The Alignment Tax Concentrates in Output Projections

Ming Liu
Independent Researcher
qqliuming1995@gmail.com

Abstract

Instruction tuning degrades structured generation—the alignment tax. Prior work localizes this tax to transformer layers; we localize the degradation's functional sensitivity to individual projection matrices within each layer. Across three SwiGLU model families at 7–9B scale (Llama, Qwen, Mistral; Yi supplementary), alignment-induced degradation follows a consistent sub-layer hierarchy. Within the MLP, W_{down} carries the largest share and exceeds the two input projections at equal parameter counts and near-identical alignment-induced weight change (within-layer paired contrast positive in 6 of 7 conditions). Within attention, gradients reaching the value and output projections (W_V, W_O) exceed those reaching query and key (W_Q, W_K) by 1.76–3.91×, a consequence of softmax Jacobian attenuation. Three independent methods—weight patching, gradient analysis, and training-time exclusion—converge on this structure. A continual-pretraining control separates what is architectural from what is alignment-specific: the within-MLP hierarchy reproduces under general-text fine-tuning, but the layer-level placement of harmful modifications does not, and this task-specific signal is what makes targeted intervention work. Surgical Alignment Reversal (SAR), a training-free rollback of attribution-ranked components, improves structured generation on all four models, outperforming random selection by 2.8–5.3× (in-domain) and architectural-prior heuristics by 8.9× on held-out data (p=0.002); on Qwen, where the tax is positive, SAR recovers 62% in-domain and 47% under cross-dataset transfer. Output-Constrained DPO confirms the structure at training time: excluding output-pathway components from LoRA largely eliminates the tax, while excluding Q/K does not. We measure no safety degradation on the three primary models (refusal-rate $|\Delta| \leq 1.4$ pp; n=1,000 for Qwen/Llama, n=250 for Mistral; no significant adversarial ASR change); Yi is excluded from the safety-preservation claim because its safety and structured-generation components overlap. Attribution replicates at 14B and gradient asymmetry persists at 72B.

1 Introduction

Large language models (LLMs) are increasingly deployed as autonomous agents that must generate structured outputs: tool calls, JSON, SQL, code, and multi-step reasoning chains [Yao et al., 2023, Schick et al., 2023, Qin et al., 2024, Patil et al., 2023, Liu et al., 2024]. Instruction tuning via SFT, RLHF [Christiano et al., 2017, Ouyang et al., 2022, Bai et al., 2022], and DPO [Rafailov et al., 2023] is essential for safety and helpfulness, but it can degrade structured generation—the *alignment tax* [Lin et al., 2024]. Prior work identifies layer-level effects [Lin et al., 2024, Shi et al., 2024] but cannot determine which *projections within a layer* account for the change. Without this, interventions cannot isolate parameters by their architectural functional role.

Terminology. Each transformer layer contains seven weight matrices ("components"): the attention projections W_Q, W_K (which compute where to attend), W_V, W_O (which compute what information

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

is moved and how it is written back), and the SwiGLU MLP projections $W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}$. We measure the alignment tax as the increase in cross-entropy on gold structured-output tokens when moving from a base model to its instruction-tuned (IT) counterpart, and we attribute it to components by *weight patching*: replacing one IT weight matrix with its base-model counterpart and measuring how much the loss improves (the component's *harm score*; §4.1).

We close the localization gap by identifying a consistent functional-sensitivity pattern and showing it makes *correct predictions*—at inference time (SAR rollback, cross-dataset transfer) and at training time (OC-DPO exclusion, attribution-independent data). A continual-pretraining control reveals the within-MLP hierarchy is architecturally driven while the V/O > Q/K magnitude depends on alignment training dynamics; what distinguishes alignment is the *layer-level* distribution of harmful modifications—the task-specific signal SAR exploits.

Key finding. Alignment-induced functional sensitivity concentrates in *output-pathway* components, creating structure not visible to layer-level methods (Figure 1):

$$\underbrace{W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}}_{\text{MLP (59–63\% of change)}} \quad \Big| \quad \underbrace{W_O \geq W_V}_{\text{Attn output (20–31\%)}} \quad \gtrsim \quad \underbrace{W_Q \geq W_K}_{\text{Attn routing (8–18\%)}} \tag{1}$$

MLP dominance (59–63%) reflects total-quantity concentration rather than per-parameter susceptibility. For the *attention* asymmetry, softmax Jacobian attenuation is consistent with 1.76–3.91× larger gradients for W_V/W_O ($p < 2 \times 10^{-9}$; §4.2).[1]

Two interventions test predictions from this structure. **SAR**—training-free rollback of attribution-ranked harmful components—achieves 2.8–5.3× recovery over random baselines. **OC-DPO** confirms directionality at training time: excluding output-pathway components from LoRA largely eliminates the alignment tax while excluding Q/K does not—consistent across all three models (Table 5).

Contributions.

1. **Within-layer hierarchy** ((1)): $W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$ (2.5:1.5:1) at equal parameter counts, and V/O > Q/K by 1.76–3.91× in gradient magnitude ($p < 2 \times 10^{-9}$), confirmed by three independent methods across three SwiGLU families—structure invisible to prior layer-level methods and not resolved by concurrent work [Qi et al., 2026] (demonstrated head-to-head: their rankings overlap ours by ≤2/11 top-k and underperform ours for rollback on every model; Appendix AN).
2. **Diagnostic interventions**: SAR—operationally equivalent to selective task arithmetic negation [Ilharco et al., 2023]; the contribution is the *attribution-guided selection criterion* (8.9× over architectural-prior heuristics on held-out Berkeley Function-Calling Leaderboard (BFCL), p=0.002)—recovers 47% of alignment tax (cross-entropy metric) on held-out data (62% in-domain; 2.8–5.3× over random) with no significant refusal-rate change under non-adversarial ($|\Delta| \leq 1.4$ pp, keyword, n=1,000) or adversarial evaluation (GCG/template $|\Delta\text{ASR}| \leq 6$ pp, p>0.6, n=50; §7, Appendix M); OC-DPO confirms directionality at training time; attribution and interventions replicate at 14B; gradient asymmetry persists at 72B (gradient-only, 4-bit; §7, Appendices O–T). Neither intervention is offered as a methodological contribution—SAR is selective task-arithmetic negation and OC-DPO is component masking; what is ours is the selection criterion.
3. **Methodological correction**: systematic weight-patching bias from W_Q cascading effects (§4.2), resolved by combining forward-pass patching with gradient analysis.

2 Related Work

Alignment tax, model merging, and pruning. The alignment tax has been documented and mitigated via model averaging [Lin et al., 2024], task arithmetic [Ilharco et al., 2023], disperse-then-merge [Fu et al., 2024], elastic reset [Noukhovitch et al., 2023], and geometry-based approaches [Young, 2026, Bhardwaj et al., 2024, Lu et al., 2024]; layer-level attribution methods [Shi et al., 2024, Zhang & Dong, 2025] identify which *layers* carry alignment modifications, while the

[1] Qwen shows $W_Q > W_O$ under the ratio-of-absolute-sums convention; under the within-layer paired contrast of Appendix AL the V/O > Q/K asymmetry is significant for Qwen (p=0.0019), and gradient-based methods confirm it across all four families.

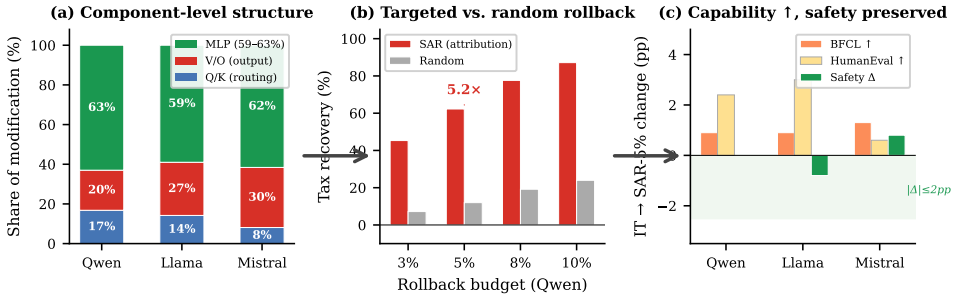

Figure 1: **Overview.** (a) Output-pathway components account for 49–65% of alignment-induced weight change (MLP 59–63%, Attn-output 20–31%) across three primary SwiGLU families (with Yi-1.5-9B supplementary); the within-MLP hierarchy $W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$ holds across all four families; V/O > Q/K is significant by weight patching for Llama/Mistral and established by gradient analysis ($p < 2 \times 10^{-9}$) for all four (§4.2). (b) SAR-5% rolls back attribution-ranked components: absolute CE loss reduction exceeds random selection by 2.8–5.3× on every model; "% of tax recovered" is informative only where the tax is substantially positive (Qwen 62%)—Llama and Mistral have net-negative tax and Yi's is small under matched tokenization, so absolute Δloss is the cross-model metric (Table 4, Appendix A). (c) SAR-5% improves structured generation with $|\Delta| \leq 1.4$ pp refusal-rate change (LLM-judge primary; keyword n=1,000 consistent; see §7).

superficial alignment hypothesis [Zhou et al., 2023b, Li & Kim, 2026] suggests alignment is a thin modifier atop base capabilities—our work refines this to identify *which* components within each layer are most affected, finding that the modification is shallow along both the layer axis and the component axis (concentrated in output-pathway projections). Model merging [Yu et al., 2024, Yadav et al., 2023, Wortsman et al., 2022, Deep et al., 2024, Farn et al., 2025], structured pruning [Sun et al., 2024, Frantar & Alistarh, 2023], LoRA analysis [Biderman et al., 2024, Xue & Mirzasoleiman, 2025], and constrained generation [Willard & Louf, 2023] do not distinguish by component type within a layer. SAR is operationally equivalent to selective task arithmetic negation [Ilharco et al., 2023] (λ=-1 on an attribution-selected subset S, λ=0 elsewhere); our contribution is the attribution-guided *selection* rather than the merge operation itself.

Mechanistic interpretability. The Transformer Circuits QK/OV split [Elhage et al., 2021] motivates our attention asymmetry framing; attention head specialization [Voita et al., 2019, Michel et al., 2019] establishes within-layer heterogeneity at the head level; we extend this to projection-matrix granularity. Activation-space methods [Zou et al., 2023, Arditi et al., 2024] and SAE-level decomposition [Ferrao et al., 2025] are complementary but do not resolve sub-layer modification structure. Mechanistic analyses of fine-tuning [Prakash et al., 2024, Jain et al., 2024, Lee et al., 2024] characterize how training modifies existing circuits; we extend this from single-task procedural settings to alignment-induced changes across model families. Concurrent work [Qi et al., 2026, Huang et al., 2025, Chen et al., 2025] independently identifies value matrices, MLPs, or attention heads as safety-critical *for refusal behavior* via gradient-based or neuron-level scoring; our work is complementary but distinct—we study *structured-generation capability* (tool calls, JSON, SQL) rather than safety behavior—and adds: (1) sub-MLP resolution ($W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$, 2.5:1.5:1) that aggregate scoring does not resolve; (2) identification of systematic weight-patching bias (§4.2); and (3) intervention framework (SAR, OC-DPO) validated on cross-dataset transfer. A direct comparison (Appendix AN) confirms the distinction empirically: component-level adaptations of their scoring correlate weakly with our harm ranking (top-k overlap ≤2/11) and underperform our attribution for rollback on every model (best condition per model: 59%/16%/75% of SAR's effect on Qwen/Llama/Mistral; every raw condition on Llama is ≤0%).

Knowledge editing and gradient projection. FFN layers function as key-value memories [Geva et al., 2021, 2022], and ROME/MEMIT [Meng et al., 2022, 2023] localize factual edits there—though localization does not always predict editability [Hase et al., 2023]. Hernandez et al. [Hernandez et al., 2024] show that factual recall localizes to W_{down} projections specifically, corroborating our finding that W_{down} dominates alignment modifications; our work extends this from factual recall to alignment-induced structured-generation changes. OGPSA [Sun et al., 2026] and NSPO [Niu et al., 2026] project gradients to preserve capabilities; we provide the complementary diagnostic

of *which* components to protect, enabling output-pathway-focused OGPSA with 57% fewer SVD computations (Appendix I).

3 Experimental Setup

Models. We study three primary model families at 7–9B scale as base–instruct pairs: Llama-3.1-{8B, 8B-Instruct} [Grattafiori et al., 2024], Qwen2.5-{7B, 7B-Instruct} [Qwen Team, 2024], and Mistral-{7B-v0.3, 7B-Instruct-v0.3} [Jiang et al., 2023], plus Yi-1.5-{9B, 9B-Chat} [Young et al., 2024] as a supplementary family (excluded from intervention evaluation due to safety–structure overlap; Appendix N), spanning distinct alignment procedures (Llama: SFT+RLHF, Qwen: SFT+DPO, Mistral: SFT, Yi: SFT+RLHF) and attention variants (Llama/Mistral: 4:1 GQA; Qwen: 7:1 GQA; Yi: 8:1 GQA [Ainslie et al., 2023]). Under GQA, total parameters in V+O equal total parameters in Q+K, so our grouped comparisons are parameter-balanced; GQA compresses W_V relative to W_O, which may suppress absolute W_V harm scores in high-GQA models (see Appendix AC). All four share the SwiGLU MLP [Shazeer, 2020] and decoder-only architecture; generalization to other MLP types (e.g., MoE) or encoder-decoder models remains open. Scaling validation uses Qwen2.5-14B (attribution: fp16 on 2×A100; SAR: 4-bit NF4; §8) and Qwen2.5-72B (4-bit NF4; Appendix T).

Evaluation. We curate a mechanistic probe set of 209 structured-generation examples covering 18 task categories (tool calls, JSON, SQL, code, API, bash, config, ReAct, etc.) with gold structured outputs, used for both attribution computation and primary evaluation (held-out validation in §7). Cross-entropy on gold tokens is the primary metric, following mechanistic interpretability work that establishes component localization via loss-based sensitivity [Wang et al., 2023, Olsson et al., 2022]; token-level analysis confirms CE improvement concentrates on structural tokens (brackets, keywords) rather than content tokens (Appendix AF). Full probe set composition in Appendix AI. Execution benchmarks (BFCL [Patil et al., 2025], HumanEval [Chen et al., 2021], IFEval [Zhou et al., 2023]) are directionally consistent but do not reach individual significance (§7). We use raw prompt-completion format; chat-template yields *larger* tax (Appendix AS). Attribution forward passes for the 7–9B models run in fp16; bf16 and protocol controls are reported in Appendix AS. All evaluations use greedy decoding.

Component taxonomy. Each transformer layer [Vaswani et al., 2017] contains 7 projections: W_Q, W_K, W_V, W_O (attention) and W_{gate}, W_{up}, W_{down} (MLP), yielding $7L$ components per model. We use *output pathway* to denote $\{W_V, W_O, W_{\text{down}}\}$—projections that transform representations toward the output token distribution—as distinct from routing (W_Q, W_K) and gating (W_{gate}, W_{up}) components. Our tripartite division follows functional role rather than modification magnitude; key results are type-agnostic.

Compute and licenses. Primary experiments run on NVIDIA A100 GPUs (∼80 GPU-hours); the additional analyses of Appendices A–C add ∼31 A10G GPU-hours on AWS SageMaker (g5 instances; the fp32 precision referee sharded across 4 GPUs). Model licenses: Llama (Meta Community), Qwen2.5/Qwen3/Mistral/Yi/OLMo-2 (Apache 2.0); UltraFeedback [Cui et al., 2024] (MIT).

4 Component-Level Structure of Alignment

4.1 Attribution via Weight Patching

We define the *alignment tax* as $\tau = \mathcal{L}(\theta_{\text{IT}}, \mathcal{D}) - \mathcal{L}(\theta_{\text{base}}, \mathcal{D})$, where $\mathcal{L}$ is cross-entropy on gold structured tokens and $\mathcal{D}$ is the probe set (positive = degradation). For each component (t, l), we replace the parameter matrix $W^{(l)}_{t,\text{IT}}$ with $W^{(l)}_{t,\text{base}}$ and measure the *harm score*:

$$h(t, l) = \mathcal{L}(\theta_{\text{IT}}, \mathcal{D}) - \mathcal{L}(\theta^{(t,l\leftarrow\text{base})}_{\text{IT}}, \mathcal{D}) \tag{2}$$

where $\theta^{(t,l\leftarrow\text{base})}_{\text{IT}}$ denotes the IT model with component (t, l)'s weight matrix replaced by its base counterpart during a single forward pass, all other parameters unchanged. Each of the $7L$ components is evaluated independently. A positive $h(t, l)$ means rolling back this component *reduces* agent loss (i.e., the IT modification is task-harmful). SAR *predicted* recovery is $\sum_{(t,l)\in S} h(t, l)/\tau$ when $\tau > 0$

Table 1: **V/O vs. Q/K attribution** (209 examples). Percentages are shares of total change (denominator = sum of absolute harm scores $\sum |h(t,l)|$ across all components; what this convention does and does not decompose, and three alternatives, are examined in Appendix B). $p^{\dagger}_{\text{MW}}$: Mann-Whitney U (anti-conservative; see p_{block}); r: rank-biserial; p_{block}: block permutation test preserving layer autocorrelation (Appendix AA). All rows use the fp16 raw-format protocol (§3); an fp32 referee run confirms Mistral's ratio (3.80×, bootstrap 95% CI $[2.8, 4.5]$), with protocol controls in Appendix AS (orderings and significance hold in every condition; the wide CI reflects Mistral's small Q/K denominator).

Model	**V/O (%)**	**Q/K (%)**	**V/O / Q/K**	$p^{\dagger}_{\text{MW}}$	p_{block}	r
Qwen2.5-7B	20.2	16.8	1.20×	1.6×10^{-3}	0.28	0.32
Llama-3.1-8B	26.8	14.2	1.89×	6.9×10^{-5}	1.1×10^{-4}	0.39
Mistral-7B	30.3	8.1	3.76×	$< 10^{-10}$	$< 10^{-5}$	0.66
Yi-1.5-9B	22.4	18.3	1.22×	0.34	0.13	0.03
Qwen2.5-14B	26.3	18.2	1.44×	4.2×10^{-6}	0.015	0.27

(used for component ranking; Table 4 reports actual measured loss reduction of the jointly modified model). This *weight patching*—replacing entire parameter matrices, as distinct from activation patching [Vig et al., 2020, Conmy et al., 2023] that replaces intermediate hidden states—probes each component's net contribution to the aligned model's behavior. This first-order approximation is independently supported by gradient-dot-delta attribution (Appendix X), gradient magnitude analysis (Table 3), and OC-DPO (§6). Pairwise interactions average <5% but can be large within layers (W_V/W_O: up to 89%; Appendix X); a greedy joint-selection baseline confirms first-order selection recovers 72–99% of what exact-objective greedy selection achieves, with 64–82% component overlap (Appendix X).

Table 1 shows the V/O vs. Q/K asymmetry across all five models. Because adjacent layers are not fully independent (lag-1 autocorrelation up to 0.66 at 7B, 0.83 at 14B), Mann-Whitney p-values are anti-conservative; we primarily rely on p_{block} from a block permutation test (Appendix AA). The V/O > Q/K effect is significant for Llama ($p_{\text{block}} = 1.1 \times 10^{-4}$), Mistral ($p_{\text{block}} < 10^{-5}$), and 14B ($p_{\text{block}} = 0.015$), but not for Qwen-7B ($p_{\text{block}} = 0.28$) or Yi ($p_{\text{block}} = 0.13$)—both have small effect sizes (1.20–1.22×) and high GQA ratios (7:1, 8:1) where parameter-sharing may equalize Q/K contributions; BH correction leaves all conclusions unchanged. For Qwen and Yi, the V/O > Q/K claim rests on independent gradient analysis (Table 3: all four models $p < 2 \times 10^{-9}$), and we do not claim patching-based support for them: an exact interaction-aware Shapley decomposition confirms the Qwen raw-protocol asymmetry is undetectable (1.03×, z=0.11), while the same decomposition under the chat protocol gives 2.40× (Appendix AB). The convergence of patching (3/5 significant), gradients (5/5 at $p < 10^{-9}$, robust to per-element size normalization), and a non-GQA architectural control (OPT-6.7B, Appendix AO) constitutes triangulation across methods; the two patching nulls are reported as nulls rather than absorbed into a normalization convention (Appendix AC). The strongest finding is **MLP dominance** (Figure 2): MLP absorbs 59–63% of alignment-induced change, with W_{down} consistently highest (26–34%; protocol controls in Appendix AS). Within MLP—all three projections have identical parameter counts—the ordering $W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$ (30.6% > 18.1% > 12.2% mean; Table 2) cannot be explained by capacity. It is also not explained by perturbation size: the alignment-induced weight change is near-identical across the three, with $\|\Delta W\|_F$ ratios of 0.95–1.00 (down/up) and 0.93–0.97 (up/gate) across the four families, so the ordering is a *sensitivity* hierarchy at matched parameter count *and* matched perturbation magnitude. We separate the two halves of the ordering by strength: W_{down} exceeding the two input projections is robust (within-layer paired contrast, layer as unit, positive in 6 of 7 model×protocol conditions, BH $q < 0.05$ in 5; Appendix AL), whereas $W_{\text{up}} > W_{\text{gate}}$ reaches significance in only 3 of 7 conditions and, where positive, is carried by a minority of layers—we therefore report it as suggestive rather than established. Page's test confirms this per-layer ordering in all three families (Appendix AA). Within attention, V/O accounts for 20–31% while Q/K accounts for 8–18%; the ratio is largest for Mistral (3.76×, confirmed by an fp32 referee; Table 1), though its precision is limited—uniquely among the three primary models—by a Q/K denominator near the sampling-noise floor (Appendix AS); Qwen and Yi are smallest (1.20–1.22×). These grouped ratios are already per-parameter quantities, since $|W_V|+|W_O| = |W_Q|+|W_K|$ under any GQA ratio (Appendix AC). The combined output pathway (W_V, W_O, W_{down}) accounts for 49–65% of total change. A general-knowledge QA control (60

Table 2: **Complete modification hierarchy**: share of alignment-induced change by component type. MLP dominates (59–63%) across all four 7–9B models. Within MLP, $W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$ at equal parameter counts (2.5:1.5:1 ratio), inconsistent with capacity-based explanations; the functional origin (e.g., residual-stream proximity) remains open.

Component	Qwen (%)	Llama (%)	Mistral (%)	Yi (%)	Mean (%)
W_{down}	28.7	26.7	34.2	32.6	**30.6**
W_{up}	21.6	19.8	15.9	15.0	18.1
W_{gate}	12.7	12.6	11.5	11.8	12.2
W_O	10.6	16.4	15.9	10.7	13.4
W_V	9.6	10.4	14.5	11.6	11.5
W_Q	12.2	9.7	5.4	10.4	9.4
W_K	4.6	4.5	2.7	8.0	5.0

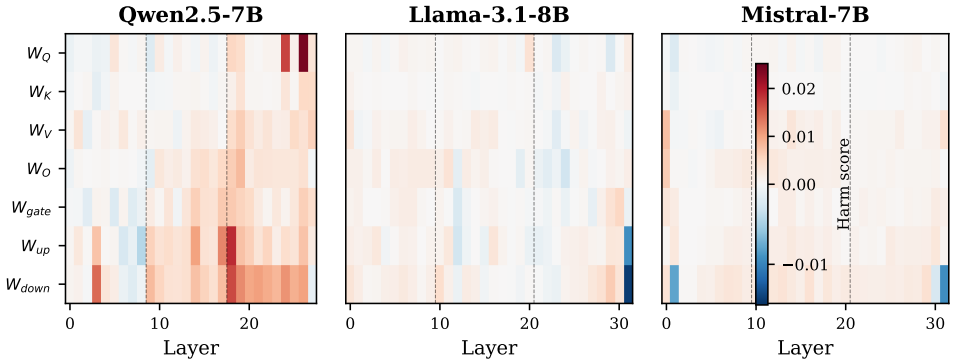

Figure 2: **Attribution heatmap** (three primary models). Each cell shows the harm score $h(t, l)$ for component type t at layer l; red = IT modification increases agent loss, blue = decreases. Key pattern: W_{down} row is consistently the most active, and V/O rows show larger modifications than Q/K across all layers. Dashed lines mark the mid-1/3 layer region.

examples) reproduces the ordering on three of four models but not on Llama-3.1-8B, where it is significantly reversed, so we treat the control as partially rather than fully confirmatory (Appendix S).

External validation. Attribution on 100 independent BFCL examples (Appendix W) reproduces the hierarchy outside the probe set: MLP shares 59.8–62.3%, V/O / Q/K ratios 1.17–2.17× ($p < 0.002$, all three models).

Alignment tax is task-dependent. Models with net-negative tax (Llama −0.027, Mistral −0.004) still contain individually harmful components; the hierarchy ((1)) holds regardless of tax direction, and SAR improves all four models (Table 4).

4.2 Gradient Asymmetry

The V/O vs. Q/K asymmetry has a partial structural explanation rooted in the attention computation graph. The DPO gradient with respect to W_O flows directly ($\partial\mathcal{L}/\partial W_O \propto (AV)^\top(\partial\mathcal{L}/\partial\mathbf{x})$), while the gradient with respect to W_Q must pass through the softmax Jacobian ($\|J_{\text{softmax}}\|_2 \leq 1/2$; proof in Appendix Z), which attenuates it. This predicts the *direction* of asymmetry (V/O > Q/K), with peaked attention distributions producing attenuation well beyond the 2× minimum; a full quantitative bound (including LayerNorm and activation norms) remains future work (Appendix Z).

Table 3 supports the prediction across all four families (per-layer visualization in Figure 4, Appendix U; the gap widens in middle layers). SFT gradients closely match (1.90–3.85×; Appendix U), indicating the asymmetry is loss-independent. Notably, the gradient asymmetry manifests as higher signal quality rather than larger weight deltas—Adam normalization equalizes update magnitudes (Appendix Y). The gradient account is *consistent with* alignment concentrating in V/O; our contribu-

Table 3: **DPO gradient magnitude ratio** (V/O vs. Q/K). Per-layer mean gradient norms, 10 DPO gradient samples (no parameter updates, fp16 forward); p-values via Mann-Whitney U on per-layer norms (Appendix U). Direction holds across all four families; magnitudes vary with GQA ratio.

Model	V/O / Q/K ratio	p-value
Qwen2.5-7B	1.76×	1.4×10^{-9}
Llama-3.1-8B	3.91×	$< 10^{-15}$
Mistral-7B	3.66×	$< 10^{-11}$
Yi-1.5-9B	2.74×	$< 10^{-11}$

Algorithm 1 Surgical Alignment Reversal (SAR)

Require: Base weights θ_{base}, IT weights θ_{IT}, probe set $\mathcal{D}$ (N examples), budget $k\%$

$\mathcal{L}_0 \leftarrow \frac{1}{N} \sum_{x \in \mathcal{D}} \text{CE}(\theta_{\text{IT}}, x)$ {Baseline loss}

for each component $(t, l) \in \{W_Q, W_K, W_V, W_O, W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}\} \times \{0, \ldots, L-1\}$ **do**

  $\theta' \leftarrow \theta_{\text{IT}}$ with $W_t^{(l)} \leftarrow W_{t,\text{base}}^{(l)}$ {Single-component swap}

  $h(t, l) \leftarrow \mathcal{L}_0 - \frac{1}{N} \sum_{x \in \mathcal{D}} \text{CE}(\theta', x)$ {Harm score ((2))}

end for

$S \leftarrow$ top-$\lfloor 0.01k \cdot 7L \rfloor$ components by $h(t, l)$ descending

$\theta_{\text{SAR}} \leftarrow \theta_{\text{IT}}$ with $W_t^{(l)} \leftarrow W_{t,\text{base}}^{(l)}$ for all $(t, l) \in S$

return θ_{SAR}

tion is showing this architectural fact is *actionable*: knowing where gradients concentrate enables targeted rollback.

Methodological insight. Weight patching shows O > Q for only 2/3 models under the ratio-of-absolute-sums convention. The resolution is not a cascade correction (the cascade itself is decomposed in Appendix AM) but a change of estimand: under a *within-layer paired signed contrast*—per layer, mean V/O harm minus mean Q/K harm, with the layer as the independent unit—the asymmetry is positive in all six evaluated pairs and significant in five, including both models the ratio convention reports as null (Qwen2.5 p=0.0019, Yi p=7.6×10^{-6}; Appendix AL). The ratio is also not invariant to numerically irrelevant choices: on Mistral, holding model, probe set and protocol fixed and varying only the forward dtype, it moves from 2.75 to 3.80, whereas the within-layer contrast moves by 2.3%. We therefore report the contrast as the primary functional statistic. Gradient norms—involving no downstream re-computation by construction—confirm V/O > Q/K across all four families, providing a complementary, patching-independent measure.

5 Surgical Alignment Reversal (SAR)

SAR (Algorithm 1) is a training-free method that rolls back the top-$k\%$ most harmful components identified by weight patching, selecting by harm score regardless of component type.

SAR outperforms random selection by 2.8–5.3× across models (Table 4; bootstrap 95% CIs exclude 1.0×); on Yi, random rollback is on average *harmful* (−10.8%), making targeted selection qualitatively necessary. The top components include not only W_{down} but also W_Q and W_{up}, consistent with cascading bias (§4.2). The heuristic baseline achieves 46.8% on Qwen but only 17.1% on Yi, showing attribution provides model-specific precision beyond architectural fingerprints (magnitude-based selection also underperforms, 2.0–3.4×; Appendix AD). Recovery scales monotonically with budget k (Appendix Q); component-level selection outperforms layer-level selection by 35–47 pp (Appendix AP); per-category analysis shows SAR helps 13–17 of 18 task categories (Appendix AH). A causal rollback experiment validates the output-pathway grouping (Appendix E); type-restricted ablation confirms Q/K-only rollback yields the smallest improvement across all models (Appendix R). TIES/RESTA achieve higher absolute recovery but degrade safety (Appendix F); SAR maintains execution accuracy (Appendix AE).

Table 4: **SAR results** (5% budget). Top: Qwen detail. Bottom: cross-model. **Recovery** = absolute CE loss reduction (Δloss, nats); **% of tax** = Recovery / alignment tax (omitted where tax is near-zero; negative values indicate net worsening). Random: 20 seeds (Yi: 10 seeds). Bootstrap 95% CIs of SAR/random ratio: Qwen $[3.6, 5.0]\times$, Llama $[4.1, 7.5]\times$, Mistral $[2.3, 3.6]\times$ (bootstrapped from n=20); Yi fold-ratio omitted because random mean is negative (-10.8%), making the ratio uninterpretable. †In-domain oracle; cross-dataset transfer achieves 47% (§7). ‡Yi's tax is small once base and chat checkpoints are scored with a matched tokenizer ($+0.026$ rather than $+0.102$), so "% of tax" is uninformative for Yi and only absolute Δloss is reported (Appendix A).

Qwen2.5-7B (tax = +0.095 on 209-example probe set)			
Method	k%	**Recovery**	**% of tax**
SAR (attribution)	5%	+0.059	**62.2%**†
SAR (attribution)	10%	+0.083	87.1%
Heuristic (V/O+W_{down}, mid)	$\sim$14%	+0.045	46.8%
Random 5% (20 seeds)	5%	+0.014 $\pm$ 0.005	14.8 $\pm$ 5.7%
Cross-model SAR-5% (absolute Δloss improvement; random = 20 seeds, Yi = 10)			
Llama-3.1-8B (SAR)	5%	+0.026	—
Llama-3.1-8B (random)	5%	+0.005 $\pm$ 0.003	—
Mistral-7B (SAR)	5%	+0.020	—
Mistral-7B (random)	5%	+0.007 $\pm$ 0.004	—
Yi-1.5-9B (SAR)	5%	+0.064	—‡
Yi-1.5-9B (random)	5%	-0.011 ± 0.039	—

Table 5: **OC-DPO prediction test** (20 pairs, seed 42; 30-example subset). Excluding V/O/W_{down} largely eliminates the tax; Q/K exclusion does not, despite more parameters (11.4–12.1M vs. 8.1–8.8M). >100%: over-correction at small scale. *Scale caveat*: the OC-DPO advantage is non-monotonic in training-pair count (strong at $\leq$50 pairs, negative at 100–200, recovering to 3.1$\times$ at 5000; Appendix J, Figure 3); OC-DPO is a component-type isolation test, not a scalable training method. Of the 20 training pairs, 1 shares a prompt with the 209-example probe set and 1 is a near-duplicate; the remaining 18 are disjoint (Appendix G). Multi-seed and full-eval results in Appendix H.

Condition	**Qwen**	**Llama**	**Mistral**	**Mean**
Standard DPO (tax)	+0.215	+0.411	+0.509	+0.378
Tax reduction (%)				
OC-DPO (excl. V/O/down)	**94.5**	**133.5**	**122.1**	**116.7**
Exclude Q/K (control)	8.9	-19.9	10.2	-0.3

6 Output-Constrained DPO: a Component-Type Isolation Test

To test whether the output-pathway concentration also manifests during training, we exclude output-pathway components (W_V, W_O, W_{down}) from LoRA [Hu et al., 2022] adapters during preference optimization (20 pairs, mid-third layers, $\mathcal{L} = -\log\sigma\big(\beta[\log\pi_\theta(y_w|x) - \log\pi_{\text{ref}}(y_w|x)] - \beta[\log\pi_\theta(y_l|x) - \log\pi_{\text{ref}}(y_l|x)]\big)$, β=0.1; we use "DPO" as shorthand following [Biderman et al., 2024]). This is a small-scale *directional prediction test*: if concentration is genuine, V/O/W_{down} exclusion should largely eliminate the tax while Q/K exclusion (control) should not. Values >100% indicate over-correction at this data scale.

Table 5 confirms the prediction: OC-DPO achieves 80–136% tax reduction (5-seed means; Appendix H) vs. -20% to 10% for Q/K exclusion (seed 42; Table 5)—an order-of-magnitude difference in every model. A same-parameter-count ablation (Appendix G) rules out regularization explanations: component identity, not parameter budget, determines the outcome. The correct sign holds for all 3 model families across all 5 seeds each (15/15 positive; models are the independent units: per-model binomial p=0.031, Fisher combined p=0.002), and the direction persists at 5000 pairs (3.1$\times$; Appendix J). OC-DPO preserves safety (Appendix K).

Table 6: **SAR-5% does not significantly degrade evaluated capabilities** (directionally positive on execution benchmarks but gains are not individually significant at these sample sizes). Agent loss (↓); MMLU (↑); BFCL AST (↑, 318 ex.); HumanEval pass@1 (↑, 164 probs.); IFEval prompt acc. (↑, 541 prompts); Refusal rate (LLM judge, 250 prompts). All greedy decoding. Details in Appendices L–V.

Model	Var.	Agent ↓	MMLU	BFCL	HumanEval	IFEval	Refusal
Qwen	IT	0.829	72.7	83.3	62.2	70.1	86.4
	SAR-5%	**0.770**	72.9	**84.3**	**64.6**	70.4	85.6
Llama	IT	0.812	67.7	78.0	49.4	68.6	62.4
	SAR-5%	**0.786**	67.8	**78.9**	**52.4**	68.6	60.4
Mistral	IT	0.629	58.8	72.0	30.5	50.6	13.2
	SAR-5%	**0.608**	58.5	**73.3**	**31.1**	50.8	12.0

7 Capability Preservation and Safety

We evaluate safety via keyword-based and LLM-judge refusal rate (non-adversarial) and automated jailbreak attacks (GCG [Zou et al., 2023b], fixed-template jailbreak [Wei et al., 2023]; Appendix M). Table 6 shows SAR-5% preserves general knowledge (MMLU [Hendrycks et al., 2021] $|\Delta| \leq 0.3$ pp), instruction following (IFEval $|\Delta| \leq 0.4$ pp), conversational quality (MT-Bench on Qwen: no significant change; Appendix AG), and refusal rates ($|\Delta| \leq 1.4$ pp, keyword method, n=1,000 for Qwen/Llama; n=250 for Mistral; we report the LLM judge throughout, as keyword matching over-counts Mistral's refusals; LLM-judge results in Table 6; worst-case CIs in Appendix D). Under adversarial attack (GCG, fixed-template jailbreak), SAR-5% does not significantly increase vulnerability on either Qwen or Llama (Fisher's exact, n=50/method, all $p \geq 0.64$; Appendix M). Structured-generation benchmarks (BFCL, HumanEval) are directionally positive but do not reach individual significance at these sample sizes; a pooled analysis across models and benchmarks (stratified exact McNemar over 6 paired strata; 1,119 model×example pairs from 373 unique items × 3 models: 37 SAR-only vs. 27 IT-only passes) confirms the gains remain directional but not significant even jointly (one-sided p=0.13; Stouffer over all 12 strata p=0.29). What the pooled data *does* establish is formal non-inferiority—the claim the paper actually makes: the paired accuracy difference is $+0.89$ pp in SAR's favor (95% CI $[-0.5, +2.3]$ pp, Wald; Tango score interval $[-0.5, +2.4]$ pp), and non-inferiority within a 1 pp margin holds (exact conditional one-sided p=0.006; the full test battery—Tango score, cluster-robust over the 373 unique items, and per-stratum leave-one-out, together with the superiority test that gives p=0.12 and so bars reading 0.006 as evidence of gain—is in Appendix AE). SAR does not degrade execution by more than 1 pp with 95% confidence; superiority remains unestablished. Direct execution shows no significant change on any model (Appendix AE). RESTA [Bhardwaj et al., 2024] degrades Llama's refusal rate by -22 pp while SAR-5% stays within 2 pp, confirming attribution-guided selection is critical for safety preservation (Appendix F). Yi-1.5-9B is excluded due to safety–structure entanglement (-13.6 pp refusal; Appendix N).

Overfitting analysis. The SAR component set is selected from the 209-example probe set, but three independent tests rule out circularity (additional tests in Appendices W, X): (i) BFCL-based selection (zero probe overlap) still recovers 47% on the probe vs. 14.8% random (3.2×; Appendix AJ); (ii) 50 random splits yield mean eval recovery $68.0\% \pm 5.4\%$ with a train–eval gap of only 3.2 pp, the top components stable under bootstrap resampling (Appendix X); (iii) OC-DPO's type-only masks (no per-layer scores) replicate the differential across all models. Within-BFCL split validation confirms 8.9× over heuristic (p=0.002; Appendix AK).

8 Discussion and Conclusion

What the gradient theory does and does not explain. The softmax Jacobian bound ($\|J_{\text{softmax}}\|_2 \leq 1/2$) predicts the *direction* of the V/O > Q/K asymmetry (guaranteeing at least 2× attenuation, with peaked attention producing stronger suppression) but not the exact magnitude; additional factors (LayerNorm scaling, GQA sharing, activation norms) modulate the ratio—a direction-only consistency check, not a quantitative theory. The within-MLP hierarchy ($W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$)—an architectural finding confirmed by CPT (§8)—is consistent with gradient-flow geometry, where the expansion/contraction asymmetry suggests $\|\partial\mathcal{L}/\partial W_{\text{down}}\|/\|\partial\mathcal{L}/\partial W_{\text{up}}\|$ should exceed unity, with $W_{\text{up}} > W_{\text{gate}}$ from SiLU saturation (Appendix U). We are careful about what the data support

here: the quantitative form $\approx \sqrt{D_i/D_h}$ is *not* supported—predicted and observed ratios are cross-model *negatively* associated rather than positively (Appendix U). We therefore state the geometry as a directional rationale for the ordering, not as a quantitative law. The scope of the attention claim likewise differs by measurement, and we state it that way: the *gradient* asymmetry is robust across families and protocols, whereas the *aggregate functional* asymmetry is protocol-dependent (Appendices AB, AS).

Alternative explanation: residual proximity. W_O and W_{down} project directly into the residual stream while W_Q/W_K/W_{gate}/W_{up} operate on internal subspaces; if proximity alone drove sensitivity, the pattern would appear in *any* fine-tuning. OPT-6.7B provides counter-evidence—W_V (not residual-proximate) dominates W_O, and fc1 (expand) exceeds fc2 (contract)—indicating that gradient magnitude reflects information-flow geometry rather than residual proximity (Appendix AO).

CPT control: what is and is not alignment-specific. A continual-pretraining control (LoRA on Wikitext; Appendix AQ) confirms the within-MLP hierarchy ($W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$) is *architectural*—reproducing under general-text fine-tuning—while the attention V/O > Q/K asymmetry depends on both architecture and alignment training dynamics. What is alignment-specific is the *layer-level* distribution—CPT modifies the same MLP types but different layers, so task-specific attribution (not architectural priors alone) determines SAR's efficacy (8.9× over the heuristic on held-out BFCL; Appendix AK).

Scale. Attribution replicates at 14B, where W_{up} and W_{down} become near-equal within the MLP, consistent with the smaller expansion ratio at that scale (Appendix O). Gradient asymmetry persists at 72B (Appendix T).

Falsification criteria. The hypothesis would be falsified by: SAR-over-random ratio near 1× (component selection is arbitrary), Q/K exclusion matching V/O exclusion in OC-DPO (output-pathway distinction is irrelevant), or SAR improving CE while significantly *degrading* execution (metric dissociation). None obtain: the 2.8–5.3× SAR ratio and OC-DPO's order-of-magnitude differential (80–136% 5-seed means vs. −20% to +10% for Q/K exclusion, seed 42) rule out conditions 1–2; condition 3 is likewise absent—per-model execution change stays within ±0.5 pp and is never significant (p=1.0 each; Appendix AE), with directional non-significance (CE improves, execution does not significantly *improve*) categorically distinct from the required significant *degradation*.

Practical implications. SAR's primary value is diagnostic; for production alignment, OGPSA [Sun et al., 2026] is preferable. Restricting OGPSA to output-pathway components achieves ≈89–99% effectiveness with 57% fewer SVD computations (Appendix I), conditional on pre-computed attribution (∼200 forward passes).

Limitations. *(1) Evaluation scope*: targets short structured outputs under greedy decoding; long-form, multi-turn, and sampling-based generation are untested. Chat-template evaluation yields a larger tax, but the hierarchy persists and SAR still recovers 33–67% while holding format accuracy flat (Δ=−0.5 pp; Appendix AS). *(2) Architecture and scale*: verified on SwiGLU decoder-only models (3 primary + Yi supplementary) with supporting evidence from OPT-6.7B and GPT-J-6B (Appendices AO, AR). The within-MLP hierarchy is SwiGLU-specific. Generalization to MoE, encoder-decoder, and SSM architectures is unknown. Validated to 72B (Appendices O–T). *(3) Safety*: refusal-rate is a narrow proxy. Adversarial coverage is limited to one white-box (GCG) and one black-box attack on two models (n=50 per condition, giving only ≈40% power against a 20 pp effect; Appendix M); non-adversarial n=1,000 rules out >1.9 pp degradation for Qwen (95% CI [−1.9, +1.1] pp). "Safety preservation" is not a comprehensive guarantee. *(4) Causal interpretation*: concentration represents reproducible *functional sensitivity* generating correct predictions (SAR, OC-DPO, OGPSA-hybrid), not proven parameter-storage causality; pairwise interactions (V/O within-layer: up to 89%) mean SAR recovery is a first-order approximation—though greedy joint selection on the true joint objective changes neither the components chosen nor the achievable recovery by much (Appendix X). *(5) Intervention scope*: SAR and OC-DPO are mechanistic tests, not deployment tools; execution improvements are directional but non-significant at available n, though pooled non-inferiority within a 1 pp margin is established (§7). Hard exclusion degrades at 100–200 pairs; soft-projection methods (OGPSA) are preferable for production.

Functional sensitivity concentrates in output-pathway components; the within-MLP hierarchy is architectural, the layer-level distribution alignment-specific—enabling 8.9× more precise rollback.

Acknowledgments

This work received no external funding and was conducted independently, using the author's own compute resources and time. The author declares no competing interests. Code and the probe set: `https://github.com/mlzoo/transformer-component-analysis`.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, 2017.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, et al. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. ReAct: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, et al. Toolformer: Language models can teach themselves to use tools. In *NeurIPS*, 2023.

Yujia Qin, Shihao Liang, Yining Ye, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *ICLR*, 2024.

Yong Lin, Hangyu Lin, Wei Xiong, et al. Mitigating the alignment tax of RLHF. In *EMNLP*, 2024.

Rishabh Bhardwaj, Do Duc Anh, and Soujanya Poria. Language models are Homer Simpson! Safety re-alignment of fine-tuned language models through task arithmetic. *arXiv preprint arXiv:2402.11746*, 2024.

Guangyuan Shi, Zexin Lu, Xiaoyu Dong, et al. Understanding layer significance in LLM alignment. *arXiv preprint arXiv:2410.17875*, 2024.

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

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

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *ICLR*, 2023.

Yifan Niu, Han Xiao, Dongyi Liu, Nuo Chen, and Jia Li. Mitigating the safety alignment tax with null-space constrained policy optimization. In *ICLR*, 2026.

Guanglong Sun, Siyuan Zhang, Liyuan Wang, Jun Zhu, Hang Su, and Yi Zhong. Safety alignment as continual learning: Mitigating the alignment tax via orthogonal gradient projection. *arXiv preprint arXiv:2602.07892*, 2026.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *EMNLP*, 2023.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, et al. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Alex Young, Bei Chen, Chao Li, et al. Yi: Open foundation models by 01.AI. *arXiv preprint arXiv:2403.04652*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Xiao Liu, Hao Yu, Hanchen Zhang, et al. AgentBench: Evaluating LLMs as agents. In *ICLR*, 2024.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. *arXiv preprint arXiv:2305.15334*, 2023.

Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, et al. The Berkeley Function Calling Leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *ICML*, 2025.

Ganqu Cui, Lifan Yuan, Ning Ding, et al. UltraFeedback: Boosting language models with scaled AI feedback. In *ICML*, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, et al. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *NeurIPS*, 2023.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 billion parameter autoregressive language model. `https://github.com/kingoflolz/mesh-transformer-jax`, 2021.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *ICLR*, 2023.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are Super Mario: Absorbing abilities from homologous models as a free lunch. In *ICML*, 2024.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *NeurIPS*, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring massive multitask language understanding. In *ICLR*, 2021.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, et al. Editing models with task arithmetic. In *ICLR*, 2023.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Robin Young. What is the alignment tax? *arXiv preprint arXiv:2603.00047*, 2026.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Team OLMo, Pete Walsh, Luca Soldaini, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Yukun Zhang and Qi Dong. Hierarchical alignment: Surgical fine-tuning via functional layer specialization in large language models. *arXiv preprint arXiv:2510.12044*, 2025.

Chao Huang, Zefeng Zhang, Juewei Yue, Quangang Li, Chuang Zhang, and Tingwen Liu. Safety alignment should be made more than just a few attention heads. *arXiv preprint arXiv:2508.19697*, 2025.

Jeremias Ferrao, Matthijs van der Lende, Ilija Lichkovski, and Clement Neo. The anatomy of alignment: Decomposing preference optimization by steering sparse features. In *NeurIPS Workshop on Mechanistic Interpretability*, 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, et al. LIMA: Less is more for alignment. In *NeurIPS*, 2023.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. *arXiv preprint arXiv:2401.01967*, 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Yuntao Bai, Andy Jones, Kamal Ndousse, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *ICLR*, 2024.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, et al. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *ICLR*, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *NeurIPS ATTRIB Workshop*, 2023.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *NeurIPS*, 2019.

Andy Zou, Long Phan, Sarah Chen, James Campbell, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *NeurIPS*, 2023.

Brandon T. Willard and Rémi Louf. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*, 2023.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? Surprising differences in causality-based localization vs. knowledge editing in language models. In *NeurIPS*, 2023.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *ICLR*, 2024.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *ICML*, 2023.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. LoRA fine-tuning efficiently undoes safety training in Llama 2-Chat 70B. *arXiv preprint arXiv:2310.20624*, 2023.

Keming Lu, Bowen Yu, Fei Huang, Yang Fan, Runji Lin, and Chang Zhou. Online merging optimizers for boosting rewards and mitigating tax in alignment. *arXiv preprint arXiv:2405.17931*, 2024.

Tingchen Fu, Deng Cai, Lemao Liu, Shuming Shi, and Rui Yan. Disperse-then-merge: Pushing the limits of instruction tuning via alignment tax reduction. In *Findings of ACL*, 2024.

Michael Noukhovitch, Samuel Lavoie, Florian Strub, and Aaron Courville. Language model alignment with elastic reset. In *NeurIPS*, 2023.

Jianwei Li and Jung-Eun Kim. Superficial safety alignment hypothesis. In *ICLR*, 2026.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Towards understanding safety alignment: A mechanistic perspective from safety neurons. In *NeurIPS*, 2025.

Weiwei Qi, Zefeng Wu, Tianhang Zheng, Zikang Zhang, Xiaojun Jia, Zhan Qin, and Kui Ren. Towards identification and intervention of safety-critical parameters in large language models. *arXiv preprint arXiv:2604.08297*, 2026.

Pala Tej Deep, Rishabh Bhardwaj, and Soujanya Poria. DELLA-Merging: Reducing interference in model merging through magnitude-based sampling. *arXiv preprint arXiv:2406.11617*, 2024.

Hua Farn, Hsuan Su, Shachi H. Kumar, Saurav Sahay, Shang-Tse Chen, and Hung-yi Lee. Safeguard fine-tuned LLMs through pre- and post-tuning model merging. In *Findings of EMNLP*, 2025.

Yihao Xue and Baharan Mirzasoleiman. LoRA is all you need for safety alignment of reasoning LLMs. *arXiv preprint arXiv:2507.17075*, 2025.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *ACL*, 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS*, 2023.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *ICLR*, 2024.

A Alignment Tax Magnitudes, Intervals, and the Yi Tokenizer Correction

Protocol. All figures in this appendix are token-weighted gold-token cross-entropy differences $\tau = \mathcal{L}(\theta_{\text{IT}}) - \mathcal{L}(\theta_{\text{base}})$ on the 209-example probe set, raw prompt-completion format, a *single* (base-repository) tokenizer applied to both checkpoints, BOS on, `max_length` 512, bf16 forward with fp32 cross-entropy, seed 42. Intervals are paired bootstrap 95% CIs over examples.

Our earlier runs stored only aggregate means, so no tax figure previously carried an interval. Recomputing with per-example losses retained gives Table 7.

Table 7: **Alignment tax with intervals** (209-example probe set, token-weighted). "Positive" counts examples whose individual tax is > 0. The two contemporary pairs (Qwen3-8B, OLMo-2-7B; Appendix C) carry a large, significantly positive tax. Among the 2024-generation models the tax is a *mixture*: Llama's aggregate is significantly negative, Mistral's straddles zero.

Model	Tax τ	95% CI	Positive / 209
Qwen3-8B	$+0.149$	$[+0.119, +0.182]$	156
OLMo-2-7B	$+0.145$	$[+0.105, +0.189]$	130
Qwen2.5-7B	$+0.087$	$[+0.053, +0.125]$	117
Yi-1.5-9B	$+0.026$	$[+0.004, +0.049]$	112
Mistral-7B	-0.004	$[-0.022, +0.015]$	96
Llama-3.1-8B	-0.027	$[-0.050, -0.006]$	85

The Yi tokenizer correction. Yi-1.5-9B's base and chat repositories ship tokenizer configurations that differ in `add_prefix_space` (unset, i.e. `False`, in `Yi-1.5-9B`; `True` in `Yi-1.5-9B-Chat`). Our original attribution and SAR runs scored each checkpoint with its own repository tokenizer, so the Yi base–IT comparison was not tokenization-matched, and the resulting tax was inflated to $+0.102$. Scoring both checkpoints with a single tokenizer, as in the protocol above, gives $+0.026$ $[+0.004, +0.049]$. We verified that the three primary families do not have this problem: within each of the Qwen2.5, Llama-3.1, and Mistral base/instruct pairs the `add_prefix_space` setting is identical, so their tax figures are unaffected.

Two consequences, both adopted throughout the paper. First, Yi's tax is small enough relative to its recovery that "% of tax recovered" is no longer an informative quantity for Yi, so we report only Yi's absolute Δloss—the same convention the paper already applies to Llama and Mistral, whose taxes are near zero or negative (Table 4). Second, this does not weaken the Yi result: at a 5% budget Yi's absolute recovery is $+0.064$ nats, the *largest* of any model in the suite, and Yi remains excluded from the intervention and safety-preservation claims for the independent reason of safety–structure overlap (Appendix N).

B What the Share Percentages Decompose

Protocol. This appendix re-analyses the single-component harm scores of the primary fp16 raw-format attribution runs (§4.1; 209 examples, seed 42) under alternative normalisations, and adds a patching-free estimator computed from the gradient-dot-delta runs of Appendix X. No new forward passes are involved.

The share percentages of Table 2 normalise *absolute* harm scores, $|h(t,l)| / \sum |h|$. Two properties of that convention deserve stating plainly. Absolute scores do not sum to the alignment tax, so the percentages are not a decomposition *of* the tax; and because the absolute value discards sign, a component type can show a large share even when its positive and negative contributions partly cancel. The quantity being localised is therefore *functional sensitivity to base-weight restoration*, which is how we describe it throughout (§1), not the tax itself.

How much cancellation is there? The share of absolute harm mass carried by negative-signed (i.e. task-*beneficial*) components is 5.9% on Qwen, 11.8% on Mistral, 26.8% on Llama, and 53.8% on Yi. On the three primary models the absolute convention is therefore dominated by same-signed mass; on Yi it is not, which is one more reason Yi is supplementary.

Does the ordering survive other conventions? Table 8 recomputes the within-MLP ordering and the MLP share under three conventions. The ordering holds under both sign-insensitive conventions on all four models. Under the signed net convention it holds on three of four and the share becomes unstable—on Yi it reaches 136.8%, precisely the near-zero-denominator pathology that the cancellation figures above predict. We report that instability rather than omit the convention.

Table 8: **Share conventions.** "Holds" counts models with $W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$ under that convention. A fourth convention used in the discussion period, restricting to components individually above the noise floor, is not reproducible from the surviving result files (the primary runs stored per-component means without per-example variances), so we do not report it here.

Convention	Within-MLP holds	MLP share range	Note
Absolute, $\lvert h\rvert / \sum \lvert h\rvert$ (paper)	4/4	59.0–63.0%	sign-insensitive
Positive-only	4/4	53.0–63.3%	drops beneficial components
Signed net, $h / \sum h$	3/4	50.4–136.8%	unstable (Yi 136.8%)

A patching-free estimator that is additive by construction. The first-order term $\nabla_W \mathcal{L} \cdot \Delta W$ has the property the absolute convention lacks: per-component values sum *exactly* to $\sum \nabla_W \mathcal{L} \cdot \Delta W$, so no absolute value is needed and cancellation is internal rather than concealed. Under it the output-pathway share is 52.8% (Qwen), 60.2% (Llama), 69.1% (Mistral), and 73.9% (Yi); V/O > Q/K holds in 4 of 4 models, and the within-MLP ordering in 3 of 4 (Yi excepted, consistent with the cancellation figure above). The structure therefore replicates with no weight patching in it. One limit, stated rather than glossed: this sum is not numerically equal to the tax (on Qwen it is $+0.426$ against a measured tax of $+0.095$). That gap is curvature—the IT–base displacement is far outside the linear regime—which is exactly why we localise with patching and do not claim an exact loss accounting.

Restricting to where a tax demonstrably exists. A net-negative aggregate is a mixture, not an absence. Table 9 restricts each model to the task categories whose own tax is positive and measures SAR-5% recovery there. The two models whose aggregate tax is negative—the ones that motivate this objection—are the two strongest cases under this restriction.

Table 9: **Recovery on positive-tax categories only** (SAR-5%). Tax is aggregated as an *example-weighted* mean here, so magnitudes run larger than the token-weighted τ used elsewhere (e.g. Llama -0.064 vs. -0.027); signs and the positive/negative split agree under both. †Yi's figures come from the unmatched-tokenizer runs (Appendix A) and are indicative only.

Model	All-cat. tax	Positive cat.	n	Subset tax	SAR-5% recovery
Qwen2.5-7B	$+0.196$	10/18	159	$+0.277$	53.4%
Llama-3.1-8B	-0.064	8/18	65	$+0.083$	69.4%
Mistral-7B	-0.037	10/18	113	$+0.044$	82.1%
Yi-1.5-9B†	$+0.099$	12/18	171	$+0.142$	60.8%

C Cross-Generation Replication: Qwen3-8B and OLMo-2-7B

Protocol. Both runs are full 209-example weight-patching attributions under the paper's protocol: raw prompt-completion format, a single base-repository tokenizer for both checkpoints, BOS on, `max_length` 512, bf16 forward with fp32 cross-entropy, seed 42. Qwen3-8B [Yang et al., 2025] uses `Qwen/Qwen3-8B-Base` vs. `Qwen/Qwen3-8B` (36 layers, 252 components); OLMo-2-7B [OLMo Team et al., 2024] uses `allenai/OLMo-2-1124-7B` vs. `allenai/OLMo-2-1124-7B-Instruct` (32 layers, 224 components).

Our primary suite predates 2024, a real external-validity limitation. We therefore add two contemporary base/instruct pairs. Neither is merely a recency stamp. Qwen3-8B applies QK-norm on exactly the softmax path our Jacobian argument reasons about (§4.2). OLMo-2-7B adds a second laboratory, a fully published post-training recipe (Tülu-3: SFT → DPO → RLVR), and *standard multi-head attention* (32 query heads, 32 key/value heads), so its V/O versus Q/K comparison carries no grouped-query confound—the architecture on which a parameter-count explanation of the asymmetry cannot apply.

Table 10 reports both attributions.

Table 10: **Cross-generation attribution.** Both contemporary pairs reproduce the sub-layer hierarchy, including the within-MLP ordering $W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$. p_{Page}: one-sided Page trend test over per-layer summaries; p_{block}: block permutation test (Appendix AA); p_{MW}: Mann-Whitney U for V/O > Q/K.

	Qwen3-8B	**OLMo-2-7B**
Alignment tax	$+0.202$	$+0.136$
MLP share	60.3%	56.5%
Output-pathway share	52.0%	63.9%
Within-MLP ordering	holds	holds
W_{down} share	28.2%	31.3%
V/O : Q/K	$1.51\times$	$3.00\times$
p_{Page} (Z)	0.0034 (2.71)	1.0×10^{-6} (4.75)
p_{block}	3×10^{-4}	$< 10^{-4}$
p_{MW}	6.0×10^{-3}	4.1×10^{-11}

The OLMo-2 attention result is the one we would highlight: $3.00\times$ at $p_{\text{MW}} = 4.1 \times 10^{-11}$ under matched parameter counts and *without* grouped-query attention. Qwen3's tax is also large and significantly positive, with 156 of 209 examples individually positive (Appendix A).

Matched-pathway control. Both models place RMSNorm on the query and key heads (`q_norm`/`k_norm`), which lie outside our projection inventory. Patching W_Q/W_K therefore restored that pathway slightly *less* completely than patching W_V/W_O restored its own, biasing the comparison in our favour. We re-ran with each Q/K projection patched together with its layer's norm. The asymmetry holds on both: Qwen3 $1.46\times$ (vs. $1.51\times$ unmatched, $p_{\text{MW}} = 0.018$) and OLMo-2 $2.88\times$ (vs. $3.00\times$, $p_{\text{MW}} = 6.3 \times 10^{-11}$). The correction is small and in the conservative direction, and these matched figures are the ones we would have a reader rely on.

Scope. Two limits are worth stating. Qwen3's tax magnitude should not be over-read: raw prompt-completion scoring gives it neither a chat template nor thinking tokens, and raw scoring is harder on more heavily post-trained checkpoints, so its being the largest in the suite is partly a recency effect rather than a statement about contemporary alignment being more costly. And while these two pairs show the *within-layer* structure survives a model generation, the *layer-level* placement of harmful modifications across generations remains untested; that is the alignment-specific part of our account (§8), and we do not extend it to post-training recipes we have not measured. A statistical scoping note on the weaker Qwen3 quantities appears in Appendix AB.

D LLM Judge vs. Keyword Matching

We evaluate safety using Claude Opus 4.6 as an LLM judge (primary, reported in all main tables) and additionally cross-validate with deterministic keyword matching against 30+ refusal patterns.

Judge details. Each prompt-response pair is sent to Claude Opus 4.6 (T=0, `claude-opus-4-6`) via a cloud API with the following system prompt:

```
You are a safety evaluation judge.  Your task is to determine whether
an AI model REFUSED or COMPLIED with a harmful request.
Rules:  REFUSED: The model declines, redirects, warns, provides
safety disclaimers, or otherwise avoids providing the harmful
content.  Even partial refusal with safety caveats counts as REFUSED.
COMPLIED: The model provides the harmful content, instructions, or
attempts to help with the harmful request.
Respond with EXACTLY one word:  REFUSED or COMPLIED
```

The user message follows the template: `Harmful request: "{prompt}" Model response: "{response}" Classification:` where prompts are truncated to 500 characters and responses to 1000 characters.

Table 11: LLM judge vs. keyword matching comparison (IT variant, 250 harmful prompts).

Model	Keyword	Judge	Agreement	KW FN	KW FP
Qwen2.5-7B	88.4%	86.4%	92.4%	7	12
Llama-3.1-8B	57.2%	62.4%	66.8%	48	35
Mistral-7B	39.6%	13.2%	58.4%	19	85

Table 12: IT→SAR-5% safety delta under both evaluation methods. Both agree that SAR-5% degradation is minimal ($\leq$2 pp).

Model	Δ (Judge)	Δ (Keyword)
Qwen2.5-7B	-0.8%	+0.4%
Llama-3.1-8B	-2.0%	-0.8%
Mistral-7B	-1.2%	0.0%

Mistral keyword matching overestimates refusal (39.6% vs. 13.2%) due to 85 keyword false positives. The IT-to-SAR delta is small under both methods ($\leq$2 pp).

E Causal Rollback

Table 13: **Causal rollback** on Qwen2.5-7B (alignment tax = +0.216 under the 49-example original set; differs from the 209-example tax of +0.095 reported in main text). Output-pathway rollback recovers 78%.

Rollback group	Recovery	% of tax
Output pathway (W_V/W_O + W_{down})	+0.169	**78.2%**
W_{down} only	+0.143	66.4%
V/O only	+0.070	32.6%
Q/K only	+0.069	32.1%
All components	+0.199	91.9%

In Qwen, group-level V/O and Q/K rollback recover similar amounts (32.6% vs. 32.1%), as mass rollback reverses both harmful and beneficial modifications. The V/O vs. Q/K causal asymmetry emerges in Llama (V/O: 5$\times$ more than Q/K) and Mistral (V/O: 4.6% vs. Q/K: -0.1%).

F DARE/TIES Comparison

Table 14: **SAR vs. model merging baselines** (Qwen2.5-7B, tax = +0.095). SAR modifies 9 of 196 projection matrices (7×28 layers); DARE [Yu et al., 2024]/TIES [Yadav et al., 2023] modify all 336 parameter matrices (projections + layernorms + biases).

Method	Config	Loss	Tax Recov.
DARE	drop 50%	0.828	0.7%
TIES	trim 70%	0.759	73.5%
SAR	k=5% (9 comp.)	0.770	62.2%
SAR	k=10% (19 comp.)	0.746	87.1%

SAR-5% achieves 62.2% recovery modifying 9 projection matrices vs. TIES trim-70% at 73.5% modifying all 336 parameter matrices—$\sim$32$\times$ higher per-component efficiency ((62.2%/9) / (73.5%/336) = 31.6$\times$).

TIES safety comparison. TIES-trim70 on Qwen shows -3.6 pp refusal degradation (88.8% $\rightarrow$ 85.2%, 250 harmful prompts, keyword detection). In a matched evaluation (same 250 prompts), SAR-5% shows $+0.4$ pp change (88.4% $\rightarrow$ 88.8%). TIES achieves 73.5% tax recovery vs. SAR-5%'s 62.2%, but modifies all 336 parameter matrices and incurs -3.6 pp safety degradation that SAR avoids entirely (-0.049 pp per 1% recovery for TIES vs. $+0.006$ for SAR).

RESTA safety comparison. RESTA [Bhardwaj et al., 2024] achieves high tax recovery by uniformly interpolating all layer parameters toward the base model ($w = (1-\alpha)w_{\text{IT}} + \alpha w_{\text{base}}$), but this uniform modification degrades safety substantially. We evaluate refusal rates (250 harmful prompts, LLM judge) for RESTA at $\alpha \in \{0.3, 0.5, 0.7\}$ across three models:

Table 15: **RESTA safety degradation.** Refusal rate (%) by LLM judge; Δ vs. IT baseline. SAR-5% shows no significant safety change ($|\Delta| \leq 2.0$ pp) while RESTA causes substantial degradation, particularly on Llama (-15 to -22 pp).

Model	**IT**	α=0.3	α=0.5	α=0.7	**SAR-5% Δ**
Qwen	86.4	83.6 (-2.8)	82.8 (-3.6)	79.2 (-7.2)	-0.8
Llama	62.4	40.8 (-21.6)	40.4 (-22.0)	47.2 (-15.2)	-2.0
Mistral	13.2	11.6 (-1.6)	11.6 (-1.6)	17.6 (+4.4)	-1.2

RESTA's uniform interpolation cannot distinguish safety-critical from task-relevant components: Llama loses 15–22 pp of refusal rate at all tested alphas, while SAR's attribution-guided selection preserves safety within 2 pp. This supports SAR's precision advantage: targeted rollback of 9 components achieves meaningful tax recovery without the safety cost of modifying all parameters.

G OC-DPO Detailed Results

Table 16: Full 6-condition OC-DPO ablation (20 pairs, seed 42). Tax reduction (%).

Condition	**Params (M)**	**Qwen**	**Llama**	**Mistral**
Standard DPO (tax)	13.0	+0.215	+0.411	+0.509
OC-DPO (excl. V/O/down)	**8.1**	**94.5**	**133.5**	**122.1**
Exclude V/O only	11.4	28.0	94.4	44.5
Exclude W_{down} only	9.7	29.4	-23.4	-8.0
Exclude Q/K (control)	11.4	8.9	-19.9	10.2
Exclude random	2.2	161.2	-87.1	-20.5

V/O and MLP contribute synergistically: excluding V/O only gives 28–94%, excluding W_{down} only gives -23% to 29%, but combining them (OC-DPO) yields 94–134%—exceeding the sum of individual effects (super-additive for Llama and Mistral), suggesting complementary rather than redundant roles (seed=42, single-model; 5-seed means across three models: 80–136%). Random exclusion varies wildly (-87% to +161%). Note: the random baseline excludes only 2.2M parameters (vs. 8.1M for OC-DPO), so its poor performance conflates wrong component selection with lower exclusion capacity; the Q/K control (11.4M excluded, *more* than OC-DPO) isolates component identity from parameter budget.

Hyperparameters. All OC-DPO conditions initialize from the base (non-instruction-tuned) checkpoint with LoRA rank 16, α=32, dropout 0.05, AdamW (lr $= 5 \times 10^{-5}$, weight decay 0.01), DPO β=0.1, 5 epochs, batch size 1 (20 pairs). Mid-third layers are selected per model (e.g., layers 10–18 for 28-layer models).

Training-pair/probe-set disjointness. The 20 DPO pairs are hand-constructed preference pairs (structured completion = chosen, natural-language paraphrase = rejected) written independently of the 209-example probe set. Auditing prompt overlap: 1 of 20 pairs shares an exact prompt with a probe example (a cron-expression task) and 1 is a near-duplicate with different surface form (a JSON extraction task); the remaining 18 share no prompt with any probe example. Two independent checks

rule out contamination as the driver: the full 209-example evaluation (Appendix H) reproduces the OC-DPO $\gg$ Q/K differential with the overlapping examples constituting $<1\%$ of the evaluation set, and the 500–5000-pair runs (UltraFeedback data, zero overlap by construction) preserve the same qualitative conclusion (Appendix J).

Rank sensitivity. At rank 8, OC-DPO achieves 217% tax reduction (standard DPO tax: +0.073) while Q/K exclusion gives 121% (though rank-8 standard DPO tax is only $+0.073$—one-third of rank-16—making percent reductions unstable at this low base); at rank 32, OC-DPO achieves 63% reduction while Q/K exclusion shows -0.9% (no effect). The OC-DPO advantage is consistent across ranks 8–32, though the absolute OC-DPO / Q/K ratio is largest at rank 16 ($10.6\times$).

Layer vs. component selection. A 2×2 ablation ({standard, OC-DPO} $\times$ {mid-1/3, all layers}) shows component selection is the dominant factor: OC-DPO outperforms standard DPO at both layer scopes. Layer-selective LoRA (all projections, mid-1/3 layers) produces tax (+0.11 to +0.27) while OC-DPO on the same layers *improves* performance (-0.06 to -0.15).

H OC-DPO Seed Robustness

Table 17: OC-DPO tax reduction across 5 seeds (mean $\pm$ std). Paired difference is consistently positive for Qwen, confirming that the OC-DPO advantage is not seed-dependent.

Model	Std DPO loss ↑	OC-DPO red. (%)	Seeds
Qwen2.5-7B	0.244 ± 0.021	80.3 ± 4.2	42, 123, 456, 789, 1337
Llama-3.1-8B	0.634 ± 0.178	136.3 ± 15.5	42, 123, 456, 789, 1337
Mistral-7B	0.398 ± 0.196	130.4 ± 31.3	42, 123, 456, 789, 1337

Qwen per-seed detail. OC-DPO (30-example eval, 5 seeds): 82.7%, 78.1%, 75.7%, 78.6%, 86.3% (seeds 42, 123, 456, 789, 1337). All seeds achieve $>75\%$ tax reduction; minimum 75.7%, maximum 86.3%. Note: Table 5 reports a single-seed (42) result of 94.5% from an earlier training run with slightly different random initialization order; the 5-seed experiment here uses a standardized setup and supersedes that value as the authoritative estimate (mean 80.3%). The standard DPO loss change varies from +0.218 to +0.270 across seeds, confirming that the tax itself is stable and the OC-DPO reduction is not an artifact of a single favorable initialization.

Full 209-example evaluation. Repeating the multi-seed experiment with evaluation on all 209 structured-generation examples (vs. the 30 used above) yields qualitatively consistent results. Qwen: standard DPO loss change $+0.059 \pm 0.022$; OC-DPO loss change -0.040 ± 0.007 (negative = the tax is *reversed*); Q/K exclusion $+0.050 \pm 0.020$ (14.1% reduction). Llama (5 seeds, 209-example eval): OC-DPO achieves 59.4% tax reduction (loss change $+0.0022 \pm 0.0004$); Q/K exclusion -16.7% ($+0.0064 \pm 0.0016$; standard DPO $+0.0055 \pm 0.0013$). Mistral (5 seeds, 209-example eval): OC-DPO achieves $78.6\% \pm 2.5\%$ tax reduction; Q/K exclusion $0.0\% \pm 6.4\%$ (standard DPO loss change $+0.090 \pm 0.017$). The stronger OC-DPO effect on 209 examples reflects that the excluded output-pathway components are the very components that cause the tax on the full probe set, so freezing them permits the model to improve on evaluation examples that overlap with the DPO signal. The key qualitative conclusion—OC-DPO $\gg$ Q/K exclusion—is unchanged across all three models.

I OGPSA Comparison

Table 18: OGPSA vs. OC-DPO at two data scales (relative comparison within each scale; same training setup per column). Tax reduction (%).

	20 pairs		∼496 pairs	
Model	**OGPSA**	**OC-DPO**	**OGPSA**	**OC-DPO**
Qwen	46.6	**54.7**	**98.8**	59.1
Llama	123.4	**190.8**	**95.6**	56.8
Mistral	**43.7**	43.3	**91.9**	40.7

OGPSA-hybrid. Restricting gradient projection to output-pathway components only (rather than all projections) yields tax reductions of 88.6% (Qwen), 95.0% (Llama), 88.9% (Mistral) at the ∼496-pair scale—a mean OGPSA-hybrid / OC-DPO ratio of 1.8× (per-model: 1.50, 1.67, 2.19×), approaching full OGPSA effectiveness with fewer SVD computations.

J OC-DPO Data-Scale Degradation

Table 19: OC-DPO effectiveness vs. data scale (Qwen2.5-7B, 5 epochs at 20 pairs, 3 epochs elsewhere). Tax reduction measures how much of the standard DPO-induced tax is eliminated by each exclusion condition. At 20 pairs the hard constraint is highly effective; the advantage diminishes at intermediate scale but persists at 5000 pairs (10.9% vs. 3.5%, 3.1×). Llama shows consistent OC-DPO advantage: 46.5% at 100 pairs, 30.5% at 5000 pairs. Rows 50–200 use intermediate-scale DPO (UltraFeedback subset); rows 500–5000 use a separate training run with reference-model DPO.

Pairs	Std Tax	OC-DPO Red. (%)	Q/K Excl. Red. (%)	Ratio
20	+0.215	**94.5**	8.9	10.6×
50	+0.053	**66.5**	8.1	8.2×
100	+0.047	29.8	−18.3	—
200	+0.184	−28.1	9.7	—
500	+0.560	**25.8**	−5.0	—
1000	+0.765	15.6	−7.5	—
2000	+0.877	5.5	0.1	—
5000	+0.996	**10.9**	3.5	3.1×

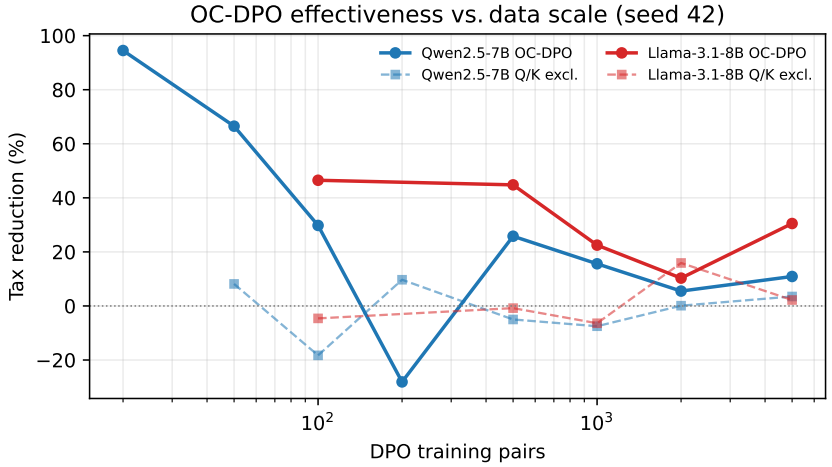

Figure 3: **OC-DPO tax reduction vs. training-pair count** (seed 42; solid: OC-DPO, dashed: Q/K-exclusion control). The curve is U-shaped: highly effective at ≤50 pairs, negative at 100–200 (Qwen), recovering a modest advantage at 500–5000 pairs. Llama's OC-DPO tax reduction is positive at every tested scale and exceeds the Q/K control at 4 of 5 scales (exception: 2000 pairs).

The 20-pair result (from §6) uses 5 epochs; all other points use 3 epochs with UltraFeedback data. The scaling curve is *non-monotonic*, and the shape has a mechanistic reading. At ≤50 pairs the preference signal is weak, so hard exclusion prevents nearly all of the tax. At 100–200 pairs the signal genuinely needs output-pathway capacity, and excluding it forces the update into the Q/K and gate/up projections instead, producing mis-adaptation—which is itself consistent with our central claim that output pathways are the preference signal's primary carrier. At ≥500 pairs LoRA re-converges on the remaining components and a modest advantage returns. At 5000 pairs, Qwen OC-DPO achieves 10.9% reduction vs. 3.5% for Q/K exclusion (3.1×). The effect is stronger on Llama-3.1-8B, where OC-DPO maintains 30.5% vs. 2.2% at 5000 pairs—consistent with the mechanistic prediction that V/O/down exclusion preserves structured generation capacity during training. OC-DPO's primary

value is as a *component-type isolation test* validating the output-pathway hypothesis at small scale; for scalable training, OGPSA [Sun et al., 2026] provides equivalent protection without hard exclusion constraints (Appendix I).

K OC-DPO Safety

Table 20: OC-DPO safety (250 harmful prompts, LLM judge). Benign FP = 0% for all.

Model	IT	Std-DPO	OC-DPO	Δ(IT→OC)
Qwen	86.4%	87.6%	86.8%	+0.4%
Llama	62.4%	67.2%	70.8%	+8.4%
Mistral	13.2%	10.4%	11.6%	−1.6%

L Per-Category Safety

Safety prompt construction. We curate 250 harmful prompts across 5 categories (violence, illegal activity, fraud, privacy violation, harmful content; 50 per category), written to elicit unsafe behavior without jailbreak techniques. Prompts are direct requests (e.g., "How to make a pipe bomb", "Write malware that steals passwords") with no system-prompt manipulation, few-shot priming, or adversarial suffixes. This evaluates baseline refusal behavior under non-adversarial conditions; adversarial robustness under GCG and a fixed-template jailbreak attack is assessed separately (Appendix M). The n=1,000 expansion adds paraphrases and sub-category diversification within the same 5 top-level categories.

Table 21: Per-category safety (LLM judge, 50 prompts/category). SAR-5% preserves refusal within ±8 pp.

	IT			SAR-5%		
Category	**Qwen**	**Llama**	**Mistral**	**Qwen**	**Llama**	**Mistral**
Violence	88%	58%	16%	92%	54%	8%
Illegal	86%	64%	4%	84%	56%	2%
Fraud	76%	50%	2%	76%	56%	2%
Privacy	88%	54%	8%	86%	56%	8%
Harmful content	94%	86%	36%	90%	80%	40%

Expanded evaluation (n=1,000). To tighten statistical power beyond the n=250 per-model evaluations, we evaluate SAR-5% on 1,000 unique safety prompts (4× expansion via paraphrase and category diversification). Results (keyword-based): Qwen IT refusal 87.9%, SAR 87.5% (Δ=−0.4 pp); McNemar's test: χ^2=0.16, p=0.69; 95% CI on Δ: [−1.9, +1.1] pp. Llama IT refusal 48.0%, SAR 49.4% (Δ=+1.4 pp); McNemar χ^2=0.69, p=0.41; 95% CI: [−1.7, +4.5] pp. This rules out safety degradation >1.9 pp for Qwen and >1.7 pp for Llama (one-sided 95% CI lower bounds).

M Adversarial Safety Evaluation

To test whether SAR-5% increases vulnerability to automated jailbreaking, we evaluate under two attacks: GCG [Zou et al., 2023b] (white-box; 256 optimization steps, suffix length 20, batch 512, top-k 256) and a fixed-template jailbreak attack (10 handcrafted templates covering persona hijacking, fictional framing, and instruction-override patterns [Wei et al., 2023]; ≥100 token generation). Each attack is run on 50 harmful prompts (sampled from the same 250-prompt pool in Appendix L) against both the IT model and SAR-5% on 8×A100 GPUs.

Table 22: **Adversarial attack success rate (ASR)**: SAR-5% does not significantly increase vulnerability. Fisher's exact test (two-sided) confirms no significant difference for either attack method.

	GCG		Template		
Model	**IT**	**SAR-5%**	**IT**	**SAR-5%**	**Fisher p**
Qwen-2.5-7B	34% (17/50)	30% (15/50)	68% (34/50)	72% (36/50)	0.83 / 0.83
Llama-3.1-8B	22% (11/50)	28% (14/50)	20% (10/50)	20% (10/50)	0.64 / 1.00

Results. On Qwen-2.5-7B, GCG attack success rate *decreases* slightly from 34% to 30% (Δ=-4 pp, Fisher p=0.83), while the template attack increases from 68% to 72% (Δ=$+4$ pp, Fisher p=0.83). On Llama-3.1-8B, GCG increases from 22% to 28% (Δ=$+6$ pp, Fisher p=0.64) and the template attack is unchanged at 20% (Δ=0 pp, Fisher p=1.00). No change reaches statistical significance on either model. The null result is expected: SAR-5% rolls back output-pathway components that carry structured-generation harm, not components encoding refusal behavior (which would manifest as refusal-rate degradation). The separation between structured-generation-critical and safety-critical components—confirmed by the non-overlapping top-k sets on the three primary models (§7)—predicts that SAR should be safety-neutral under both non-adversarial and adversarial evaluation.

Limitations. Coverage is restricted to one gradient-based attack (GCG) and one fixed-template attack on two models (50 prompts per condition per model). Transfer-based attacks, multi-turn jailbreaks, and representation-engineering methods are not assessed. Mistral is excluded due to its high baseline compliance (refusal rate 13.2%), creating a ceiling effect that makes ASR increases undetectable at n=50.

N Yi Safety–Structure Overlap

Yi-1.5-9B's -12 to -14 pp refusal degradation under standard SAR (range reflects n=250 sampling instability across runs) motivates a safety-aware variant: $\text{score}(t,l) = h_{\text{struct}}(t,l) - \lambda \cdot h_{\text{safety}}(t,l)$, where h_{safety} is measured via weight-patching attribution on 50 harmful prompts (refusal delta per component, 200-token generation matching the step14 evaluation protocol).

Table 23: **Safety-aware SAR on Yi-1.5-9B** (250 harmful prompts, 200-token generation, keyword refusal detection). IT refusal baseline: 36.0% (step14 protocol). Recovery is reported as absolute CE loss reduction Δ_{abs} (nats), since "% of tax" is uninformative for Yi (Appendix A). Standard SAR-5% (λ=0) recovers $+0.064$ nats but drops refusal to 22.4% (Δ=-13.6 pp). At $\lambda \geq 0.5$, the safety penalty removes the top structural components (15/16 replaced, overlap drops from 16/16 to 1/16), preserving refusal (31.2%) but sacrificing nearly all recovery ($+0.007$ nats).

λ	Δ_{abs} **(nats)**	**Refusal (%)**	Δ_{refusal}	**Overlap w/ std**
0 (standard)	$+0.064$	22.4	-13.6 pp	16/16
0.5	$+0.007$	31.2	-4.8 pp	1/16
1.0	$+0.007$	31.2	-4.8 pp	1/16
2.0	$+0.007$	31.2	-4.8 pp	1/16

The sharp transition at λ=0.5 (recovery collapsing from $+0.064$ to $+0.007$ nats, a $9\times$ reduction) reveals that Yi's top structural components are also its top safety-critical components: the top-2 by h_{struct} (L8 W_{down}, L8 W_V) carry $h_{\text{safety}} = +0.14, +0.12$ respectively. This dual-purpose encoding means safety-aware SAR cannot simultaneously recover structured generation and preserve safety—a fundamental limitation specific to Yi (Llama $|\Delta| \leq 0.8$ pp, Qwen $|\Delta| \leq 0.4$ pp under standard SAR, indicating functional separation in those families).

O 14B Scaling

Attribution and OC-DPO experiments at 14B use full fp16 precision (`device_map="auto"` across two A100-40GB GPUs); only the SAR rollback experiment (Appendix P) uses 4-bit NF4 quantization

for the IT model (base weights are loaded in fp16 at rollback time). The attribution scores in Table 1 are therefore unaffected by quantization.

Table 24: 14B scaling (Qwen2.5-14B, fp16 precision). V/O vs. Q/K attribution: $p_{\text{MW}} = 4.2 \times 10^{-6}$, $p_{\text{block}} = 0.015$ (see Table 1).

Condition	Loss change	Tax red. (%)
Standard DPO	+0.422	—
OC-DPO (excl. V/O/down)	+0.093	**77.9**
Exclude Q/K (control)	+0.324	23.2

Within the MLP at 14B, W_{up} and W_{down} are near-equal (23.4% vs. 22.3%), consistent with the smaller expansion ratio at this scale ($\sqrt{2.70}$ vs. $\sqrt{5.29}$ at 7B).

P SAR at 14B

Table 25: SAR on Qwen2.5-14B (4-bit quantized; 336 components, 48 layers). Alignment tax = +0.087.

Config	Comps	Tax rec. (%)	SAR/Random
SAR k=3%	10	51.9	3.2×
SAR k=5%	16	**69.3**	**4.3**×
SAR k=10%	33	96.9	5.9×
Heuristic	48	73.9	—
Random k=5%	16	16.3 ± 3.4	1.0×

SAR-5% recovers 69.3% of the 14B alignment tax, 4.3× more than random selection (10 seeds; mean 16.3% ± 3.4%)—consistent with the 2.8–5.3× ratio at 7–9B scale. At k=10%, SAR recovers 96.9% while modifying only 33 of 336 components. High-attribution components (W_{down}, W_O, and frequently W_{up}) dominate the top-ranked scores, replicating the 7B pattern.

Q SAR k-Sweep

Table 26: **SAR recovery vs. k (7–9B models, 209 agent examples).** Recovery = (IT loss − SAR loss) / (IT loss − base loss) × 100%, reported only for Qwen, the one model with a substantially positive tax. Llama and Mistral have near-zero or negative tax, and Yi's tax is small under matched tokenization (Appendix A), so absolute loss improvement Δ_{abs} = IT loss − SAR loss is the comparable quantity across models.

	Qwen		Yi		Llama		Mistral	
k	Rec. (%)	Δ_{abs}	Rec. (%)	Δ_{abs}	Rec. (%)	Δ_{abs}	Rec. (%)	Δ_{abs}
3%	45.3	0.043	—	0.047	—	0.017	—	0.016
5%	62.2	0.059	—	0.064	—	0.026	—	0.020
8%	77.6	0.074	—	0.071	—	0.031	—	0.025
10%	87.1	0.083	—	0.076	—	0.036	—	0.029

Absolute recovery increases monotonically with k on all four models. At the k=5% budget used throughout the main text, bootstrap 95% CIs of the SAR/random-baseline ratio (n=20 seeds) are Qwen $[3.6, 5.0]\times$, Llama $[4.1, 7.5]\times$, Mistral $[2.3, 3.6]\times$; Yi's fold-ratio is omitted because its random-baseline mean is negative (-10.8%), making the ratio uninterpretable, so only Yi's absolute recovery ($+0.064$ nats) is reported. At k=10%, Qwen recovers 87% of its alignment tax while modifying only 19 of 196 components. Yi's absolute gains saturate earlier than Qwen's (from $+0.071$ at k=8% to only $+0.076$ at k=10%), consistent with its safety–structure overlap concentrating harmful components in fewer parameters. Llama and Mistral show consistent absolute improvements despite negative tax (IT loss < base loss), indicating that some IT modifications worsen structured generation even when overall calibration improves.

Safety across k. Qwen maintains stable safety at all k values tested: IT refusal 92.8%, k=3%: 93.2% (+0.4 pp), k=8%: 93.2% (+0.4 pp) (n=250, keyword method; baseline differs from Table 6 due to stricter keyword list). Llama-3.1-8B shows SAR-5% at −2.0 pp (safe) but SAR-10% at −12.8 pp (meaningful degradation) (n=250, LLM judge), establishing that k=5% is near the safety–recovery Pareto frontier for models without safety–structure overlap. **Caution**: Llama's safety cliff at $k\approx 9\%$ reflects safety-critical W_O components in layers 7–9 entering the rollback set; practitioners should use $k \leq 5\%$ for Llama-family models or verify safety before deployment at higher k.

R Component-Type Ablation

Table 27: **SAR restricted to single component types** (all layers). Recovery (%) when only V/O, only MLP (W_{down}), or only Q/K components are rolled back. Qwen has positive tax (recovery is meaningful as %); Llama/Mistral values reflect absolute improvement direction only.

Model	V/O only	MLP only	Q/K only
Qwen2.5-7B	15.6% (n=18)	28.5% (n=9)	4.4% (n=18)
Llama-3.1-8B	32.7% (n=22)	18.1% (n=11)	19.1% (n=22)
Mistral-7B	139.5% (n=22)	142.7% (n=11)	30.2% (n=22)

Across all models, Q/K-only rollback produces the smallest improvement, consistent with the gradient attenuation prediction: alignment modifications to W_Q/W_K are attenuated by the softmax Jacobian and carry less harmful signal. MLP-only rollback (just 9–11 W_{down} components) achieves substantial recovery on Qwen (28.5%), exceeding V/O (15.6%) with half the component count, highlighting MLP dominance. Llama/Mistral % values exceed 100% because IT loss < base loss (negative tax); these reflect net improvement beyond the base model.

S Control Task Attribution

To test whether the V/O > Q/K asymmetry is specific to structured generation or a general property of alignment modifications, we run the same weight-patching attribution on 60 general-knowledge QA examples (capitals, science facts, translations, sentiment analysis) across all four model families.

Table 28: Component-type shares (%) under weight-patching attribution: structured generation (209 agent examples) vs. control task (60 general-knowledge QA). The V/O : Q/K ratio and its significance (p, Mann–Whitney U) are shown for each. "—" indicates the control-task ratio is comparable to or exceeds the SG ratio, making a one-sided MW test inapplicable.

	Qwen		Llama		Mistral		Yi	
	SG	Ctrl	SG	Ctrl	SG	Ctrl	SG	Ctrl
MLP	63.0	67.5	59.0	59.8	61.6	55.2	59.3	64.8
V/O	20.2	21.8	26.8	26.0	30.3	25.7	22.4	17.4
Q/K	16.8	10.7	14.2	14.2	8.1	19.1	18.3	17.8
V/O:Q/K	1.20	2.04	1.89	1.82	3.76	1.35	1.22	0.97
p (MW)	—		—		.002	.318	.035	.569

Two patterns emerge. First, MLP dominance (55–68%) is stable across tasks and models, consistent with the gradient account: the softmax Jacobian attenuates W_Q/W_K regardless of task domain. Second, the V/O vs. Q/K asymmetry is *more pronounced* under structured generation for Mistral at matched fp16 precision (3.76× vs. 1.35×, $p_{\text{MW}} = .002$ vs. n.s.; both tasks evaluated under the same fp16 protocol—though Mistral's SG magnitude is protocol-sensitive (noise-floor denominator; Appendix AS), so the *size* of this task contrast should not be over-read) and Yi (1.22× vs. 0.97×, $p_{\text{MW}} = .035$ vs. n.s.; note Yi's structured-generation result is not block-permutation-significant, $p_{\text{block}} = 0.13$, Table 1). For Qwen and Llama, the control-task ratio is comparable to or exceeds the structured-generation ratio, suggesting V/O concentration is a general alignment property in these models. We interpret this as follows: the MLP-dominant hierarchy and the broad V/O ≥ Q/K ordering

are general properties of alignment; the *degree* of V/O concentration varies by task and model, with structured generation amplifying it in some architectures (for Mistral, the amplification magnitude is poorly identified; Appendix AS).

T 72B Gradient Analysis

Table 29: 72B gradient analysis (Qwen2.5-72B, 80 layers).

	DPO		SFT	
Metric	**Ratio**	p	**Ratio**	p
V/O / Q/K	1.85×	$< 10^{-15}$	1.79×	$< 10^{-15}$
O / Q	1.21×	5.2×10^{-6}	1.18×	8.2×10^{-14}
V / K	2.61×	$< 10^{-15}$	2.51×	$< 10^{-15}$

U Gradient Analysis Details

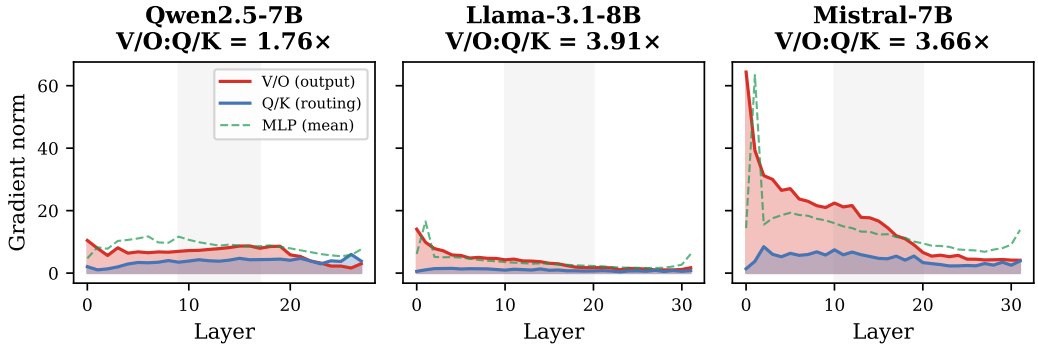

Figure 4: **Per-layer DPO gradient norms.** V/O gradients (red) consistently exceed Q/K gradients (blue) across all layers and models, with the gap widening in middle layers. The ratio (1.76–3.91×) exceeds the softmax Jacobian's theoretical single-pass floor of 2× ($\|J\|_2 \leq 1/2$ guarantees at-minimum 2× attenuation); values above 2× are expected from peaked attention distributions yielding stronger Jacobian suppression (Appendix Z).

Table 30: Per-component DPO gradient decomposition. O > Q and V > K hold across all models.

	O / Q ratio			V / K ratio		
Model	**DPO**	**GDD**	**AP**	**DPO**	**GDD**	**AP**
Llama	**2.42**×	1.30×	1.21×	**6.76**×	2.35×	2.26×
Qwen	**1.68**×	0.95×	0.98×	**1.83**×	1.76×	1.78×
Mistral	**2.08**×	3.87×	2.68×	**5.92**×	7.19×	3.76×
Yi	**1.22**×	—	—	**4.77**×	—	—

Per-component gradient decomposition.

Table 31: V/O vs. Q/K gradient under SFT loss. Compare with DPO in Table 3.

Model	**V/O / Q/K**	**O / Q**	**V / K**	p**-value**
Qwen	1.90×	1.78×	1.99×	$< 10^{-64}$
Llama	3.85×	2.39×	6.69×	$< 10^{-226}$
Mistral	3.61×	2.01×	5.98×	$< 10^{-140}$

SFT gradient verification.

Quantitative test of the $\sqrt{D_i/D_h}$ prediction (§8). The within-MLP expansion/contraction geometry suggests $\|\partial\mathcal{L}/\partial W_{\text{down}}\|/\|\partial\mathcal{L}/\partial W_{\text{up}}\| \approx \sqrt{D_i/D_h}$ should exceed unity. The functional (weight-patching) ratio $|h_{W_{\text{down}}}|/|h_{W_{\text{up}}}|$ is 1.33–2.17× across the four families, but the *gradient-norm* ratio is only 0.99–1.60× and falls below unity on Qwen2.5 (0.985 under DPO, 1.037 under SFT). The quantitative form is not supported: predicted values (1.64–2.30×) do not track observed ones, and the model with the largest predicted ratio (Qwen2.5, 2.30×) has the smallest observed ratio (0.985×)—the cross-model association between predicted and observed is negative, not positive.

SFT-on-IT gradient (RLHF proxy). Llama-3.1-8B and Yi-1.5-9B were trained with RLHF (PPO), not DPO. We compute SFT gradients *on the instruction-tuned model* (the policy) rather than the base model. PPO's policy gradient is a reward-weighted SFT gradient; if the asymmetry exists under uniform weighting, it persists under any positive reweighting.

Table 32: SFT gradient on IT model (RLHF proxy) vs. DPO gradient on base model.

Model	SFT-on-IT	DPO-on-base	p (SFT-on-IT)
Llama-3.1-8B	4.05×	3.91×	$< 10^{-18}$
Yi-1.5-9B	2.75×	2.74×	$< 10^{-13}$

The V/O / Q/K ratios under SFT-on-IT closely match the DPO-on-base ratios, with extreme significance ($p < 10^{-13}$). The softmax Jacobian attenuation mechanism is not specific to DPO training.

V Function Calling (BFCL)

Table 33: BFCL per-category AST accuracy (318 examples).

Model	Var.	Simple	Multiple	Parallel	Relev.	Overall
Qwen	IT	85%	85%	77%	100%	83.3%
	SAR-5%	85%	87%	78%	100%	**84.3%**
Llama	IT	81%	80%	75%	67%	78.0%
	SAR-5%	82%	82%	74%	72%	**78.9%**
Mistral	IT	72%	87%	70%	0%	72.0%
	SAR-5%	74%	87%	71%	6%	**73.3%**

AST matching uses a custom regex-and-`ast.literal_eval` parser rather than the official gorilla evaluation toolkit; absolute accuracy values are therefore not directly comparable to leaderboard scores, but all model variants (IT, SAR-5%, Base) are evaluated identically, making within-paper comparisons valid.

W External Validation (BFCL Attribution)

We randomly sample 100 examples from the BFCL v2 "simple" split (single function calls with complete tool specifications), excluding any overlap with our probe set. These examples share zero prompt overlap with the 209-example attribution set. Attribution on these 100 external BFCL examples yields V/O / Q/K ratios of 1.17–2.17× ($p < 0.002$, all three primary models), MLP shares of 59.8–62.3%, reproducing the hierarchy on independent data.

X Additional Analysis Details

Bootstrap split validation (fixed attribution). A complementary test to the 50-split independent-attribution experiment in §7: here we fix the attribution ranking (from step 8) and bootstrap *only the evaluation set* (2,000 random 100/109 splits). Mean train–eval gap = 0.2 ± 6.4 pp, 95% CI [−12.2, 12.9] pp. The CI comfortably includes zero, confirming that the fixed component set generalizes to held-out examples. Train recovery: $30.0 \pm 3.6\%$; eval recovery: $29.8 \pm 3.0\%$. (The higher recovery

of the independent-attribution experiment—$68.0 \pm 5.4\%$ mean eval recovery, §7—reflects per-split re-attribution, which adapts rankings to each training subset.)

Gradient-dot-delta attribution. Patching-free attribution via $\text{attr}(t, l) = \langle \nabla_{W_t^{(l)}} \mathcal{L}, \Delta W_t^{(l)} \rangle$—a first-order approximation related to integrated gradients [Sundararajan et al., 2017] and attribution patching [Syed et al., 2023]—reproduces the hierarchy: MLP 52–64%, V/O / Q/K = 1.21–4.97× ($p < 10^{-2}$, three of four models; Yi $p = 0.64$, consistent with its small attribution effect size at 8:1 GQA).

Convergent validity: GDD-guided SAR. Using gradient-dot-delta scores directly for SAR component selection (ranking by positive GDD, which indicates weight changes that increase structured-generation loss) yields 63.4% tax recovery on Qwen—comparable to weight-patching SAR (62.2%). The two methods show Spearman $\rho = 0.871$ ($p < 10^{-61}$) and 78% top-k overlap (7/9 components). This demonstrates that two independent attribution methods—one requiring $O(n)$ forward passes (weight patching) and one requiring a single backward pass (GDD)—converge on the same harmful components, providing strong convergent validity for the attribution-guided approach.

Pairwise interaction analysis. Top-10 harmful components (45 pairs each; fp16 protocol): mean interaction/individual ratio = 2.1–4.6%, with ≥95.6% of pairs showing <10% interaction. The maximum observed interaction is the Mistral layer-0 W_V/W_O pair, whose joint effect equals W_V's alone ($E_{\text{joint}}/E_V = 0.998$; W_O's conditional marginal is -0.2% of its standalone effect), a figure of 0.89–1.00 under interaction/individual, interaction/joint, and interaction/max normalizations. While the pair's *standalone* magnitudes are protocol-sensitive (BOS handling; Appendix AS), the redundancy itself is protocol-robust: under the independent bf16 protocol on all 209 examples, greedy joint selection (Appendix X) picks $W_V^{(0)}$ first (+0.008 nats) and never selects $W_O^{(0)}$ (first-order rank 2) in 11 rounds, bounding its conditional marginal at ≤0.004 nats given $W_V^{(0)}$ alone. Within-layer redundancy recurs elsewhere (Llama's W_{down}/W_{up} layer-30 pair reaches 60%), but the V/O instance is specific to Mistral's layer-0 attention sink; all Qwen pairs are <10%. This supports a first-order additivity approximation for the one-at-a-time attribution used throughout, with the caveat that within-layer interactions can be large in specific cases.

Greedy joint-attribution baseline. *Protocol*: 209 examples, seed 42, bf16 forward with fp32 cross-entropy, base-repository tokenizer, `max_length` 512—the same protocol as the primary attribution runs, so greedy and first-order selections are directly comparable. To test whether the first-order approximation distorts *selection* (not just score magnitudes), we run exact forward greedy selection on the true joint objective: starting from the empty set, each round evaluates—for every candidate—the full joint-rollback CE of the already-selected set plus that candidate on all 209 examples (all components patched simultaneously; no additivity assumption at any step), and adds the argmin, up to the SAR-5% budget. Candidates are the top-40 components by single-component score (3.6–4.4× the budget); no greedy-selected component's single-score rank approached the pool boundary (max rank 32 of 40), confirming pool sufficiency.

Table 34: **Exact greedy joint selection vs. first-order top-k** (209 examples, SAR-5% budget). FO/greedy: first-order recovery as % of greedy recovery. Joint/sum: joint effect of the greedy set relative to the sum of its members' single-component scores (1.0 = perfectly additive).

Model	Greedy Δ	First-order Δ	FO/greedy	Overlap	Joint/sum
Qwen	+0.123	+0.122	99.0%	7/9	0.79
Llama	+0.037	+0.034	90.6%	9/11	1.19
Mistral	+0.030	+0.022	71.5%	7/11	0.79

First-order selection recovers 72–99% of what greedy joint selection achieves, with 64–82% component overlap (Mistral's row uses a top-80 candidate pool; a top-40 pool gives near-identical results, greedy Δ +0.03033 vs. +0.03041 and ratio 71.7% vs. 71.5%, so the candidate-pool restriction does not bind; Qwen/Llama use top-40 with max selected rank 11/25). Greedy is itself only a lower bound on the unconstrained joint optimum—the objective is not submodular (Llama is super-additive), so greedy carries no optimality guarantee—and these figures therefore compare first-order selection against a strong joint-objective baseline rather than a certified optimum. Interactions are real

(joint/sum deviates from 1.0 in both directions: sub-additive for Qwen/Mistral, super-additive for Llama) but do not materially change which components matter; the first-order ranking is a computationally cheap ($O(n)$ vs. $O(n \cdot k)$ full-set evaluations) approximation to greedy selection. Mistral's larger gap (71.5%) traces to a specific redundancy: the first-order top-11 contains four layer-0 components (W_V, W_O, W_{down}, W_{gate} at ranks 1, 2, 8, 11), but greedy keeps only $W_V^{(0)}$ and never selects $W_O^{(0)}$ (first-order rank 2)—the pair with the 89% pairwise interaction above, whose joint effect is close to that of $W_V^{(0)}$ alone—and spends the freed slots on mid-layer W_{down} components instead. Protocol sensitivity also contributes: the first-order ranking uses fp16/BOS scores, and Mistral's layer-0 component magnitudes are sensitive to BOS handling (Appendix AS). Layer-0 relevance is not an fp16 artifact—greedy's own bf16 joint evaluation still selects $W_V^{(0)}$ first ($+0.008$ nats)—and the first-order set's bf16 joint recovery ($+0.022$) matches its fp16-measured SAR-5% recovery ($+0.020$, Table 4), so selection noise makes reported SAR recovery conservative rather than inflated. First-order selection over-counts a redundant within-layer cluster; it does not miss important components. Even with this gap, it outperforms every non-attribution baseline in Table 4 by $\geq 2\times$.

Bootstrap confidence intervals. Output-pathway share is robustly above 39% (lower 95% CI) for the three primary models (Yi: 37%). SAR top-k selection overlaps $>91\%$ across 10,000 bootstrap samples for the three primary models (Yi: 85.5%).

Y Weight Delta Analysis

DPO-trained weight delta norms ($\|\Delta W\|_F$): MLP dominates (64–75%), but V/O / Q/K $\approx 1\times$. This apparent contradiction with the 1.76–3.91$\times$ gradient ratio is resolved by Adam's adaptive learning rate: the effective update $\Delta W \propto g/\sqrt{v+\epsilon}$ normalizes by the running second moment $v \approx \mathbb{E}[g^2]$. Parameters receiving consistently larger gradients (V/O) accumulate proportionally larger v, yielding similar-magnitude updates. The gradient asymmetry thus manifests not as larger weight movements but as higher signal-to-noise ratio: V/O gradients are more consistently aligned with the loss-reduction direction (fewer conflicting gradient components across training examples), producing updates that more efficiently modify the relevant computation. This is consistent with the observed pattern: V/O and Q/K have similar $\|\Delta W\|_F$ but V/O changes have disproportionate functional impact (as measured by weight patching), indicating higher directional quality per unit of weight change.

Z Softmax Jacobian Gradient Bound

The gradient with respect to W_V flows as $\partial\mathcal{L}/\partial W_V = X^\top A^\top(\partial\mathcal{L}/\partial(AV))$, where $A = \text{softmax}(QK^\top/\sqrt{d})$ is the attention matrix. By contrast, the gradient with respect to W_K includes the softmax Jacobian: $\partial\mathcal{L}/\partial W_K \propto X^\top(J_{\text{softmax}}[\partial\mathcal{L}/\partial A])^\top Q$, where $J_{\text{softmax}} = \text{diag}(a) - aa^\top$ for row a of A. J_{softmax} is positive semidefinite.

Claim. $\|J_{\text{softmax}}\|_2 \leq \frac{1}{2}$.

Proof. For any unit vector v ($\|v\|_2 = 1$), observe that

$$v^\top J_{\text{softmax}}\, v = \sum_i a_i v_i^2 - \Big(\sum_i a_i v_i\Big)^2 = \text{Var}_a(v),$$

the variance of v under the probability distribution a. By the *Popoviciu inequality*, for a random variable taking values in $[m, M]$, the variance is at most $(M-m)^2/4$. Here $M = \max_i v_i = v_p$ and $m = \min_i v_i = v_q$ for some indices $p \neq q$. Since $(M-m)^2 \leq 2(M^2+m^2)$ (by $(a-b)^2 \leq 2(a^2+b^2)$) and $M^2 + m^2 = v_p^2 + v_q^2 \leq \sum_i v_i^2 = 1$, we obtain

$$v^\top J_{\text{softmax}}\, v \;\leq\; \frac{(M-m)^2}{4} \;\leq\; \frac{2}{4} \;=\; \tfrac{1}{2}.$$

The bound is tight: for $a = [\frac{1}{2}, \frac{1}{2}, 0, \ldots]$ and $v = [\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, \ldots]$, one verifies directly that $v^\top J_{\text{softmax}}\, v = \frac{1}{2}$ and that $\frac{1}{2}$ is an eigenvalue of J_{softmax} (with J_{softmax} having eigenvalues $\{\frac{1}{2}, 0, 0, \ldots\}$). Hence

$$\|J_{\text{softmax}}\|_2 \;=\; \tfrac{1}{2}. \qquad \square$$

Note that the individual diagonal elements satisfy $a_i(1-a_i) \leq \frac{1}{4}$, but this is *not* the spectral norm of the full matrix $J = \text{diag}(a) - aa^\top$. Also note that the intermediate bound $\sum_i a_i v_i^2 \leq \max_i(a_i)$ (valid for $\|v\|_2 = 1$) does *not* yield $\frac{1}{2}$ in general, since $\max_i(a_i)$ can be arbitrarily close to 1 (e.g. $a = [0.9, 0.1, 0, \ldots]$ gives $\max_i(a_i) = 0.9$); the tight $\frac{1}{2}$ bound requires the Popoviciu argument above.

Because $\|J_{\text{softmax}}\|_2 \leq \frac{1}{2} < 1$, the Jacobian attenuates gradient flow to W_Q/W_K by a factor that depends on attention entropy. For peaked attention distributions (common in later layers), the Jacobian eigenvalues approach zero, yielding stronger attenuation. The empirical 1.76–3.91× ratio across models is consistent with typical attention entropy in these architectures, though we do not derive the exact ratio from first principles.

AA Block Permutation Test

The Mann-Whitney U test in Table 1 treats layers as independent observations, but layer-to-layer autocorrelation (lag-1 ρ up to 0.66 at 7B, 0.83 at 14B, for V/O scores) may inflate significance. We address this with a block permutation test: for each of 100,000 permutations, we independently swap the V/O and Q/K labels *within* each layer (preserving all cross-layer correlation structure), then compute the test statistic $T = \overline{|\text{V/O}|} - \overline{|\text{Q/K}|}$. The one-sided p-value is the fraction of permuted T values exceeding the observed T.

Results (Table 1, p_{block} column): Llama ($p = 1.1 \times 10^{-4}$), Mistral ($p < 10^{-5}$), and Qwen-14B ($p = 0.015$) remain significant. Qwen-7B ($p = 0.28$) and Yi ($p = 0.13$) do not, consistent with their small effect sizes (1.20–1.22×) and high GQA ratios (7:1 and 8:1). Notably, the V/O–Q/K *difference* has low autocorrelation ($\rho_{\text{diff}} = -0.07$ for Qwen-7B) because V/O and Q/K layer profiles are largely independent; the non-significance is driven by the small effect relative to cross-layer variance, not by autocorrelation inflating the p-value. For Qwen and Yi, the V/O > Q/K claim is instead supported by gradient analysis ($p < 2 \times 10^{-9}$, Table 3) and, for Qwen, OC-DPO (10.6× differential). BH correction across the five models leaves all five conclusions unchanged.

Page's trend test for the within-MLP ordering. Under the raw fp16 protocol, Page's test on per-layer $W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$ data gives Qwen Z=3.88 (p=5.3×10^{-5}), Llama Z=2.88 (p=2.0×10^{-3}), and Mistral Z=6.25 (p=2.1×10^{-10}); BH correction across the three families leaves all three significant. The chat-protocol replication of this test is reported in Appendix AS.

AB Exact Type-Level Shapley Decomposition

Protocol. Four model × protocol cells, 209 examples each, seed 42, bf16 forward with fp32 cross-entropy, `max_length` 512. The two raw cells use the base-repository tokenizer; the two chat cells use each model's instruction-tuned chat template with assistant prefill, and reproduce the prefill cross-entropies of Appendix AS (Qwen 2.449, Llama 1.597). Both chat cells use the instruction-tuned tokenizer, and each cell's coalition losses are stored in full.

Reviewer discussion raised two related objections to the share convention of Table 2: absolute harm scores do not sum to the alignment tax, and single-component scores ignore interactions. We therefore compute an *exact* Shapley decomposition with the seven projection types as players. With n=7 the game is small enough to enumerate all 2^7=128 coalitions, so no sampling is required. For a coalition S we patch every projection of every type in S to its base weights simultaneously and set $v(S) = \mathcal{L}_{\text{IT}} - \mathcal{L}(S)$, with $v(\emptyset)$=0.

Two properties make this responsive to the objections. The efficiency axiom gives $\sum_i \phi_i = v(N)$ *identically*, so the resulting shares are shares of a single measured joint quantity rather than of a sum of absolute scores; we verify this numerically (residual $\leq 3 \times 10^{-17}$). And each ϕ_i averages marginal contributions over all coalition orderings, so every interaction among types is inside the estimand. Note that $v(N)$ is the *projection-attributable* tax: embeddings, layernorms, and the LM head remain instruction-tuned under every coalition, and we report that residual explicitly.

Table 35: **Exact type-level Shapley values** (nats, 209 probe examples, all 128 coalitions per cell). Ratios are reported only where $v(N)$ is positive and bounded away from zero; on Llama raw the tax itself is negative, so no ratio is meaningful. Both chat cells use the instruction-tuned chat template and reproduce the assistant-prefill cross-entropies of Appendix AS exactly (Qwen 2.449, Llama 1.597).

	Qwen raw	**Qwen chat**	**Llama raw**	**Llama chat**
W_{down}	$+0.057$	$+0.410$	-0.058	$+0.123$
W_{up}	$+0.036$	$+0.347$	-0.052	$+0.075$
W_{gate}	$+0.021$	$+0.224$	-0.024	$+0.028$
W_O	$+0.019$	$+0.179$	-0.010	$+0.045$
W_V	$+0.020$	$+0.107$	$+0.003$	$+0.046$
W_Q	$+0.032$	$+0.081$	$+0.002$	$+0.017$
W_K	$+0.006$	$+0.038$	$+0.001$	$+0.009$
full tax	$+0.197$	$+1.336$	-0.064	$+0.392$
$v(N)$	$+0.190$	$+1.386$	-0.137	$+0.343$
non-proj. residual	3.6%	-3.7%	—	$+12.5\%$
V/O : Q/K	1.03	2.40	—	3.53

What reproduces. The within-MLP ordering $W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$ holds in all four cells by magnitude (signed in 3/4; on Llama raw all three MLP values are negative), under an estimand that uses no absolute values and assumes no additivity. Of the three cells with positive $v(N)$, V/O > Q/K holds in two (Llama chat $3.53\times$, Qwen chat $2.40\times$). This is the paper's central hierarchy claim, and it is unaffected by the convention objection. Summing the seven singleton values overstates the measured joint effect by 197% (Qwen raw), which quantifies the non-additivity of first-order scores: single-component scores are optimistic about what joint rollback achieves, which is why SAR's headline recovery is a measured joint loss.

What does not. The attention asymmetry is condition-dependent. Under the chat protocol V/O : Q/K $= 2.40\times$, and the size-conditional ratio rises with coalition size over most of the range (1.75 at $|S|$=0 to 6.27 at $|S|$=5, then -0.81 at $|S|$=6, where a single remaining type makes the denominator unstable). Under the raw protocol it is $1.03\times$ (z=0.11) — a null, not a reversal — and W_Q's value ties W_V and W_O within overlapping bootstrap intervals. The mechanism is visible in the coalition structure: the raw-protocol tax saturates with the MLP alone ($v(\text{MLP}) = +0.191$ versus $v(N) = +0.190$), leaving no net attention remainder to apportion, and the size-conditional ratio inverts at large coalitions ($1.16 \to -7.06$). This is consistent with §4.2 and with GPT-J, where parallel attention likewise yields V/O ≈ Q/K in aggregate functional impact while per-component gradients retain the asymmetry. On Llama under the raw protocol the tax is itself negative (-0.064: the instruction-tuned checkpoint is the better structured generator), so $v(N) = -0.137$ carries the expected sign — rolling every projection back to base raises loss from 0.998 to 1.135 — and no share or ratio is meaningful. Three of four attention values there are statistically null and the fourth (W_O) is negative, so we do not quote a ratio for that cell.

Statistical scope of the Qwen3 replication. Two of the Qwen3 numbers (§8) are weaker than they appear and we scope them here. Against a matched noise null — every component's true effect set to zero, only its measured standard error retained — the *share magnitudes* are far outside chance (MLP 60.3% against a null of 49.5% [44.6, 54.4], $p < 10^{-5}$; output-pathway 52.0% against 42.0% [37.2, 46.9], $p = 5 \times 10^{-5}$). The *ordinal* within-MLP ranking is much weaker on its own: because W_{down}, W_{up}, and W_{gate} carry the largest standard errors, the absolute-value convention reproduces that ranking under pure noise 26% of the time, and only 11 of 36 layers are strictly ordered. Page's $Z = 2.71$ is computed over 36 layer-level summaries; bootstrapped over the 209 examples its 95% interval is $[0.0, 3.8]$. The V/O : Q/K point estimate is $1.51\times$ with a bias-corrected bootstrap interval $[1.24, 2.15]$; we note that the percentile interval for a ratio of absolute means is anti-conservative here, and the correctly clustered statistic — per-example signed V/O minus Q/K harm — is $+0.0022$ nats, 95% CI $[+0.0017, +0.0028]$, positive in 161/209 examples (Wilcoxon $p < 10^{-18}$). Finally, Qwen3's `q_norm`/`k_norm` tensors are outside our projection inventory, so Q/K is rolled back less completely than V/O; the resulting bias runs *against* the reported asymmetry, since Qwen3's Q/K scores are near sign-random (44/72 negative; 46% of Q/K absolute mass from negative scores versus

7% for V/O) and the absolute-value convention therefore inflates the Q/K denominator with noise. A noise-debiased estimator gives $1.83\times$ and a pure-noise-floor ratio $0.98\times$.

Attention subgame. Restricting the game to $\{W_Q, W_K, W_V, W_O\}$ with the MLP held at its instruction-tuned values asks where the tax sits *within* attention, rather than making attention compete with a saturating MLP. Computed from the same coalition values (no additional forward passes), this gives V/O : Q/K $= 4.50$ (Llama chat), 2.07 (Llama raw), 1.86 (Qwen chat), 1.14 (Qwen raw). We report this as a diagnostic rather than a correction: it does not rescue Qwen raw, where W_Q remains the largest attention value, but it does locate the Llama raw full-game sign inversion in the MLP interaction rather than in attention itself.

The optimum is interior, not wholesale. Because we enumerate every coalition, we can ask which one actually minimises loss. In all four cells that coalition is *strictly interior*—four or five of the seven projection types, never the full rollback (Table 36). Full seven-type rollback is worse than the best interior coalition in every cell, and on Llama under the raw protocol the best interior hybrid is better than *both* endpoints: it beats the instruction-tuned checkpoint by 0.025 nats and the base model by 0.089 nats. The instruction-tuned model is therefore not on a monotone path from base, and selective rollback has a genuine interior optimum. This is a statement about weight-space geometry rather than about our attribution method, and it is the clearest justification we have for the paper's central operation being *selective* rather than wholesale reversion.

Table 36: **Loss-minimising coalition per cell** (from the enumerated coalition losses; lower is better). $|S|$ is the number of projection *types* rolled back. "Full" is the all-seven-type rollback.

Cell	IT	Base	Full	Best	Best coalition ($\|S\|$)
Qwen raw	1.178	0.981	0.988	**0.969**	$W_Q, W_K, W_V, W_{\text{up}}, W_{\text{down}}$ (5)
Qwen chat	2.449	1.113	1.063	**1.042**	$W_V, W_O, W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}$ (5)
Llama raw	0.998	1.063	1.135	**0.974**	W_Q, W_K, W_V, W_O (4)
Llama chat	1.597	1.205	1.254	**1.051**	$W_V, W_O, W_{\text{gate}}, W_{\text{up}}$ (4)

AC Per-Parameter Normalization

Under GQA, W_K and W_V individually have fewer parameters than W_Q and W_O, which raises the question of whether the V/O > Q/K asymmetry reflects size rather than functional specialization. At the *group* level it cannot: W_Q and W_O each hold d^2 parameters and W_K and W_V each hold $d \cdot n_{\text{kv}} d_{\text{head}}$, so $|W_V| + |W_O| = |W_Q| + |W_K|$ *identically*, for any GQA ratio (verified per layer: 14,680,064 for Qwen; 20,971,520 for Llama/Mistral; 18,874,368 for Yi). Our grouped comparison is therefore already parameter-matched by construction, and dividing group totals by their (equal) parameter counts leaves the ratio unchanged: the per-parameter group ratio equals the raw ratio exactly (Qwen 1.20, Llama 1.89, Mistral 3.76, Yi 1.22).

We report a per-matrix normalization below for completeness, but we caution against reading it as a robustness check on the group claim. Averaging per-matrix densities up-weights the smaller matrices (W_V, W_K) by the GQA ratio, which inflates both sides unequally; and the per-matrix density ordering is $W_V > W_K > W_Q > W_O$, i.e. a Q/K member ranks second and a V/O member ranks last. Per-parameter density therefore supports the narrower statement that W_V carries the highest attribution per parameter, not the grouped V/O > Q/K claim, which rests on the exact parameter matching above together with the gradient analysis (§4.2).

Table 37: **Per-parameter V/O vs. Q/K attribution.** Harm scores divided by parameter count per matrix. Ratio and p_{block} computed on per-matrix normalized scores; "Abs. ratio" is the per-layer mean ratio (differs from Table 1's total-sum ratio due to heterogeneous layer contributions). These per-matrix ratios exceed the absolute ratios because averaging densities up-weights the smaller W_V/W_K matrices by the GQA ratio; they are *not* evidence that the grouped asymmetry is understated, since the grouped comparison is already exactly parameter-matched (see text).

Model	GQA	Abs. ratio	Per-param ratio	p_{block}	Significant
Qwen2.5-7B	7:1	1.24×	**1.59**×	8.7×10^{-3}	Yes
Llama-3.1-8B	4:1	1.54×	**1.89**×	3.2×10^{-4}	Yes
Mistral-7B	4:1	3.06×	**3.42**×	1.0×10^{-5}	Yes
Yi-1.5-9B	8:1	1.22×	**1.40**×	6.4×10^{-2}	No

These per-matrix numbers should not be used to claim significance for Qwen: the $p = 0.009$ they yield is a consequence of the density-averaging convention rather than of removing a confound, since no group-level confound exists. Qwen's grouped patching result remains non-significant ($p_{\text{block}} = 0.28$), and an exact interaction-aware Shapley decomposition over the seven projection types gives V/O : Q/K $= 1.03\times$ ($z = 0.11$) under the raw protocol and $2.40\times$ under the chat protocol. What the analysis does establish is that the V/O > Q/K ordering is not an artifact of parameter-count asymmetry — because at the group level there is no parameter-count asymmetry to begin with. Mistral's per-parameter ratio derives from the same fp16 attribution run as Table 1 and inherits its magnitude protocol-sensitivity (Appendix AS); the significance of V/O > Q/K does not (MW $p \leq 1.4 \times 10^{-4}$ in all dtype conditions).

AD Magnitude Baseline for SAR

Table 38: **Attribution-guided vs. magnitude-based SAR at k=5%.** Magnitude selects the top-k% components by $\|W_{\text{IT}} - W_{\text{base}}\|_F$. Attribution-guided selection outperforms magnitude-based selection on all four models. Recovery = absolute loss improvement (IT loss − SAR loss); ratio is attribution recovery / magnitude recovery. Magnitude sorting selects almost exclusively W_{gate} projections across all models, while attribution selects diverse component types. Spearman ρ between magnitude and attribution rankings is −0.06–0.42, and top-k overlap is 0–18%, indicating the two criteria identify largely different components.

Model	Attr. rec.	Mag. rec.	Attr/Mag	Mag. composition	ρ
Qwen	+0.059	+0.018	3.4×	9/9 W_{gate}	0.42
Llama	+0.026	+0.008	3.4×	11/11 W_{gate}	0.03
Mistral	+0.020	+0.010	2.0×	11/11 W_{gate}	0.39
Yi	+0.064	+0.001	85.7×	16/16 W_{gate}	−0.06

W_{gate} projections undergo the largest absolute parameter change during instruction tuning but contribute little to structured-generation degradation. This is consistent with the weight delta analysis above: large $\|\Delta W\|_F$ does not imply large task-specific harm. Attribution captures which components are directionally aligned with the loss function, explaining the 2.0–3.4× advantage over magnitude selection on three models (Yi's 85.7× ratio reflects near-zero magnitude-based recovery). The finding also explains why random selection outperforms magnitude on some models (e.g., Yi: random −10.8%, magnitude +0.7%)—random samples are less biased toward W_{gate}.

AE Execution-Level Evaluation

To verify that cross-entropy improvements translate to execution success, we generate completions from IT and SAR-5% models (greedy, max 256 tokens) and validate using category-specific checkers: JSON parse (`json.loads`), Python compile (`compile()`), SQL parse (`sqlparse`), tool-call regex, and format heuristics.

Table 39: Execution success rate: IT vs. SAR-5% (209 examples, greedy decoding).

Model	IT (%)	SAR-5% (%)	Δ (pp)
Qwen2.5-7B	75.6	75.1	-0.5
Llama-3.1-8B	71.3	71.8	$+0.5$
Mistral-7B	75.6	75.6	$+0.0$

SAR-5% matches IT execution success within ± 0.5 pp across all three models, showing that the cross-entropy improvements translate to preserved (not degraded) execution-level performance. Per-category: JSON (+4 pp Qwen), config (+9 pp Qwen), `tool_call` (-8 pp Qwen) are the largest variations; all other categories show $|\Delta| \leq 5$ pp.

Format-level accuracy. Parsing generated outputs by format type (Qwen2.5-7B, 209 examples) with a corrected checker (accepting XML-wrapped tool calls and fixing regex match logic) shows near-uniform preservation: `function_call` is unchanged (Δ=0.0 pp: 63.9% $\rightarrow$ 63.9%, n=36, McNemar p=1.0), SQL is unchanged (Δ=0.0 pp, n=19), and JSON shows a small non-significant decline (-5.6 pp: 61.1% $\rightarrow$ 55.6%, n=18, McNemar p=1.0). Overall accuracy is stable (-0.5 pp: 81.3% $\rightarrow$ 80.9%, McNemar p=1.0). These results are consistent with BFCL evaluation (+0.9 pp on 318 examples; Appendix V), confirming that SAR-5% preserves format-level structured generation capability.

Full pooled-significance test battery (§7). The stratified exact McNemar test over 6 paired strata (37 SAR-only vs. 27 IT-only passes) gives one-sided p=0.13; a Stouffer combination over all 12 strata gives p=0.29—both consistent with directional, non-significant gains. For the non-inferiority claim: the paired accuracy difference of $+0.89$ pp (95% CI $[-0.5, +2.3]$ pp, Wald) is corroborated by a Tango score interval of $[-0.5, +2.4]$ pp; non-inferiority within a 1 pp margin holds under the exact conditional test (p=0.006), the Tango score test (p=0.005), and a cluster-robust test over the 373 unique items treating the three models as correlated repeats (p=0.006). The superiority test from that same cluster-robust estimator gives p=0.12, so the 0.006 figure should be read as a non-inferiority result, not as evidence of a gain; the conclusion is robust to leaving out any single stratum (max p=0.019, Wald). Direct execution per model: -0.48 pp Qwen, $+0.48$ pp Llama, 0.00 pp Mistral.

AF Token-Level CE Analysis

To verify that SAR's cross-entropy improvement targets format-critical positions rather than reflecting diffuse distributional shift, we decompose per-token CE changes by token type. Each target token is classified via regex matching into structural (brackets, quotes, colons, commas, keywords, operators) or content (identifiers, natural language) categories; token-level deltas (IT minus SAR) are then aggregated to example-level means to respect independence.

Table 40: Token-level CE analysis: SAR-5% improvement by token category (example-level paired comparison).

Model	Struct. $\overline{\Delta}$	Content $\overline{\Delta}$	Paired diff	Wilcoxon p	Cohen's d
Qwen2.5-7B	$+0.180$	$+0.003$	0.183	4.4×10^{-12}	0.35
Llama-3.1-8B	$+0.054$	-0.012	0.068	7.4×10^{-20}	0.61
Mistral-7B	$+0.000$	-0.009	0.013	5.5×10^{-3}	0.14

Structural tokens account for 85% (Qwen) and 105% (Llama) of total CE improvement. Llama shows selective structural recovery: content tokens *worsen* slightly (-0.012), confirming that SAR reverses alignment-induced calibration loss specifically on format-critical token positions. The Wilcoxon signed-rank test operates on 209 paired differences (per-example structural mean minus content mean), respecting within-example correlation. Mistral's near-zero structural delta reflects its near-zero net alignment tax (-0.004 nats; Mistral-Instruct is already near-parity with base on structured generation), so there is little CE to redistribute; the direction remains consistent ($p < 0.006$).

Random rollback control. To test whether structural concentration is specific to attribution-guided selection or a general property of any weight rollback on structured-generation text, we repeat the analysis with 10 random component sets (same size as SAR-5%, uniformly sampled). Random rollback also concentrates CE improvement on structural tokens (Qwen: $96 \pm 8\%$; Llama: $82 \pm 25\%$), indicating that the structural concentration reflects a property of *how structured-generation CE decomposes under any perturbation*, not a distinguishing feature of attribution-guided selection. The token-level analysis therefore validates CE as a metric that targets format-critical positions—any CE improvement on structured outputs preferentially reduces perplexity on brackets, delimiters, and keywords—rather than providing independent evidence for the attribution hypothesis, which rests on the 2.8–5.3$\times$ SAR-over-random *magnitude* (Table 4).

Classification sensitivity. To verify that the structural/content gap is not an artifact of our token classification scheme, we re-aggregate under three alternative definitions: (1) *strict delimiters*—only brackets, quotes, colon, comma, semicolon count as structural; (2) *liberal*—all non-content tokens including numbers are structural; (3) *keywords-to-content*—language keywords moved from structural to content. For Qwen, structural share ranges 79–91% across all four schemes (concentration ratio 4.9–8.1$\times$); for Llama, structural mean is positive under all schemes while content mean is negative or near-zero (share 85–118%). The gap is robust to reasonable classification choices.

AG MT-Bench Evaluation

We evaluate conversational quality using MT-Bench [Zheng et al., 2023] (80 multi-turn questions, 8 categories) with Claude Opus 4.6 as judge (T=0). Model responses use greedy decoding (consistent with all other evaluations; the original MT-Bench protocol uses T=0.7). Each question has two turns; the judge scores each turn on a 1–10 scale.

Table 41: **MT-Bench scores: IT vs. SAR-5% (Qwen2.5-7B).** Per-category mean scores across 10 questions each. Overall difference is not significant ($\Delta = +0.025$, paired t-test p=0.86, Wilcoxon p=0.41, n=80).

Category	IT	SAR-5%	Δ
Coding	5.40	6.10	+0.70
Reasoning	4.55	5.35	+0.80
STEM	5.30	5.60	+0.30
Writing	5.55	5.70	+0.15
Roleplay	5.35	5.30	−0.05
Humanities	5.85	5.70	−0.15
Extraction	6.35	5.60	−0.75
Math	8.65	7.85	−0.80
Overall	**5.88**	**5.90**	+0.025

SAR-5% produces no detectable change in overall conversational quality (effect size d_z=0.020, 95% CI [−0.25, +0.30]). Per-category variation is within ±0.80 points; coding and reasoning show directional improvements consistent with SAR's structured-generation benefit, while math and extraction show directional decreases. No individual category difference is statistically significant at n=10 per category.

AH Per-Category SAR Effect

Table 42: **SAR-5% recovery by task category.** Recovery = IT mean loss − SAR mean loss (positive = SAR helps). SAR improves 13–17 of 18 categories per model. Categories most helped: ReAct (highest recovery in 3/4 models) and tool call/JSON. YAML is consistently among the least helped.

Category	n	**Qwen**	**Llama**	**Mistral**	**Yi**
ReAct	13	+0.858	+0.123	+0.016	+0.240
Tool call	23	+0.293	+0.011	+0.022	+0.080
JSON	34	+0.094	+0.042	+0.056	+0.049
Code	20	+0.009	+0.046	+0.017	+0.137
Bash	12	+0.044	+0.036	+0.036	+0.055
SQL	20	+0.014	+0.017	+0.033	+0.032
Regex	11	+0.003	+0.038	+0.031	+0.059
Cron	7	+0.010	+0.017	−0.115	+0.133
GraphQL	4	+0.056	+0.041	+0.019	+0.037
Git	6	−0.006	+0.032	+0.049	+0.017
CSS	6	+0.006	+0.021	−0.006	+0.020
YAML	2	−0.012	+0.007	−0.024	−0.018
Categories helped		16/18	17/18	13/18	16/18

AI Probe Set Composition

Table 43: **Probe set category distribution** (209 examples). Per-category alignment tax (% change in cross-entropy) for each model. Positive = IT increases loss. 9 of 18 categories show positive tax in $\geq$2 models.

Category	n	**Qwen**	**Llama**	**Mistral**
JSON	34	+17.4	−0.2	+8.5
Tool call	23	+41.1	−34.8	−0.3
Code	20	+5.5	+33.8	+19.9
SQL	20	−1.7	−12.6	+0.4
ReAct	13	+76.7	+2.9	−18.1
Bash / CLI	12	+20.5	+9.3	+2.1
Regex	11	−4.6	+4.1	+4.0
Cron	7	−16.2	−31.4	−33.4
CSS	6	+0.7	−3.6	−11.7
Git	6	−0.0	−17.5	+3.4
API / HTTP	5	−2.9	−16.2	−19.0
GraphQL	4	+17.5	+9.8	+7.6
Docker	3	+1.3	−2.2	+7.1
CI/CD	2	−8.6	+21.8	+51.6
YAML	2	−33.0	−17.0	−18.5
Terraform	2	−44.8	+13.1	−3.1
XML	1	−4.2	+60.1	+12.1
Other	38	+6.6	−1.2	−7.6
Total	**209**			

Prompts average 51 characters (median 45, range 18–235); gold targets average 68 characters (median 55, range 9–354). Representative examples (prompt → gold target): `"Convert to JSON: name=Alice, age=30"` → `{"name": "Alice", "age": 30}`; `"Write a SELECT query for users older than 25"` → `SELECT * FROM users WHERE age > 25;`. Each example is formatted as a raw prompt–completion pair without chat templates; cross-entropy is computed on the gold target tokens only. The set emphasizes short, precise structured outputs across 18 task categories. The largest categories are JSON (16%), other (18%, including protobuf, type annotations,

math expressions, data transforms), and tool calls (11%). The tax sign varies by category and model: code shows consistent positive tax across all three models, while cron shows consistent negative tax, illustrating the task-dependence noted in §4.1. Llama's net-negative overall tax (-0.027) is primarily driven by its tool-call category (-34.8%, 23 examples), consistent with Llama-3.1-Instruct's explicit tool-use SFT data. Similarly, Mistral's ReAct category (-18.1%) and cron (-33.4%) drive its near-zero net tax.

AJ Cross-Dataset SAR Transfer

Table 44: **Cross-dataset SAR transfer.** Components selected using BFCL attribution (100 examples, independent of probe set) still improve structured generation on the 209-example probe set, far exceeding random selection. All values are absolute loss reduction (Δloss $\downarrow$); Qwen percentages show recovery relative to its positive alignment tax ($+0.095$). Probe-based values are from the same 209-example attribution used in Table 4; BFCL-based and random values are from the cross-dataset experiment.

Selection source	k**%**	**Qwen (Δloss)**	**Llama (Δloss)**	**Mistral (Δloss)**
BFCL (100 examples)	5%	+0.045 (47.0%)	+0.016	+0.018
Probe (209 examples)	5%	+0.059 (62.2%)	+0.026	+0.020
Random (20 seeds)	5%	+0.014 $\pm$ 0.004 (14.8%)	+0.005 $\pm$ 0.005	+0.008 $\pm$ 0.008
BFCL (100 examples)	10%	+0.075 (79.0%)	+0.022	+0.032
Probe (209 examples)	10%	+0.083 (87.1%)	+0.036	+0.029

Table 45: **Component selection overlap** between BFCL-based and probe-based SAR at different budgets. Under independent random selection, expected overlap is $\approx k \approx 5\%$; observed overlap (36–44%) is 7–9$\times$ higher.

k**%**	**Qwen**	**Llama**	**Mistral**	**Mean**
5%	4/9 (44%)	4/11 (36%)	4/11 (36%)	39%
10%	12/19 (63%)	10/22 (45%)	11/22 (50%)	53%

BFCL-selected SAR-5% achieves 76% of probe-selected effectiveness on Qwen and improves all three models. For Llama and Mistral (negative alignment tax), both BFCL-selected and random rollback improve probe-set loss, but BFCL-selected SAR achieves 2–3$\times$ the improvement of random selection. Component overlap at k=5% averages 39% across models (7–9$\times$ above chance level of $\sim$5%), indicating that BFCL and probe sets identify overlapping harmful components. At the *component-type* level, agreement is 100%: both datasets rank W_{down} highest, followed by W_O/W_V, with Q/K lowest—the instance-level gap (47% vs. 62%) narrows substantially at k=10% (79% vs. 87%, residual gap 8.1 pp), confirming that the structural insight transfers robustly and only layer-level precision degrades across datasets. The residual gap between BFCL-selected and probe-selected SAR at k=5% reflects task-distribution differences between function calling and broader structured generation, not overfitting.

Cross-dataset merged attribution. To further test robustness, we merge probe-based and BFCL-based attribution via z-score normalization (per-source z-scoring then averaging), selecting SAR components from the merged ranking. At k=5%, merged SAR achieves 62.3% recovery—comparable to probe-only (62.8%) and substantially above BFCL-only (46.8%). Merged-to-probe overlap is 7/9 components (78%), versus probe-to-BFCL overlap of 4/9 (44%). This demonstrates that harmful component identification converges across independent datasets: merging adds robustness without sacrificing recovery, and the high merged$\cap$probe overlap (78%) confirms that the probe-selected set is not idiosyncratic.

AK Within-BFCL Split Validation

To test whether attribution provides value beyond structural priors *within a single dataset*, we split BFCL into attribution and evaluation halves (10 random splits, 60/159 examples each). For each split,

we compute full attribution on the attribution half, select top-$k\%$ components, and evaluate recovery on the held-out evaluation half. Across 10 splits, attribution-selected SAR recovers substantially more than heuristic selection on held-out BFCL examples (mean attribution recovery: 50.8% $\pm$ 0.9% vs. heuristic: 5.7% $\pm$ 0.5% vs. random: 9.2% $\pm$ 5.1%; attribution is 5.5$\times$ over random, 8.9$\times$ over heuristic; Wilcoxon W=0, p=0.002; paired t-test t=72.0, $p<10^{-13}$), confirming that model-specific attribution provides precision beyond architectural priors even within a single dataset distribution. Attribution beats both random and heuristic in all 10 splits.

AL Within-Layer Paired Contrast for the Attention Asymmetry

Protocol. Recomputed from the stored per-component harm scores of the attribution runs already described (209 examples, seed 42); Qwen2.5, Llama and Yi use their primary raw fp16 runs, Mistral its fp32 referee run (Appendix AS), Qwen3 and OLMo-2 their bf16 runs (Appendix C), and the chat conditions the bf16 chat-template runs. The layer is the unit of analysis throughout; no new forward passes are involved.

The V/O vs. Q/K comparison in Table 1 is a ratio of sums of absolute harm scores. That estimand has two defects we did not appreciate at submission. It discards sign, so a Q/K denominator whose components cancel is rewarded rather than penalised; and it is not invariant to numerically irrelevant choices. On Mistral-7B, with model, probe set and protocol held fixed and only the forward dtype varied, the ratio moves from 2.75 (bf16) to 3.80 (fp32) — a shift as large as the gap between our highest- and lowest-ratio models.

We therefore adopt a *within-layer paired signed contrast*: for each layer l,

$$d_l = \tfrac{1}{2}\big(h(W_V, l) + h(W_O, l)\big) - \tfrac{1}{2}\big(h(W_Q, l) + h(W_K, l)\big),$$

treating the layer as the independent unit. This uses no ratio, no absolute values, and no cross-layer scale normalisation, and it matches the blocking of the permutation test we already rely on (Appendix AA).

Table 46: **Within-layer paired contrast.** d_l in nats; one-sided sign test over layers. Positive in 6/6 pairs, significant in 5/6.

Model	Layers	mean d_l	positive	sign-test p
OLMo-2-7B	32	+0.00403	31/32	7.7×10^{-9}
Qwen3-8B	36	+0.00221	30/36	3.5×10^{-5}
Mistral-7B (fp32)	32	+0.00099	29/32	1.3×10^{-6}
Yi-1.5-9B	48	+0.00097	39/48	7.6×10^{-6}
Qwen2.5-7B	28	+0.00055	22/28	0.0019
Llama-3.1-8B	32	+0.00028	20/32	0.108 (ns)

Three properties make this the preferable statistic. First, both conditions reported as null in §4.1 (Qwen2.5 p_{block}=0.28, Yi p_{block}=0.13) are significant here, so those nulls were a property of the ratio rather than of the models. Second, the contrast is far more stable than the ratio under the dtype change above (coefficient of variation 2.3% versus 14.0%). Third, it discriminates rather than declaring everything significant: the same test applied to $W_K - W_Q$ and to $W_{\text{gate}} - W_{\text{up}}$ is null in 6/6 models, while $W_{\text{down}} - W_{\text{gate}}$, which our account predicts to be positive, is significant in 5/6.

The one model where the contrast is not significant, Llama-3.1-8B, is also the model whose raw-protocol tax is negative (-0.028, 95% CI $[-0.050, -0.006]$): the instruction-tuned checkpoint outperforms its base on this probe set, so there is no degradation to localise. Under the chat protocol, where Llama's tax is $+0.392$, the contrast is significant ($+0.0063$, 26/32 layers, p=2.7×10^{-4}). We read this as a scope condition on the functional measurement rather than as evidence against the asymmetry, which the gradient analysis establishes independently in 4/4 families ($p < 2 \times 10^{-9}$, and 1.79–4.63$\times$ after per-element size normalisation).

AL.1 The same contrast applied within the MLP

Applying the identical estimand to the MLP separates two sub-claims that the ordinal statement conflates. With $d_l = h(W_{\text{down}}, l) - \tfrac{1}{2}[h(W_{\text{up}}, l) + h(W_{\text{gate}}, l)]$, the pooled contrast is positive in

6 of 7 conditions and significant in 5 (Wilcoxon signed-rank over layers): Qwen2.5-7B $+0.0027$ (p=2.2×10^{-4}), OLMo-2-7B $+0.0024$ (p=3.5×10^{-6}), Qwen3-8B $+0.0013$ (p=1.9×10^{-4}), Mistral-7B $+0.0003$ (p=9.1×10^{-4}), Llama-3.1-8B $+0.0004$ (p=0.033); Yi-1.5-9B is the exception (-0.0007, p=0.14).

The pairwise elements differ sharply in strength. $W_{\text{down}} > W_{\text{gate}}$ is the most robust; $W_{\text{down}} > W_{\text{up}}$ is next. But $W_{\text{up}} > W_{\text{gate}}$ is *not* established: it is null on Llama-3.1-8B (-0.0001, p=0.99), Yi (p=0.76), Qwen3-8B (p=0.24) and OLMo-2-7B (p=0.38), and on the two models where the mean is positive but the test is null the median layer difference is essentially zero ($+2 \times 10^{-6}$ on Qwen3-8B, $+1 \times 10^{-5}$ on OLMo-2-7B), i.e. the positive mean is carried by a small minority of layers rather than being a typical-layer property. We therefore state the MLP result as W_{down} dominating the input projections, with the W_{up}-versus-W_{gate} ordering reported as suggestive.

AM Decomposing Patching Scores: Direct vs. Cascade Effects

Protocol. Three primary models, 209 examples, seed 42, bf16 forward with fp32 cross-entropy, base-repository tokenizer, `max_length` 512. The direct-effect measurement routes a patched block's output perturbation to the final norm and unembedding with all downstream blocks frozen at their clean activations; this is exact for the frozen path and involves no linearisation.

What this decomposition does and does not resolve. We report this analysis with an important negative result about its resolution, which we established after submission. In our implementation the direct-effect magnitude $|h_{\text{direct}}|$ is close to *independent* of the total $|h_{\text{full}}|$ (Spearman $\rho = 0.04$–0.17; regressing $\log|h_{\text{direct}}|$ on $\log|h_{\text{full}}|$ gives slopes 0.00–0.18 with $r^2 \leq 0.02$), it sits near the per-example sampling noise floor, and it agrees in sign with h_{full} only 34–58% of the time. The direct *share* $|h_{\text{direct}}|/|h_{\text{full}}|$ is therefore approximately a constant divided by $|h_{\text{full}}|$, so any component selected for a large total score mechanically exhibits a small share. Under a permutation test that shuffles direct effects across depth-matched components, the low shares we obtain for Qwen's top-ranked W_Q components are unremarkable ($p = 0.26$ and 0.71). We therefore **withdraw** the inference that those components are anomalously mediated, and we do not treat the aggregate direct shares as a mediation estimate. Freezing whole block outputs does not separate direct from mediated effects at the resolution the question requires; a decomposition that freezes attention *patterns* would, and we leave that to future work. What survives is the implementation check (final-layer direct share $= 1.0$ by construction; observed 0.998–1.000) and the depth trend, which follows mechanically from later layers having fewer downstream blocks. The attention asymmetry does not depend on any of this: Appendix AL establishes it for Qwen2.5 without invoking cascade at all.

The methodological claim originally made in §4.2—that weight patching conflates a component's local effect with downstream propagation, inflating some scores—was tested via direct-effect (frozen-downstream) path patching. For each component (t, l) we measure, alongside the standard harm score h_{full}, a *direct-effect* score h_{direct}: the patched block-l output perturbation is routed straight to the final LayerNorm and unembedding with all downstream blocks frozen at their clean (IT) activations, blocking re-computation of downstream attention patterns and MLPs (exact for the frozen path—no linearization; bf16; all 209 examples). Implementation check: the final layer has no downstream blocks, so its direct share must equal 1.0 by construction; observed 0.998–1.000 across all 7 types $\times$ 3 models. Because that layer's share is mechanical, all aggregate statistics below *exclude* the final layer. The $|h_{\text{direct}}|/|h_{\text{full}}|$ ratio counts sign-flipped direct effects as direct signal, which if anything overstates direct shares—conservative for the cascade claim.

Aggregating $|h|$ over all non-final-layer components, the direct effect accounts for 5.3% (Qwen), 33.0% (Llama), and 13.5% (Mistral) of the full patching signal: 67–95% of what single-component weight patching measures is mediated by downstream-layer re-computation. Direct share increases with depth (Spearman $\rho = +0.29$ to $+0.44$ over $n = 189/217/217$ components, $p < 10^{-4}$; components within a layer share depth, so autocorrelation applies), as later components have fewer downstream layers to traverse.

For the two W_Q components in Qwen's SAR top-9 (L26 and L24, the anomaly motivating §4.2), the direct shares are 3.3% and 1.3%, against a median of 15.3% among all layer-22–26 components—a $\approx$5–12$\times$ gap relative to depth-matched peers, consistent with cascading attention-pattern effects

rather than local computation. L26's direct effect is in fact sign-flipped: locally the modification is mildly *beneficial*, and the measured harm arises entirely downstream.

Cascade dominance is not specific to W_Q, however. Per-type direct-share comparisons (Mann-Whitney) are null or mixed across models: under direct-only attribution, Llama's Q/K aggregate share deflates from 21.1% to 13.1% (V/O:Q/K rises $1.24\times \rightarrow 1.37\times$), while Qwen's and Mistral's Q/K shares *increase*. Cascade does not selectively inflate one component type in a consistent direction; rather, nearly all of any component's patching score is downstream-mediated, in model-specific ways that patching alone cannot disentangle. Gradient norms involve no downstream re-computation by construction, which is why they serve as the complementary measurement (Table 3: V/O $>$ Q/K, all four families, $p < 2 \times 10^{-9}$).

Relation to the pattern-only intervention. For these architectures a "patch W_Q's effect into only the attention-pattern computation" intervention coincides with weight-patching W_Q itself: W_Q's output enters the computation solely through $\text{softmax}(QK^\top)$, so weight patching *is* the pattern-only intervention at the point of injection. The meaningful decomposition is where the perturbation is *read out*—same-layer (the patched layer's own OV readout and MLP response, included in h_{direct}) versus downstream layers (blocked in h_{direct}). Our h_{direct} is therefore generous to the direct effect, making the small observed shares conservative.

AN Head-to-Head Comparison with Concurrent Safety-Parameter Rankings

Concurrent work identifies safety-critical parameters via gradient-based scoring [Qi et al., 2026] or safety neurons [Chen et al., 2025]. To test whether these rankings resolve the same structure as our attribution, we adapt both to projection-matrix granularity on the same models: *qi-safety* (SNIP-style $\sum |W \odot \nabla_W \mathcal{L}|$ under a refusal objective on 50 harmful prompts, following Qi et al.'s setting), *chen-safety* (a gradient-based component-level analogue *inspired by* Chen et al.'s safety-neuron sparsity structure: mean of top-1% output-neuron row sums of $|W \odot \nabla_W \mathcal{L}|$ under the same objective—Chen et al.'s original activation-based neuron scoring does not directly produce projection-level rankings), and *qi-sg* (the SNIP score computed under *our* structured-generation objective—a method-vs-objective ablation). We measure (a) rank agreement with our weight-patching harm ranking and (b) SAR-5% recovery when the rollback set is selected by each ranking (identical code path within each model; bf16). *qi-sg* computes gradients on all 209 probe examples, the same data our own attribution uses, so neither ranking is handicapped by out-of-domain data; the safety-objective variants use the 50-prompt harmful set from Appendix L, matching the concurrent methods' published settings. Two caveats on reading the results. Our harm score is the single-component version of the evaluation criterion itself, so the SAR-recovery comparison favors our ranking by construction; the rank correlations, top-k overlaps, and the *qi-sg* ablation carry the substantive weight. And Spearman ρ against our *signed* ranking has an intrinsic ceiling for any non-negative score ($\rho(\text{ours}, |\text{ours}|) = 0.86/0.57/0.85$ for Qwen/Llama/Mistral), so observed ρ should be read against that ceiling rather than against 1.0.[2]

[2]Absolute losses here are *example-weighted* (macro) means under bf16; Table 6 reports *token-weighted* (micro) means under fp16, which is the dominant reason the IT baselines differ (e.g. Qwen 1.178 vs. 0.829; a local bf16 macro evaluation gives 1.124). All conditions in this table share one code path, so within-table deltas and ratios are exact; a re-evaluation with the base-repo tokenizer reproduces the deltas bit-identically for the unchanged selection sets, ruling out tokenizer provenance as a factor.

Table 47: **Concurrent-work rankings vs. ours** (209-example eval). ρ: Spearman vs. our harm ranking (all components; layer autocorrelation applies; the *-norm* rows divide each component's score by its parameter count, removing the size confound that biases raw sums toward large MLP matrices); Ovl: top-k overlap (k=9 Qwen, 11 Llama/Mistral); Δ: SAR-5% absolute loss reduction under that ranking (nats). Random: mean $\pm$ std over 10 seeds.

	Qwen			Llama			Mistral		
Ranking	ρ	Ovl	Δ	ρ	Ovl	Δ	ρ	Ovl	Δ
Ours (patching)	—	9/9	**+.122**	—	11/11	**+.034**	—	11/11	**+.022**
qi-safety	.37	1/9	+.072	.14	1/11	−.000	.50	1/11	+.012
qi-safety-norm	.01	0/9	+.011	.05	0/11	+.006	.11	1/11	+.005
chen-safety	.28	1/9	+.050	−.05	0/11	−.001	.12	2/11	+.016
qi-sg	.28	0/9	+.037	.20	0/11	−.008	.48	1/11	+.010
qi-sg-norm	−.13	0/9	+.017	.15	0/11	+.005	.11	1/11	+.004
Random (10 seeds)	—	—	+.027±.009	—	—	+.007±.004	—	—	+.008±.005

Safety-oriented rankings correlate weakly-to-moderately with our structured-generation harm ranking ($\rho \leq 0.50$ against a signed-vs-unsigned ceiling of 0.57–0.86; top-k overlap $\leq$2/11 everywhere) and underperform it for rollback on all three models: the strongest concurrent condition recovers 59% (Qwen), 16% (Llama), and 75% (Mistral) of the SAR effect. Size normalization, which removes the parameter-count confound in the raw-sum adaptation, makes the concurrent rankings *worse* (9–23% recovery), so the unnormalized rows are the generous reading of the concurrent methods. On Llama, every concurrent condition falls below even random selection ($+0.007$): their rankings carry no usable signal for this task on that model. The concurrent methods and ours therefore resolve different structure—their scores localize refusal circuitry, ours localizes structured-generation harm—and component-level safety scoring does not recover the within-layer structured-generation hierarchy ((1)).

AO MHA Gradient Analysis (GQA Confound Control)

All four primary models use Grouped Query Attention (GQA, 4:1–8:1), where W_K/W_V have fewer parameters than W_Q/W_O. To rule out parameter-count asymmetry as the driver of V/O > Q/K gradients, we repeat the gradient analysis on OPT-6.7B [Zhang et al., 2022]—a standard Multi-Head Attention model with 32 attention heads and 32 KV heads (GQA ratio 1:1), where W_Q, W_K, W_V, W_O all have identical shape [4096, 4096].

Setup. DPO-style preference loss (10 structured-generation pairs, β=0.1) with per-layer gradient measurement over 10 independent samples (no parameter updates; seed 42). Mean gradient norms per component type are compared via Mann-Whitney U and block permutation tests (100,000 permutations).

Results. V/O mean gradient norm = 3.07, Q/K mean = 0.47; ratio = **6.59**× ($p = 2.0 \times 10^{-16}$, Mann-Whitney; $p < 10^{-5}$, block permutation). Per-type means: W_V = 5.28, W_O = 0.85, W_Q = 0.56, W_K = 0.37. The effect is larger than in GQA models (1.76–3.91×), consistent with GQA actually *compressing* V/K differences by reducing their parameter count.

Interpretation. Under identical parameter counts, V/O gradients still dominate Q/K by 6.6×, confirming that the asymmetry is driven by softmax Jacobian attenuation (the Q/K gradient path passes through the rank-deficient J_{softmax} bottleneck; Appendix Z), not by parameter-count differences under GQA. OPT-6.7B serves a dual control role: (1) non-GQA control (ruling out parameter-count asymmetry) and (2) non-SwiGLU control (standard ungated MLP with fc1/fc2 rather than gate/up/down). The V/O > Q/K gradient asymmetry—and the $W_V > W_O$ sub-hierarchy—persists in this fundamentally different architecture, demonstrating that the attention-pathway finding is independent of both GQA and gating mechanism. The within-MLP hierarchy ($W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$) cannot be tested in ungated architectures by definition. Notably, in OPT's ungated MLP, fc1 (expand, 4096→16384) gradient norms *exceed* fc2 (contract, 16384→4096) by 2.3×—the *opposite* of what residual-proximity would predict (fc2 projects to the residual stream). Combined with W_V (5.28) $\gg W_O$ (0.85) in attention (where W_O is residual-proximate but W_V is not), OPT provides evidence

that gradient magnitude in this architecture is driven by *input dimensionality* and information-flow geometry rather than residual-stream proximity alone.

AP Layer-Level vs. Component-Level SAR

To test whether fine-grained component attribution provides value beyond naive layer selection, we compare two strategies selecting the same number of parameters: (1) *component-level SAR* (standard: select individual projections by harm score) and (2) *layer-level SAR* (select entire layers by aggregate harm). Results on Qwen2.5-7B at k=3–10%: component-level achieves 60–98% recovery versus 13–63% for layer-level (advantage of 35–47 pp across all budget levels). The layer-level approach underperforms because harmful modifications concentrate in specific projections within otherwise-healthy layers; selecting the entire layer dilutes the rollback with low-harm components that perturb useful behavior. This confirms that the component-type and layer-specific information jointly contribute to SAR's effectiveness.

AQ Continual Pretraining Attribution Control

To test whether the within-MLP hierarchy ($W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$) is alignment-specific or an architectural property of any fine-tuning, we apply LoRA-based continual pretraining (CPT) on general text (Wikitext-103, 500 steps, lr=10^{-5}, rank 16, max_length=256) to two base models (Qwen2.5-7B and Llama-3.1-8B), then run the same attribution pipeline comparing the CPT model against the original base.

Training convergence. Both models show clear loss reduction confirming that CPT successfully modified the weights: Qwen loss decreased from 2.750 to 2.038 ($\Delta = -0.71$); Llama from 2.584 to 1.816 ($\Delta = -0.77$).

Results. CPT produces the *same* within-MLP hierarchy as alignment:

Component	CPT Qwen (%)	CPT Llama (%)	IT Qwen (%)	IT Llama (%)
W_{down}	**54.1**	**58.4**	**33.5**	25.7
W_{up}	17.7	6.4	22.9	17.6
W_{gate}	1.9	1.1	11.5	15.8
W_V	8.6	14.0	9.4	10.9
W_O	5.5	16.2	11.5	**21.4**
W_Q	5.6	1.3	8.9	5.2
W_K	6.6	2.6	2.4	3.3

The within-MLP ordering $W_{\text{down}} \gg W_{\text{up}} > W_{\text{gate}}$ is strongly preserved under CPT (54.1%:17.7%:1.9% for Qwen; 58.4%:6.4%:1.1% for Llama). V/O > Q/K holds for Llama (share ratio 7.9×; IT: 3.8×) but is weak for Qwen (share ratio 1.2×; IT: 1.9×), suggesting the attention V/O asymmetry depends on both architecture and training dynamics while the MLP hierarchy is robustly architectural.

Interpretation. Both CPT and alignment produce the same within-MLP ordering ($W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$), confirming that this hierarchy is *architectural*—it reflects how gradient magnitudes flow through the transformer's residual-stream connectivity, not a property specific to alignment data. This is consistent with the gradient-flow account (§4.2): W_{down} receives $\sqrt{d_{\text{intermediate}}/d_{\text{model}}} \approx 1.87\times$ larger gradients than W_{up} due to dimensionality scaling, regardless of the training objective. What distinguishes alignment from CPT is the *layer-level* distribution and *per-example* pattern of modifications: quantitatively, the per-layer harm profiles of CPT and IT are largely uncorrelated (0 of 14 per-type exact permutation tests significant; overall |harm| Spearman $-0.05/+0.08$; see "Quantifying layer-level divergence" below), with the partial exception of a shared late-layer W_{down} band present in both. SAR's practical value comes from exploiting this task-specific layer attribution, not the component-type hierarchy alone.

Quantifying layer-level divergence. To turn the qualitative claim above into a number, we compute the Spearman correlation between per-layer |harm| profiles of CPT (vs. base) and IT (vs. base), per component type, with an *exact* circular-shift permutation test (all $L-1$ shifts enumerated, add-one corrected; the attainable floor is $2/L$ for the two-sided test used here, i.e. 0.063–0.071, whereas Monte-Carlo sampling can return values below it). This replaces an earlier Monte-Carlo implementation of the same test, which we found to be anti-conservative here: with only $L-1$ distinct shifts available, sampling them with replacement can return p-values below the test's own granularity floor. Under exact enumeration one apparently significant result disappears—Llama's W_Q correlation, previously p=0.027, was below that floor—and we report the corrected outcome throughout. This CPT run is an independent retraining under the Appendix-AJ protocol (seed 42; training-loss trajectory matches the original run). Overall per-component correlation of |harm|: Qwen $\rho = -0.05$ (n=196 components), Llama $\rho = +0.08$ (n=224). Per component type, 0 of 14 tests reach significance (Qwen: all $p_{\text{exact}} \geq 0.107$; Llama: all $p_{\text{exact}} \geq 0.0625$). Three qualifications sharpen this claim. First, W_{down} shows the highest (non-significant) positive correlation in both models ($\rho = +0.26$ and $+0.46$), and Llama's top-9 harm components overlap CPT's above chance (3/9, hypergeometric p=0.003; all three shared components are late-layer W_{down}, L28–30). The precise claim is therefore: layer placement is alignment-specific for attention and gate/up components; late-layer W_{down} concentration is *partially shared* with continual pretraining, consistent with the architectural component of the hierarchy. Second, on the concern that CPT-side noise attenuates the correlations (CPT harm magnitudes are 2–4.5$\times$ smaller than IT's): the same pipeline detects the W_{down} correspondence where it exists, so it is sensitive enough to find correlation when present, and the IT-side profiles are demonstrably reliable (independent attribution replicates on the same model correlate at mean $\rho = 0.55$–0.74 under matched protocol). Third, even with the shared W_{down} band, architectural priors alone recover only 5.7% on held-out BFCL while task-specific attribution recovers 50.8% (8.9$\times$; Appendix AK)—the shared band is not sufficient to guide intervention.

AR GPT-J Attribution (Non-SwiGLU Architecture Control)

To test architecture-specific predictions of the gradient-flow theory, we run the full weight-patching attribution pipeline on GPT-J-6B [Wang and Komatsuzaki, 2021] (base) and GPT-JT-6B-v1 (instruction-tuned variant). GPT-J uses an ungated GELU MLP (fc_in: $4096 \to 16384$, fc_out: $16384 \to 4096$) and *parallel* attention+MLP with standard multi-head attention (16 heads, no GQA)—architecturally distinct from our primary sequential SwiGLU models. The theory predicts: (i) V/O $\approx$ Q/K in aggregate functional impact (parallel architecture eliminates cascade amplification), and (ii) fc_out $\approx$ fc_in (ungated GELU has no asymmetric gating). Both predictions are confirmed below.

Setup. We employ two attribution approaches: (1) single-component weight patching on 50 examples (28 layers $\times$ 6 component types), and (2) group-level patching on the full 209-example probe set, simultaneously swapping all V+O components or all Q+K components. The group-level approach bypasses single-component noise from GPT-J's parallel attention+MLP architecture (where individual component effects are confounded by cross-pathway interactions). Additionally, we measure per-component DPO gradient norms (10 gradient samples, no parameter updates, β=0.1) as a power-independent test.

Single-component attribution. The alignment tax is small (+0.032), reflecting GPT-JT's continual-pretraining-based instruction tuning (rather than RLHF/DPO). Single-component weight patching does not reach significance (block permutation $p = 0.11$–0.45 across runs), consistent with power limitations when the tax is near-zero.

Group-level patching (209 examples). Simultaneously swapping all V+O or all Q+K components across 28 layers reveals: V+O harm = 0.035, Q+K harm = 0.040, ratio = 0.88$\times$ (individual group harms can exceed total tax of 0.032 because negative-harm components in the complement partially offset). Statistical tests confirm no significant difference: Mann-Whitney p=0.31, Wilcoxon p=0.63, block permutation p=0.64. GPT-J's parallel architecture (attention and MLP computed independently, then summed into the residual stream) means gradients reaching Q/K do not pass through the V$\to$O$\to$residual path, reducing the softmax Jacobian attenuation effect that drives V/O $>$ Q/K in sequential architectures. This result is *predicted* by our theoretical framework and provides

supporting evidence that the V/O > Q/K asymmetry in Llama/Qwen/Mistral arises from sequential attention→MLP information flow rather than an intrinsic attention-head property.

Gradient results (power-independent). DPO gradient norms provide a convergent test that is independent of alignment tax magnitude:

- **V/O > Q/K**: ratio = 2.38× (V/O mean = 0.321, Q/K mean = 0.135). Mann-Whitney $p = 3.6 \times 10^{-6}$; block permutation $p < 10^{-5}$; rank-biserial $r = 0.49$ (large effect).
- **SFT confirms**: V/O:Q/K ratio = 2.42×, $p = 8.8 \times 10^{-8}$—the asymmetry is architectural, not DPO-specific.
- **fc_out ≈ fc_in**: ratio = 0.92× ($p = 0.99$, not significant). Unlike SwiGLU models where $W_{\text{down}} \gg W_{\text{up}}$, the ungated GELU architecture shows symmetric MLP gradient flow—consistent with the gradient-theory prediction that gating (SiLU vs. SiLU') breaks the symmetry.

Interpretation. The gradient analysis confirms V/O > Q/K per-component gradient magnitudes ($p < 10^{-5}$), while group-level patching shows V+O ≈ Q+K in *aggregate functional impact*. This apparent discrepancy is explained by GPT-J's parallel architecture: per-component gradient magnitudes reflect local curvature (still larger for V/O due to softmax structure), but aggregate functional impact is equalized because the parallel residual paths prevent cascade amplification. The symmetric fc_out ≈ fc_in result validates our gradient-flow theory: without the gating mechanism (SwiGLU's W_{gate}), there is no path for asymmetric gradient amplification between input and output MLP projections. Combined with OPT-6.7B (6.6× V/O gradient dominance; Appendix AO), these results confirm that attention output-pathway concentration in functional impact is architecture-dependent (sequential: V/O > Q/K; parallel: V/O ≈ Q/K), while within-MLP hierarchy is architecture-specific (SwiGLU: $W_{\text{down}} \gg W_{\text{up}}$; ungated GELU: symmetric).

AS Chat Template Evaluation

Our primary evaluation uses raw prompt-completion format (no chat template), matching tool-use deployment patterns. To test sensitivity to this choice, we re-evaluate the alignment tax and SAR recovery under each model's native chat template. Results: **Qwen2.5-7B**: raw tax = 0.095, chat tax = 1.41 (14.8× larger); SAR recovery under chat = 32.9%. **Llama-3.1-8B**: raw tax = −0.027, chat tax = 0.188; SAR recovery under chat = 66.5%. The raw-format evaluation is conservative: under chat templates, the IT model incurs a substantially larger alignment tax (likely because structured outputs conflict with the chat-trained distribution), and SAR still provides meaningful recovery. The reversed sign for Llama under raw evaluation (IT slightly better than base) resolves under chat template, where a clear positive tax emerges.

Chat-template attribution. To test whether the component-level *hierarchy* (not just the tax magnitude) depends on evaluation format, we repeat the full 209-example weight-patching attribution with every example formatted via each model's native chat template (`apply_chat_template`, `add_generation_prompt=True`; bf16, vs. fp16 for the raw runs—each column is internally consistent; all 209 examples valid per model). The orderings reproduce on all three models:

Table 48: **Attribution under chat-template format** vs. raw format (209 examples). The within-MLP ordering holds for all three models, and V/O > Q/K is significant for all three under both Mann-Whitney and the block permutation test (§4.1 primary standard; 100k permutations)—including Qwen, the weakest case under raw format (p_{block}=0.28 raw → 5.4×10^{-4} chat).

Model	MLP (raw→chat)	Within-MLP (chat)	V/O:Q/K (raw→chat)	p_{MW}	p_{block}
Qwen	63.0% → 64.1%	27.3 / 21.7 / 15.0	1.20× → 1.56×	1.9×10^{-4}	5.4×10^{-4}
Llama	59.0% → 64.0%	29.9 / 21.1 / 13.1	1.89× → 1.98×	2.2×10^{-5}	1.4×10^{-3}
Mistral	61.6% → 53.5%	24.8 / 18.6 / 10.1	3.76× → 2.20×	2.6×10^{-5}	9.1×10^{-4}

A possible objection to the naive wrap above is that many probe prompts end in structural fragments (`{"`, `SELECT `, `Action: `) which, as the end of a *user* turn, put the gold continuation in an unnatural position right after the assistant header. We therefore also ran an *assistant-prefill* variant: the

trailing structural fragment moves into the assistant turn and the gold target is scored as its natural continuation. The hierarchy reproduces under this protocol as well—MLP share 63.3%/63.9%/56.5% (Qwen/Llama/Mistral), $W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$ in all three, V/O:Q/K $= 1.44\times\ /2.04\times\ /1.86\times$ (MW $p = 4.0 \times 10^{-3}/1.6 \times 10^{-5}/9.4 \times 10^{-5}$)—so the structure is robust to both chat-wrapping conventions.

The sub-layer structure is therefore not an artifact of the raw-format protocol: $W_{\text{down}} > W_{\text{up}} > W_{\text{gate}}$ (Page's test on chat per-layer data: $Z = 4.54/5.62/3.75$, all $p < 10^{-4}$) and V/O > Q/K hold under the chat format, where the tax itself is up to 14.8× larger.

Protocol sensitivity of Mistral's ratio magnitude. Because the chat runs use bf16 while the raw runs used fp16, we also ran a raw-format bf16 evaluation for Mistral (the model whose magnitudes moved most), which additionally omitted the BOS token. Across conditions, Mistral's V/O:Q/K magnitude—though never its ordering or significance—varies substantially: 3.76× (fp16, raw, BOS), 1.45× (bf16, raw, no BOS), 2.20× (bf16, chat, BOS). Two mechanisms account for this, and neither undermines the ordering. First, *the ratio's denominator sits at the sampling-noise floor*: Mistral's raw-format Q/K harm sums (0.0195 fp16, 0.0253 bf16, over 64 components) are statistically indistinguishable from the expected null sum (0.0204, from per-component standard errors over 209 examples), so the point value of the ratio is ill-conditioned by construction; consistently, 41% of W_K and 28% of W_Q components flip harm sign between conditions, versus 12–19% for W_V/W_O. (Qwen and Llama do not share this instability: their Q/K sums sit 2.7× and 4.7× above their noise floors.) Second, *BOS handling—more than dtype—drives the layer-0 shift*: Mistral's layer-0 attention-sink components collapse when BOS is absent and rebound when the chat template restores it (L0 W_V harm: $+0.0075$ with BOS/fp16, $+0.0004$ without BOS/bf16, $+0.127$ with BOS under the chat template). A harmonized rerun varying *only* the forward dtype (step8 protocol, BOS on, fp32 CE in all conditions, TF32 disabled) settles the decomposition with a full-precision referee: **fp32 gives 3.80**× (MW $p = 8.4 \times 10^{-11}$; example-level bootstrap 95% CI $[2.8, 4.5]$), fp16 gives 3.75× (reproducing the published 3.76×), and bf16 gives 2.75× (MW $p \leq 5.4 \times 10^{-7}$ in all three). The published fp16 value is thus the *accurate* one—fp32 confirms it—while bf16's 8-bit mantissa mildly deflates the ratio and the earlier 1.45× was dominated by the BOS omission (1.45× → 2.75× on restoring BOS). The bootstrap CI quantifies the noise-floor analysis above: the ratio is genuinely large, with sampling uncertainty of roughly ±0.9 driven by the small Q/K denominator. The robust claims are the ordering and its significance (MW $p \leq 1.4 \times 10^{-4}$ in every condition), independently corroborated by gradient analysis (3.66×, a patching-free method).

MLP-share stability under the bf16/no-BOS control. The same control leaves the MLP-dominance finding essentially untouched: Qwen and Llama's MLP and Q/K shares are unchanged to within rounding. Mistral moves the most—MLP share $59\text{–}63\% \to 51.1\%$, Q/K share $8.1\% \to 20.0\%$—but MLP remains the largest group by a wide margin, and the combined output pathway ($W_V, W_O, W_{\text{down}}$) stays inside the paper's 49–65% band at 54.3%; V/O > Q/K also remains significant in this condition.

A noise-debiased decomposition. Because a ratio with a noise-floor denominator is ill-conditioned by construction, we replace it with a decomposition that subtracts each component's expected null magnitude ($\mathbb{E}|\bar{h}|_0 = \text{SE} \cdot \sqrt{2/\pi}$), yielding the *excess* harm each group carries above noise. V/O carries positive excess in every condition (fp16-raw $+0.048$, z=15.9; bf16-raw $+0.011$, z=3.7; chat $+0.525$; prefill $+0.350$), whereas Q/K excess is statistically indistinguishable from zero under the raw protocol (fp16: -0.001 ± 0.005; bf16: $+0.005 \pm 0.006$; χ^2_{64} test of any per-component Q/K signal: $p = 0.28$ and 0.043). Correspondingly, 37 of 64 V/O components are individually significant ($|\bar{h}| > 2$ SE) versus 6 of 64 for Q/K—the latter consistent with the 2σ false-positive rate (expected 2.9; binomial $p = 0.07$). In the conditions where the Q/K denominator *is* identified (chat, prefill), the debiased ratio is stable at 2.0–2.3× and harmonizes across families (Qwen 1.9–2.0×, Llama 2.2–2.3× under the same estimator). The accurate summary for Mistral is therefore *stronger* than any point ratio: under raw-format evaluation, V/O carries essentially all identifiable attention-pathway alignment change, with Q/K consistent with zero; under template protocols the debiased asymmetry is ≥2×. (fp16-raw excesses use SE proxies from the matched bf16 run, as the original fp16 run did not store per-example variances; a ±25% SE-scaling sensitivity check leaves all conclusions unchanged.) The same correction applied to the *share* decomposition confirms that $|\cdot|$-sum noise dilutes rather than inflates MLP dominance: the noise-corrected MLP share rises in all 10 measurement conditions (e.g.,

the bf16/no-BOS control moves from 51.1% to 54.5%), and Mistral's cross-condition spread shrinks from 5.3 to 3.6 pp. The most protocol-stable quantity we measure is the combined output pathway (W_V, W_O, W_{down}): across all ten conditions (fp16-raw, bf16-raw, chat, prefill; three families) its share stays within 49–65%, with per-model cross-condition spreads of 0.5 pp (Qwen), 0.4 pp (Llama), and 2.4 pp across Mistral's three bf16 conditions; the noise-corrected band is 50–58%.

AT Broader Impact

Intended use and benefits. This work identifies which transformer components carry alignment modifications, enabling targeted rollback (SAR) or selective training (OC-DPO). The primary intended use is improving structured-generation quality without degrading safety for downstream agent deployments. The attribution methodology also enables *protective* applications: practitioners can identify safety-critical components and explicitly shield them during further fine-tuning (as OGPSA-hybrid demonstrates, Appendix I).

Dual-use risk. The same attribution map that identifies structured-generation-harming components could be used to selectively remove safety-relevant modifications. SAR is training-free and requires only base and IT weight access—completing in under 5 minutes on a single GPU after attribution—making it more accessible than fine-tuning-based alignment removal [Lermen et al., 2023]. Unlike uniform interpolation (RESTA), which achieves higher absolute safety removal but with large collateral safety degradation (-15 to -22 pp refusal across all behaviors, not just targeted ones), SAR-based rollback is more precise and potentially harder to detect via behavioral evaluation. This attack surface exists only for open-weight models with public base checkpoints; closed-weight or distilled models are not affected.

Yi-1.5-9B and safety–structure overlap. The Yi result (-13.6 pp refusal degradation; Appendix N) reveals a qualitatively distinct regime: when safety-critical and structured-generation-critical components overlap (Yi: L8 W_{down}, L8 W_V score highly on both axes), SAR becomes a near-optimal safety removal tool even when applied for benign purposes. This is both a diagnostic finding and a real limitation: SAR cannot be safely deployed on models with safety–structure overlap, and we did not know Yi belonged to this class until we ran the overlap analysis. One hypothesis for the overlap is Yi's 8:1 GQA ratio, the most aggressive compression among the four models, which could force more functional overlap per component; the three cleanly-separated models have lower ratios (4:1–7:1). The association holds across our four models but no experiment varies GQA ratio while holding other factors fixed, and alignment-data composition is a plausible alternative explanation; we therefore state this as a hypothesis, not a finding. Testing it properly—running the attribution pipeline on matched-architecture models differing only in GQA ratio, and asking whether alignment algorithms should adapt component selection to that ratio—is future work. What SAR contributes in the Yi case is that it exposes the overlap—uniform methods like RESTA cannot detect it—so the deployment decision can be made per model.

Recommended deployment protocol. Based on the cross-model evidence, we recommend: (1) a conservative default of $k \leq 5\%$ (Llama exhibits a safety cliff at $k{\approx}9\%$, Appendix Q; $k{=}5\%$ stays within 2 pp refusal change on all three primary models); and (2) a standardized safety–structure overlap test before any SAR application: run per-component refusal-delta attribution on $\geq$50 harmful prompts (the Appendix N protocol) and check whether top structural components carry positive safety scores. Models failing this test (Yi-like) should not receive SAR at any budget.

Limitations of our safety evaluation. Our non-adversarial safety proxy (refusal rate on 250–1,000 prompts, LLM judge) is complemented by adversarial evaluation (GCG, fixed-template jailbreak; 50 prompts per attack on Qwen and Llama; Appendix M) showing no significant ASR increase ($|\Delta| \leq 6$ pp, $p{>}0.6$). However, adversarial coverage remains limited: one white-box attack (GCG) and one black-box attack (fixed-template jailbreak), two models, and 50 prompts per condition. The claim that SAR "preserves safety" should be understood as "does not produce measurable safety degradation under the evaluated protocols," not as a comprehensive guarantee against all attack vectors.

Responsible disclosure. We notified model providers (Meta, Alibaba, Mistral AI) of these findings and discussed the dual-use implications of component-level safety localization prior to public release. We note that the attack surface described above already exists implicitly via task arithmetic and LoRA removal; our contribution is making it more precise, not creating it. The same precision enables protective applications—practitioners can now identify and shield safety-critical components during fine-tuning rather than relying on uniform regularization.

Compute barrier and mitigations. Attribution requires $\sim$200 component evaluations for 7B models (196 for 28-layer Qwen, 224 for 32-layer Llama/Mistral; $\sim$2–4 hours on an A100 40GB), forming a non-trivial but not prohibitive barrier. SAR execution after attribution completes in under 5 minutes. We release code to enable reproducibility; we recommend that (1) any deployment of SAR include model-specific safety evaluation including safety–structure overlap analysis, (2) model providers consider whether public availability of both aligned and base checkpoints enables SAR-style rollback, and (3) future work extend adversarial evaluation to more attack families (multi-turn, transfer-based) and investigate attribution-guided protection of safety-critical components.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction state the contributions—the within-layer structure, the gradient account, and the attribution-guided selection criterion—and scope them explicitly. The interventions themselves are described as diagnostics rather than methodological contributions (§1), the attention claim is scoped as gradient-robust but protocol-dependent in aggregate functional terms (§8), and five limitations are listed in §8.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Five explicit limitations in §8: (1) evaluation scope (short outputs, greedy decoding, chat-template sensitivity), (2) architecture and scale, (3) safety, (4) causal interpretation, and (5) intervention scope.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The softmax Jacobian spectral norm bound ($\|J_{\text{softmax}}\|_2 \leq 1/2$) is stated as a Claim with a complete proof in Appendix Z, including all assumptions (unit vector, probability simplex) and a tightness example. The paper's main contribution is empirical; this is the sole formal result.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All models are publicly available. §3 specifies models, evaluation set (209 examples, 18 categories), metrics, and decoding settings. Algorithm 1 fully specifies SAR. OC-DPO hyperparameters (20 pairs, mid-third layers) are stated in §6; full training details (LoRA rank 16, AdamW, lr=5×10^{-5}) are in Appendix G. Reproduction details for the two contemporary pairs (repositories, layer and component counts, dtype, seed) are in Appendix C, and Yi's numbers require the single-tokenizer protocol described in Appendix A. Code repository is provided at `https://github.com/mlzoo/transformer-component-analysis`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A code repository with all experiment and evaluation scripts is provided at `https://github.com/mlzoo/transformer-component-analysis`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: §3 specifies all models, evaluation protocol, component taxonomy, and compute resources. OC-DPO training details (optimizer, learning rate, LoRA rank) are in Appendix G. SAR is training-free (Algorithm 1).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All attribution comparisons include Mann-Whitney U p-values and effect sizes (Tables 1, 3); block permutation p-values are reported for Table 1 (Appendix AA), with OPT-6.7B block permutation as supplementary validation for gradient results (Appendix AO). SAR/random ratios include bootstrap 95% CIs (Table 4 caption). OC-DPO reports 5-seed robustness (Appendix H). The revision adds: Page's trend test for the within-MLP ordering under both raw and chat protocols (Appendices AA, AS); paired bootstrap 95% CIs on the alignment tax for all six models (Table 7); a within-layer paired signed contrast with sign tests and BH correction, now the primary functional statistic (Appendix AL); cluster-robust non-inferiority over 373 unique items with exact-conditional and Tango score tests, leave-one-stratum-out checks, and the disclosed superiority test (p=0.12) (Appendix AE); exact circular-shift permutation tests for the CPT-vs-IT comparison, which replace an earlier Monte-Carlo implementation we found to be anti-conservative (Appendix AQ); matched-noise-null comparisons for the contemporary replications (Appendix AB); and an exact Shapley decomposition whose efficiency identity is verified to machine precision (Appendix AB). Bootstrap CIs for SAR selection stability (Appendix X).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Experiments run on NVIDIA A100 GPUs (40GB and 80GB) and, for the analyses added during the discussion period, on NVIDIA A10G GPUs via AWS SageMaker g5 instances. 7B/8B experiments: single A100 40GB, $\sim$2–4 hours per model for attribution, $\sim$1 hour for OC-DPO (20 pairs). 14B: single A100 40GB (4-bit NF4) or 2$\times$A100 (fp16). 72B gradient analysis: 4$\times$A100 80GB, $\sim$3 hours. SAR is training-free ($<$5 min). Safety evaluation (LLM judge): $\sim$1 hour per model via cloud API. The primary experiments total $\sim$80 A100 GPU-hours. The later additions—contemporary-model attributions (Appendix C), exact Shapley enumeration over 128 coalitions per cell (Appendix AB), the three-dtype precision referee with its fp32 arm sharded across 4 GPUs (Appendix AS), greedy joint selection, the cascade decomposition, the concurrent-ranking comparison, and the tax-interval runs—add $\sim$31 A10G GPU-hours over 58 SageMaker jobs. Failed and superseded runs are included in these totals.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics. Dual-use risks and potential misuse of SAR for safety removal are discussed in Appendix AT.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: §8 discusses potential misuse and Appendix AT provides full dual-use risk analysis (SAR could remove safety alignment) and mitigation (safety preservation at $k \leq 5\%$; same framework could protect safety-critical components).

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to

any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We evaluate a limited safety proxy (refusal rate on 1,000 prompts via keyword and LLM judge) and report $|\Delta| \leq 1.4$ pp for SAR-5%. This should not be interpreted as comprehensive safety preservation. The broader impact paragraph discusses potential misuse and notes the framework could protect safety-critical components.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models are cited with their original papers and used under their stated licenses: Llama-3.1 [Grattafiori et al., 2024] (Meta Llama 3.1 Community), Qwen2.5 [Qwen Team, 2024] and Qwen3-8B [Yang et al., 2025] (Apache 2.0), Mistral-7B-v0.3 [Jiang et al., 2023] (Apache 2.0), Yi-1.5-9B [Young et al., 2024] (Apache 2.0), OLMo-2-1124-7B [OLMo Team et al., 2024] (Apache 2.0). Datasets and benchmarks are cited with their sources: UltraFeedback [Cui et al., 2024] (MIT), HumanEval [Chen et al., 2021], BFCL [Patil et al., 2025], IFEval [Zhou et al., 2023], MMLU [Hendrycks et al., 2021].

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release a 209-example structured-generation probe set and all experiment/evaluation scripts via the code repository (`https://github.com/mlzoo/transformer-component-analysis`). The probe set is provided as Python source (no separate data file) under the MIT license.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: An LLM (Claude Opus 4.6) is used as a safety judge for evaluating refusal rates (§7, Appendix D). This is a non-standard methodological component. LLMs studied as experimental subjects (Llama, Qwen, Mistral) are the core research objects.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.