.0141, −0.0100]	−0.0055 [−0.0078, −0.0032]
Gemma-2-2B	+0.1687 [+0.1582, +0.1791]	−0.1479 [−0.1584, −0.1374]	+0.0238 [+0.0095, +0.0383]
Gemma-2-9B	+0.0574 [+0.0538, +0.0613]	−0.1160 [−0.1235, −0.1085]	−0.0595 [−0.0680, −0.0510]
Gemma-3-12B	+0.0687 [+0.0633, +0.0741]	−0.0746 [−0.0830, −0.0663]	−0.0236 [−0.0330, −0.0143]
Llama-3.2-1B	+0.1114 [+0.0995, +0.1232]	−0.1708 [−0.1829, −0.1588]	−0.0468 [−0.0613, −0.0324]
Llama-3.2-3B	+0.2146 [+0.2036, +0.2253]	−0.1559 [−0.1644, −0.1473]	+0.0732 [+0.0598, +0.0866]
Llama-3.1-8B	+0.0301 [+0.0249, +0.0354]	−0.0546 [−0.0637, −0.0457]	−0.0505 [−0.0597, −0.0412]
Llama-3-8B	+0.1149 [+0.1033, +0.1267]	−0.0622 [−0.0720, −0.0527]	+0.0576 [+0.0464, +0.0692]

Table 17: Bootstrap 95% CIs on the L-BBQ score, $B = 10,000$, stratified by category. **Bold** = interval excludes zero. Necessity is significant and positive in 11 of 11; sufficiency is significant and negative in 11 of 11. The third column is the active-suppression test: all 11 intervals exclude zero, but the sign is what matters — it is *positive* in exactly 3 (Gemma-2-2B, Llama-3.2-3B, Llama-3-8B), the three models reported in the main text, and negative in the other 8, which remain less biased than base even without the mid-third. Contrast intervals use independent draws rather than paired ones; since the paired correlation is positive, this widens the intervals and is conservative. [†]The base condition here is the zero-injection control (an apply/unapply round trip), not the base model read directly; this control runs 0.001–0.002 higher, so the active-suppression test is if anything understated. Every contrast marked significant also survives Bonferroni correction for 11 models.

P Bootstrap Confidence Intervals

Table 17 gives 95% bootstrap confidence intervals for the three central contrasts on the 11-model controlled set. Resampling is stratified by BBQ category over the 29,246 ambiguous items with $B = 10,000$ replicates.

Model	Scorer	Block removal			Block-only		
		—first / —mid / —last	first / mid / last				
Qwen2.5-3B	L-BBQ	0.000 / 0.019 / 0.000	0.034 / 0.001 / 0.029				
	Official	0.037 / 0.125 / 0.037	0.156 / 0.042 / 0.139				
Gemma-2-9B	L-BBQ	0.001 / 0.057 / 0.000	0.040 / 0.001 / 0.124				
	Official	0.031 / 0.157 / 0.024	0.148 / 0.031 / 0.121				
Llama-3-8B	L-BBQ	0.056 / 0.182 / 0.062	0.175 / 0.063 / 0.134				
	Official	0.071 / 0.101 / 0.068	0.097 / 0.068 / 0.136				

Table 18: Ambiguous-context bias under L-BBQ and the official polarity-aware BBQ scorer, computed from identical predictions for three models. **Bold** marks the largest block-removal score (necessity) and the lowest block-only score (sufficiency). Results are limited to the shown models and block-level conditions.

Q Robustness to the Official BBQ Scorer

L-BBQ assigns its biased/anti-biased labels without regard to question polarity (§3.2), whereas the official BBQ scorer of Parrish et al. (2022) reverses the two roles for non-negative questions. To test whether the paper’s block-level conclusions depend on this choice, we re-scored the *same* per-example predictions under an independent reimplementation of the official metadata- and polarity-aware scorer. The reimplementation was validated exhaustively against the pinned reference implementation (lm-evaluation-harness v0.4.11, BBQ task utilities): across all $58,492 \times 3 = 175,476$ (example, candidate) cases it agrees with the reference on every case, with zero mismatches in every category, polarity, and context condition.

The three models of the main-text panel (Figure 1) — Qwen2.5-3B, Gemma-2-9B, and Llama-3-8B — were re-scored on the block-level conditions: the three block removals and the three single-third injections, on the full 58,492-item BBQ set with per-example predictions saved. The check covers necessity (which third, when removed, raises bias most) and sufficiency (which third, injected alone, gives the lowest bias) for all three evaluated models, and, for Llama-3-8B specifically — the one model among the three whose outer-only removal raises bias above the unaligned base (§5.2) — whether the active-suppression verdict itself is preserved (below). We report official-score robustness only for these block-level conditions and do not extend the claim to the exact per-layer profiles. Table 18 reports the necessity and sufficiency scores computed from identical predictions.

Response-mapping exception. The original implementation assigned L-BBQ labels by matching the selected answer’s group description against the `stereotyped_groups` metadata. Due to surface-form discrepancies between some answer descriptions and metadata labels, a condition-dependent subset of non-UNKNOWN predictions did not receive a biased or anti-biased label and was consequently excluded from N . This behaviour was an implementation-level exception rather than part of the intended L-BBQ definition. Across the 33 robustness conditions for the three evaluated models, the unassigned rate ranged from 0.69% to 20.05%, with a prediction-weighted mean of 8.22%.

To assess the sensitivity of our conclusions to this exception and to the scoring definition, we rescored the same predictions using a metadata-aware, polarity-aware implementation verified against lm-evaluation-harness v0.4.11. The block-level necessity and sufficiency conclusions were preserved for all three evaluated models — Qwen2.5-3B, Gemma-2-9B, and Llama-3-8B — although the absolute score magnitudes differed. The active-suppression verdict is a separate matter and is not preserved for Llama-3-8B under this rescoring; we report that divergence explicitly below rather than extending the necessity/sufficiency robustness claim to it. This analysis serves as a targeted robustness check and does not replace the exhaustive L-BBQ per-layer sweeps across all 11 models.

Active suppression is scorer-sensitive for Llama-3-8B. The audit definition (§5.2) is Fail if outer-only bias exceeds the unaligned base; we compare the two quantities under both scorers, from the same predictions. Under L-BBQ, outer-only bias (0.182) exceeds the unaligned base (0.125) by +46%, matching the audit failure already reported for this model (Table 5). Under the official scorer, the same predictions give outer-only = 0.101 against base = 0.132 — the verdict flips to Pass. The reversal

reflects scorer-dependent shifts in both quantities: the unaligned base rises slightly ($0.125 \rightarrow 0.132$), whereas outer-only decreases substantially ($0.182 \rightarrow 0.101$), reversing the sign of their margin. Necessity and sufficiency remain scorer-consistent for all three evaluated models. For Llama-3-8B, the active-suppression audit is scorer-sensitive: the outer-only model exceeds the unaligned base under L-BBQ but not under the official scorer. We therefore restrict the official-score robustness claim to block-level necessity and sufficiency and do not claim scorer robustness for the active-suppression audit.