WHERE FUNCTION GOES: FUNCTIONAL CORRESPONDENCE BY PARTIAL COUPLING ACROSS MECHANISTIC LEVELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Comparing a pretrained model with its fine-tuned descendant requires deciding what counts as the same function. Similarity in weights, activations, or features does not guarantee the same behavioral role, so it cannot by itself tell whether a component's function was preserved, relocated, or replaced. We define functional equivalence by intervention: under a behavioral lens, a component's function is the signed change in behavior its removal causes, split into an amount and a role. We cast correspondence as a capacity-respecting partial coupling of these amounts that leaves unmatched function explicit. We apply it to two Base-to-SFT lineages, at weight groups and activation writes, and to two vision transformers that share no native coordinates. Most matched mass stays at its original locus, whereas Tülu under an instruction lens relocates up to 19% of matched weight mass. The coupling also predicts what interventions do: it locates where independently removed weight packets take effect, restoring them better than layer-matched random writes, and it finds targets whose removal reproduces a component's removal better than representation matching. In OLMo, ranking SFT activations by unmatched mass recovers more of the Base–SFT gap than ranking by amount; restoring the top 1% removes every opening `<think>` tag at half the pretraining-text cost of amount ranking. Between DeiT-3 and Swin, signed roles still rank targets better than CKA, although absolute reconstruction remains weak, so we claim ranking transfer, not architectural equivalence. To summarize, we provide a correspondence tool for targeting interventions and diffing fine-tuning changes without shared coordinates.

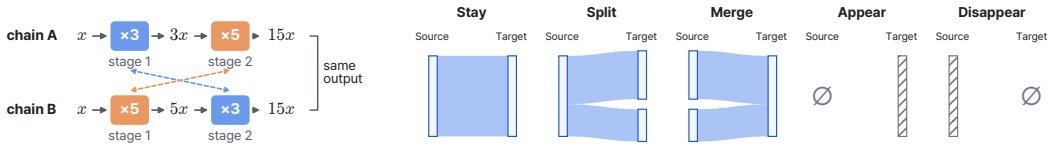
1 INTRODUCTION

Building on a pretrained model, whether by fine-tuning it or reusing it as a component, requires considering which of its learned functions to preserve, which to adapt, and which to combine. These decisions arise both along a model's training lineage and in larger systems assembled from pretrained components. Along a training lineage, better pretrained performance does not necessarily lead to better performance after adaptation (Öncel et al., 2024; Springer et al., 2025). Yet performance alone does not reveal how individual components (*e.g.*, weights or activations) support downstream behavior, or how that support changes under adaptation. Understanding the model adaptation therefore requires capturing how function changes between a pretrained model and its adapted variant.

Prior work relates trained models through weights, activations, or learned features: neuron matching aligns weights before merging (Singh and Jaggi, 2023; Ainsworth et al., 2023), CKA compares representations (Kornblith et al., 2019), and model diffing trains an autoencoder on a shared activation basis to trace fine-tuning changes (Lindsey et al., 2024; Bricken et al., 2024; Minder et al., 2025; 2026). Each judges 'equivalence' by what components compute or encode, not by whether behavior depends on them in the same way. This gap raises a question: *what in two models should count as the same, and what as different in the first place?*

Consider two networks computing $15x$: one multiplies by 3 then 5, the other by 5 then 3 (Figure 1a). Yet removing the $\times 3$ stage from either leaves $5x$, so the two stages play the same role at different depths. We therefore define functional equivalence by intervention: a component's *functional support* is the change in behavior its removals

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



(a) (b)

Figure 1: **Functional support can stay, split, merge, appear, or disappear.** (a) Two chains compute $15x$ through different orders and intermediate states; removing the $\times 3$ stage from either leaves $5x$, so the dashed links pair stages by their effect on the output rather than by depth. (b) Partial coupling allocates functional mass among compatible components and leaves unmatched support explicit.

produce the same pattern of effects. Correspondence also need not be one-to-one: function can *stay*, *split*, *merge*, *appear*, or *disappear* (Figure 1b).

To this end, we introduce a capacity-constrained partial coupling as a correspondence, which represents all five cases in one allocation. We summarize each component's removal effects by two quantities: its *role*, the normalized signed pattern of effects across inputs, and its *mass*, its share of the model's total effect magnitude. Correspondence then becomes a transport problem: optimal transport moves mass from one distribution to another along the cheapest routes, and its partial variant may leave mass unmoved when no route is cheap enough. Our coupling moves mass only between components whose roles are close, never assigning more than either component carries, and is solved as a linear program. Because the transport is fractional, one component's mass can split across several partners or merge with others at one target; because it is partial, mass without a compatible partner stays unmatched rather than being forced onto a poor match.

Under the proposed functional correspondence, we investigate how instruction tuning reorganizes functional support in two Base-to-SFT lineages, OLMo-3-7B to OLMo-3-7B-Think-SFT (Olmo et al., 2026) and Llama-3.1-8B (Grattafiori et al., 2024) to Tülu-3-8B-SFT (Lambert et al., 2025), under two behavioral lenses, Dolma 3 pretraining text and Dolci assistant responses (Olmo et al., 2026), and at two intervention levels, weight groups and activation writes, giving eight settings. In six of eight settings, 97% or more of matched mass stays at its original locus. Tülu under the Instruction lens relocates 19% of matched weight mass (Section 4.1).

We further validate a coupling by examining if it predicts what interventions do, and we test this in three ways. First, *cross-level consistency* (Section 4.2): the weight and activation couplings come from separate interventions, so if the activation coupling locates an independently removed weight change, correspondence reflects the model rather than the measurement. Restoring the activation writes it predicts recovers 0.93–1.00 of the removal effect, more than matched random writes. Second, *target selection* (Section 4.3): the corresponding SFT target reproduces a Base component's removal better than representation matching in 28 of 32 cases. Third, *a functional diff* (Section 4.4): in OLMo, restoring Base activations at the 1% of SFT units with the most unmatched mass recovers more of the fine-tuning change than ranking by amount (0.889 versus 0.827 on new items) and, on the original prompts, removes every opening <think> tag at half the pretraining-text cost.

Because correspondence is defined through behavior, we extend DeiT-3 (Touvron et al., 2022) and Swin (Liu et al., 2021), two vision transformers with different architectures, where correspondence ranks intervention targets better than CKA in 32 of 40 cases (Section 5).

In summary, our main contributions are as follows:

- **Functional correspondence.** We define correspondence through signed intervention effects and use a capacity-constrained partial coupling to represent preservation, relocation, splitting, merging, appearance, and disappearance.
- **Function after fine-tuning.** Across two Base-to-SFT lineages, most matched mass stays at its original locus, while Tülu relocates up to 19% of matched weight mass under the Instruction lens.
- **Intervention-guided targeting.** Correspondence predicts where independent weight removals take effect, selects targets that reproduce source removal effects, and identifies unmatched activations for rollback. Across DeiT-3 and Swin, signed roles also improve target ranking.

2 MEASURING FUNCTIONAL CORRESPONDENCE

2.1 LENSES AND INTERVENTION EFFECTS

We compare two *endpoints*, $m \in \{0, 1\}$, under a behavioral *lens*: an ordered item set $D_{\text{func}} = (x_i)_{i=1}^{n}$ and a scalar readout R. An endpoint is a checkpoint or a subsystem within its surrounding computation, and its inventory $\mathcal{S}_m$ contains native units or groups. No shared coordinates or architecture are required, only a scalar readout both endpoints feed; first-order effects also need it differentiable along the intervention path, which holds for teacher-forced log-probabilities but not for scores of generated text.

Our language-model lenses read length-normalized log-probability. The General lens scores a pretraining document x; the Instruction lens scores the L assistant spans a_ℓ of a conversation x, each given its history H_ℓ:

$$R_m^{\text{gen}}(x) = \frac{1}{|x|} \log p_m(x), \qquad R_m^{\text{ins}}(x) = \frac{1}{\sum_{\ell=1}^{L} |a_\ell|} \sum_{\ell=1}^{L} \log p_m(a_\ell \mid H_\ell). \tag{1}$$

We use two subsets from Dolma 3 pretraining text and Dolci assistant responses (Olmo et al., 2026) for the General lens and Instruction lens, respectively. Both endpoints score the same corpus-provided Dolci-Think-SFT-7B spans, including any <think> content (Appendix C.1).

For component r, the intervention path $\mathcal{I}_{m,r}(\lambda)$ runs from the natural computation ($\lambda = 0$) to full removal ($\lambda = 1$). Writing $R_{m,r}(x; \lambda)$ for the readout along this path, we define the exact removal effect and its first-order approximation:

$$g_{m,r}^{\text{exact}}(x) = R_{m,r}(x; 0) - R_{m,r}(x; 1), \qquad g_{m,r}(x) = -\left.\frac{\partial R_{m,r}(x; \lambda)}{\partial \lambda}\right|_{\lambda=0}. \tag{2}$$

A positive effect means removal lowers the readout. First-order profiles screen large inventories (Syed et al., 2024); exact interventions test their predictions (Vig et al., 2020; Meng et al., 2022; Geiger et al., 2021).

2.2 AMOUNT AND ROLE

The *amount* $A_{m,r}$ is a component's mean absolute first-order effect, and its *mass* $M_{m,r}$ is its share of the endpoint's total amount:

$$A_{m,r} = \mathbb{E}_{x \in D_{\text{func}}} |g_{m,r}(x)|, \qquad M_{m,r} = \frac{A_{m,r}}{\sum_{r' \in \mathcal{S}_m} A_{m,r'}}, \qquad \sum_{r \in \mathcal{S}_m} M_{m,r} = 1. \tag{3}$$

Absolute effects prevent sign cancellation, and normalization gives each inventory a unit budget, so mass is relative, not absolute function.

The *role* records where and in which direction a component acts:

$$\rho_{m,r} = \frac{\left(\max\{g_{m,r}(x_i),\, 0\}\right)_i \,\|\, \left(\max\{-g_{m,r}(x_i),\, 0\}\right)_i}{\sum_i |g_{m,r}(x_i)|}. \tag{4}$$

Positive and negative effects occupy $2n$ separate slots. For example, profiles $(1, -1)$ and $(2, -2)$ share role $(\frac{1}{2}, 0, 0, \frac{1}{2})$ but have amounts 1 and 2. Roles are equivalent when $\rho_{0,r} = \rho_{1,s}$. We compute correspondence over the positive-amount inventory $\mathcal{S}_m^+$.

2.3 WEIGHT GROUPS AND ACTIVATION WRITES

WeightDiff measures parameter function by scaling row or column slices of attention and MLP projections toward zero. The primary inventory groups 16 contiguous slices of one projection into each Coarse16 component. With $C_j = e_j e_j^\top$ and identity $\mathbf{I}$, projections that read from the residual stream are sliced by output row and those that write to it by input column (seven operator classes):

$$W_{m,o}(\lambda) = (\mathbf{I} - \lambda C_j) W_{m,o} \;\text{(Q, K, V, gate, up)}, \qquad W_{m,o}(\lambda) = W_{m,o}(\mathbf{I} - \lambda C_j) \;\text{(output, down)}. \tag{5}$$

ActDiff measures input-dependent function by subtracting the residual-stream write of one MLP channel or attention head at each scored position; at site o with input h, MLP hidden activations $\phi_{m,o}(h)$, and head outputs $z_{m,o}(h)$ on head coordinates S_j, writes sum to the site write $a_{m,o}(h) = \sum_j a^A_{m,(o,j)}(h)$:

$$a^A_{m,(o,j)}(h) = \begin{cases} \phi_{m,o}(h)[j]\, W_{\mathrm{down},m,o}[:, j] & \text{(MLP)}, \\ W_{o,m,o}[:, S_j]\, z_{m,o}(h)[S_j] & \text{(head)}, \end{cases} \qquad a^{(\lambda)}_{m,o}(h) = a_{m,o}(h) - \lambda\, a^A_{m,(o,j)}(h). \quad (6)$$

In shared-coordinate comparisons, *same-locus* indicates the same activation unit or weight groups acting on the same channels within a layer; *relocation* is matched flow outside this relation.

2.4 Partial coupling

We compare roles by total-variation distance:

$$K_{rs} = \mathrm{TV}(\rho_{0,r}, \rho_{1,s}) = \frac{1}{2} \left\| \rho_{0,r} - \rho_{1,s} \right\|_1 \in [0, 1]. \quad (7)$$

Distance 0 means equal roles and 1 disjoint signed supports. At tolerance $\kappa \in (0, 1)$, compatibility means $K_{rs} < \kappa$ and need not be transitive.

Compatibility says which pairs may correspond, not how much: a source of mass 0.8 cannot be fully matched to a target of mass 0.2, and capacity-free nearest-compatible assignment can claim far more target mass than an endpoint holds (Appendix D.3). To address this capacity problem in matching, we utilize a partial optimal transport (Peyré and Cuturi, 2020; Figalli, 2010; Chizat et al., 2019) bounds rather than fixes the mass each component sends or receives, so incompatible mass may stay unmoved. Our coupling rewards each unit of matched mass by its surplus $\kappa - K_{rs}$:

$$\widehat{\Pi}(\kappa) \in \arg\max_{\Pi \geq 0} \sum_{r \in \mathcal{S}_0^+} \sum_{s \in \mathcal{S}_1^+} \Pi_{rs}\big(\kappa - K_{rs}\big) \quad (8)$$

$$\text{subject to} \quad \sum_{s \in \mathcal{S}_1^+} \Pi_{rs} \leq M_{0,r}, \qquad \sum_{r \in \mathcal{S}_0^+} \Pi_{rs} \leq M_{1,s} \quad (9)$$

$$\Pi_{rs} = 0 \quad \text{whenever } K_{rs} \geq \kappa.$$

Unmatched mass earns nothing, so κ prices leaving mass unmatched; fractional allocation permits split and merge, and the upper bounds permit unmatched mass. Matched mass $T(\kappa) = \sum_{rs} \widehat{\Pi}_{rs}(\kappa)$ leaves the residuals

$$u_{0,r}(\kappa) = M_{0,r} - \sum_s \widehat{\Pi}_{rs}(\kappa), \qquad u_{1,s}(\kappa) = M_{1,s} - \sum_r \widehat{\Pi}_{rs}(\kappa),$$
$$\sum_r u_{0,r} = \sum_s u_{1,s} = 1 - T(\kappa). \quad (10)$$

We report at $\kappa^\star$, the grid point with matched mass nearest one half: an operating point, not an estimated threshold (full tolerance path in Appendix D.1; sensitivity checks in Appendix C.3).

3 Recovering Planted Functional Equivalence

Trained models offer no ground-truth correspondence, so we first design to plant the same function in different correspondence with a toy transformer. We separate two questions: whether partial coupling represents a known relation once compatibility is given, and whether signed intervention effects recover that compatibility when no relation is given. The coupling passes both tests: it recovers all five planted relations and rejects candidates with opposite effects.

3.1 Planted models

Both endpoints are untrained two-block pre-LayerNorm Transformers with residual width four. We plant relations by rewriting a first-block MLP-channel or attention-head write a_r into target slots:

$$\text{identity: } a'_s = a_r, \qquad \text{split: } (a'_{s_1}, a'_{s_2}) = (\alpha a_r, (1-\alpha)a_r), \qquad \text{deletion: } a' = 0. \tag{11}$$

Merge reverses split; addition reverses deletion. Removing a component scales its write by $1 - \lambda$. We provide further details in Appendix C.4 about the complete model with transplantation and coefficients.

3.2 Allocation recovery

We give the methods the planted masses $M_0^\star, M_1^\star$ and binary costs $K^\star$ over n_0 source and n_1 target components, but withhold the planted coupling $\Pi^\star$. We compare four matching schemes: pairwise thresholding, one-to-one assignment, balanced optimal transport, and partial optimal transport:

$$\text{pairwise: } \{(r,s) : K^\star_{rs} < \kappa\}, \qquad \text{one-to-one: } \underset{\substack{\text{matching } \mathcal{M} \\ |\mathcal{M}|=\min(n_0,n_1)}}{\arg\min} \textstyle\sum_{(r,s)\in\mathcal{M}} K^\star_{rs}, \tag{12}$$

$$\text{balanced: } \underset{\Pi\geq 0}{\arg\min} \textstyle\sum_{rs} \Pi_{rs} K^\star_{rs} \text{ s.t. } \Pi\mathbf{1} = M_0^\star,\ \Pi^\top\mathbf{1} = M_1^\star, \qquad \text{partial: Equation (8).}$$

At $\kappa = 0.5$, partial coupling recovers $\Pi^\star$ exactly in all five constructions for both intervention inventories (edge F1 1.00; Table 5, Figure 6 for illustration.).

The other schemes succeed only in part. Pairwise thresholding identifies compatible pairs by construction but assigns no mass (Figure 6); one-to-one assignment cannot represent splits or merges; and balanced transport cannot represent deletions or additions (Table 5). Only partial coupling combines fractional allocation with an explicit unmatched residual. Without that residual, appeared or disappeared function must be assigned to a partner, creating spurious relocation. Appendix C.5 repeats the comparison at 7B scale under realistic role noise.

3.3 Role recovery

Allocation recovery assumes compatibility is known. We next ask whether signed intervention effects can recover it. To test the role of sign, each target slot writes c_j times a common direction, so that $g_j(x) = c_j q(x)$. The source has coefficients $(1, 0, 0, 0)$. Four target constructions preserve the source, split it into positive contributions, place an opposed distractor (equal magnitude, opposite direction) beside a match, or contain only opposed candidates (see Figure 7.) As baselines for ranking target slots, we use changes in slot parameters $\theta_{m,s}$ and linear CKA (Kornblith et al., 2019) of centered activations $\bar{Z}_{m,r}$:

$$S_s^{\Delta\theta} = \|\theta_{1,s} - \theta_{0,s}\|_2, \qquad S_{rs}^{\text{rep}} = \frac{\|\bar{Z}_{1,s}^\top \bar{Z}_{0,r}\|_F^2}{\|\bar{Z}_{0,r}^\top \bar{Z}_{0,r}\|_F \, \|\bar{Z}_{1,s}^\top \bar{Z}_{1,s}\|_F}, \tag{13}$$

We evaluate a selected support S with a single joint exact removal on held-out inputs. Given source effect v_0 and target effect $v_{1,S}$, the reconstruction error is defined as follows, with $E = 1$ for an empty intervention:

$$E(S) = \frac{\sum_x \big(v_0(x) - v_{1,S}(x)\big)^2}{\sum_x v_0(x)^2}. \tag{14}$$

Signed roles recover the intended support in all four constructions, with effectively zero reconstruction error whenever a counterpart exists. With an opposed distractor beside a match, the coupling matches half the source mass; when all candidates are opposed, it abstains. In both constructions with opposed candidates, parameter change and CKA select slots that act in the wrong direction, cancelling or reversing the source effect. Thus, components can act on the same inputs with equal magnitude while pushing behavior in opposite directions. Such distractors also occur in trained models (Figure 2).

4 Functional Correspondence After Fine-Tuning

The planted models in Section 3 test whether the method recovers known allocations and roles. We now compare OLMo-3-7B with OLMo-3-7B-Think-SFT (Olmo et al., 2026), and Llama-3.1-8B with Tülu-3-8B-SFT (Grattafiori et al., 2024; Lambert et al., 2025). Component inventories contain about

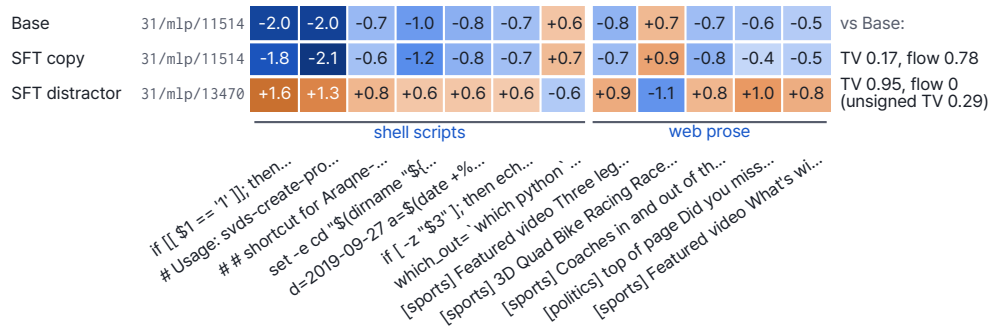


Figure 2: **Why sign matters: overlap in unsigned role does not imply the same function.** Signed roles ($\times 100$; orange positive, blue negative) of Tülu General Base channel 31/11514, its SFT copy, and SFT channel 31/13470 on the 12 items with the largest shared $|\text{role}|$. The distractor is admissible by unsigned role (TV $0.29 < \kappa^\star = 0.4$, its match threshold) but reverses the sign on nearly every item (signed TV 0.95); the coupling sends 0.78 of the Base mass to its own SFT copy and none to the distractor.

10^5–4.6×10^5 units per endpoint. Each behavioral lens uses 768 items to fit the coupling and separate panels of typically 128 items for exact interventions (Appendix C.1). First-order effects screen the inventories: their amount rankings agree closely with exact removal (Spearman ≥ 0.97), while role rankings agree less closely, so we use exact interventions to test the downstream predictions (Table 2). We ask where matched mass goes, whether the coupling predicts independent interventions, and whether unmatched mass guides selective rollback.

4.1 Mass allocation after fine-tuning

Mostly same-locus flow. At each setting's reporting tolerance $\kappa^\star$, at least 97% of matched mass stays at the same locus in six of eight settings (Figure 3). We partition matched mass by the same-locus relation $s \sim r$ from Section 2.3:

$$\text{Same-locus mass: } T_{\text{id}} = \sum_r \sum_{s \sim r} \widehat{\Pi}_{rs}(\kappa^\star), \qquad \text{Relocated mass: } T_{\text{rel}} = T(\kappa^\star) - T_{\text{id}}. \tag{15}$$

Relocation, T_{rel}/T, is the share assigned to a different activation unit or, for weights, to groups acting on different channels. Same-locus flow marks a shared coordinate, not an unchanged role; the matched roles can differ substantially even when their native units coincide (Appendix D.1).

Tülu relocation. Under the Instruction lens, 11% of Tülu's matched activation mass and 19% of its matched weight mass move to different channels. Weight relocation is concentrated in attention, particularly query/key groups. The weight share remains about five times OLMo's when the coupling is refit on either half of the items and changes little when Llama's grouped query/key heads are retied by head (Appendix D.1).

Noise calibration. In a planted OLMo activation inventory with realistic role noise, 3.5–12% of matched flow reaches an incorrect target. Tülu's 19% weight relocation exceeds this range, whereas its 11% activation relocation falls within it. This is a provisional reference: the planted inventory has fewer near-ties than the real couplings, and all relocation shares are relative allocations (Appendix D.1). Native location is therefore usually a useful guide to matched function, with the clearest exception in Tülu's Instruction-lens weights.

4.2 Cross-level consistency

Cross-level test. The activation coupling should predict where an independently removed weight packet acts, although WeightDiff removes parameter groups and ActDiff removes input-dependent residual writes. In each of four settings, we select eight disjoint Coarse16 packets from the highest-flow 1% of transported weight mass. We propagate each packet's Base activation footprint through the coupling, remove its SFT weights, and restore the SFT model's own writes at the smallest predicted

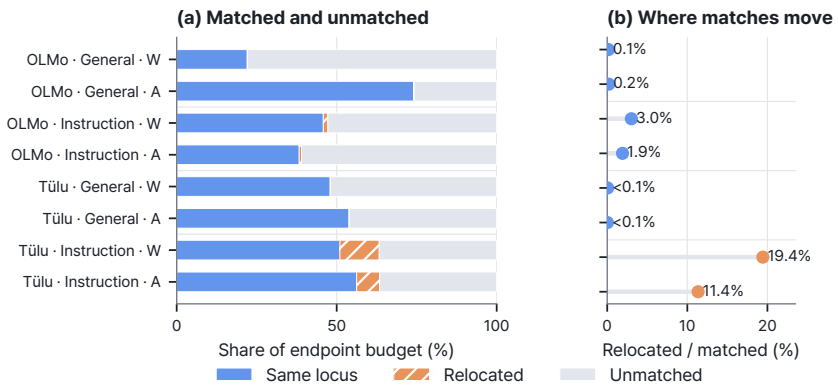


Figure 3: **Most matched mass stays in place; Tülu Instruction redistributes more.** (a) Each endpoint's budget at $\kappa^\star$: same-locus, relocated (hatched), and unmatched mass (W: weight groups; A: activation units). (b) Relocated mass as a fraction of matched mass.

set containing 95% of the energy. Let $\bar{q}_i^{\text{pred}}$ and $\bar{q}_j^{\text{obs}}$ be predicted and observed energy footprints, and v_W and v_{res} the removal effects before and after restoration:

$$\begin{aligned} \text{Footprint distance:}\quad & D_{ij} = \mathrm{TV}\big(\bar{q}_i^{\text{pred}}, \bar{q}_j^{\text{obs}}\big), \\ \text{Restoration recovery:}\quad & \gamma = 1 - \|v_{\text{res}}\|_2^2 / \|v_W\|_2^2, \\ \text{Footprint specificity:}\quad & \Delta D = \tfrac{1}{|\mathcal{P}|(|\mathcal{P}|-1)} \textstyle\sum_{i\neq j} D_{ij} - \tfrac{1}{|\mathcal{P}|} \sum_i D_{ii}. \end{aligned} \tag{16}$$

A packet is identified when $\Delta D > 0$; packet i ranks first when its own observed footprint minimizes D_{ij}.

Prediction and restoration. Predicted footprints identify their own packets in all four settings (23/32 rank first; exact permutation $p \leq 5.2 \times 10^{-4}$). Restoring predicted writes recovers 0.93–1.00 of the removal effect on average, within 0.08 of an observed-change oracle. Cardinality- and operator-matched random writes recover less; disjoint layer-matched controls recover 0.36 versus 0.97 for predicted units where those controls are available. A prospective replication with rules fixed in advance reproduces packet identification (28/32 rank first) and restoration (32/32 wins over random; mean 0.98 versus 0.72). The preregistered layer-matched comparison favors the prediction in 23/32 packets in each set (Appendix D.2). Agreement across distinct interventions supports the coupling as a predictor of where weight changes act in activation space.

4.3 Causal Target Selection

A functional counterpart should reproduce a source component's removal effect. We evaluate a candidate target support S on held-out items using effect-shape error, the reconstruction error after the best non-negative rescaling of its effect:

$$\text{Effect-shape error:}\quad E_{\text{shape}}(S) = \min_{a\geq 0} \frac{\|v_0 - a\, v_{1,S}\|_2^2}{\|v_0\|_2^2} = 1 - \max\big(\cos(v_0, v_{1,S}), 0\big)^2. \tag{17}$$

Correspondence yields lower E_{shape} than representation matching for 28 of 32 multi-unit supports and than random targets for 31. Singleton weight placement also beats parameter-change and random controls in all 32 evaluated cases. These are dependent cases sharing models, lenses, and panels, rather than independent trials (Appendix D.3). The result tests whether the selected targets act like their sources under intervention.

4.4 Unmatched Mass as a Restoration Ranking

Functional diff. Target-unmatched mass ranks SFT units whose support lacks a matched Base counterpart. Each target's SFT mass comprises same-locus flow, relocated incoming flow, and an

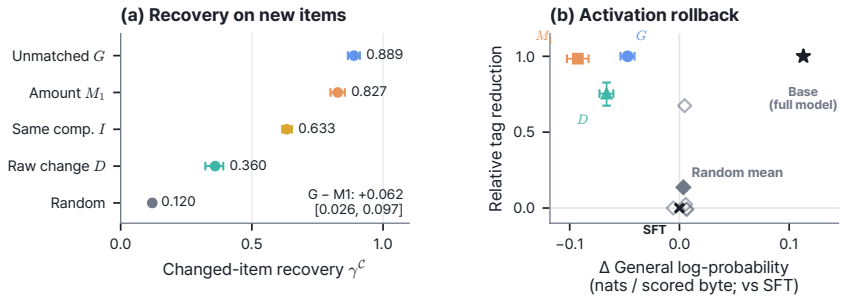


Figure 4: **Target-unmatched mass guides OLMo activation rollback.** (a) Recovery on 128 new OLMo items; the same-component arm I ranks units by same-locus matched flow $\widehat{\Pi}_{\mathrm{id}(s),s}$. (b) Opening-tag reduction and General-lens change on the original panels. Both use the fixed 1% unit budget; intervals are 95% item-bootstrap intervals. Hollow diamonds are five random draws; the random-mean interval ignores draw-to-draw variance. Base is the full-model reference.

unmatched residual G_s. We rank units by G_s, restore Base activation writes at the top 1%, and measure changed-item recovery $\gamma^{\mathcal{C}}$, the fraction of the SFT−Base readout gap reversed on items $\mathcal{C}$ with $|\Delta_i|$ at least the panel median:

$$\text{SFT mass: } M_{1,s} = \underbrace{\widehat{\Pi}_{\mathrm{id}(s),s}}_{\text{same-locus } I_s} + \underbrace{\textstyle\sum_{r\neq \mathrm{id}(s)} \widehat{\Pi}_{rs}}_{\text{relocated } R_s^{\mathrm{in}}} + \underbrace{u_{1,s}}_{\text{target-unmatched } G_s}$$
$$\text{Changed-item recovery: } \gamma^{\mathcal{C}} = 1 - \frac{\sum_{i\in\mathcal{C}} |e_i|}{\sum_{i\in\mathcal{C}} |\Delta_i|}. \tag{18}$$

Here e_i is the remaining gap after restoration. We compare this ranking with SFT mass M_1. Unlike a no-partner score, G_s also captures mass left unmatched by competition for finite capacity; capacity-aware greedy allocation gives similar rollback rankings. Tolerance and allocation checks appear in Appendix D.4.

Prospective replication. With the coupling, rankings, $\kappa^\star = 0.6$, and 1% budget fixed before evaluation, G recovers more of the SFT−Base gap than M_1 on 128 new OLMo Instruction items: 0.889 versus 0.827, a gain of $+0.062$ $[0.026, 0.097]$ (Figure 4). These items are held out from the fit and prior panels but may have appeared in Dolci SFT data.

Behavioral example. On the original OLMo Instruction prompts, restoring the top 1% by G suppresses every opening <think> tag, while amount ranking leaves two. The General-lens log-probability change is -0.047 versus -0.093 nats per scored byte for amount ranking (Figure 4). This result concerns one tag, not broad capability reversal. The ranking is also lineage-dependent: amount ranking is stronger on the Tülu readout, although G removes instruction following more selectively while preserving arithmetic better (Appendix D.4.1). The residual thus provides a useful intervention ranking for some fine-tuned behaviors.

Summary. Across both lineages, most matched support remains at its original locus, with the clearest relocation in Tülu's Instruction-lens weights. The coupling predicts the effects of independent weight interventions and selects targets that reproduce source removal effects, while its unmatched residual guides selective OLMo rollback. These findings support correspondence as an intervention-guided account of fine-tuning changes, conditional on the behavioral lens and relative mass allocation.

5 Correspondence Without Shared Coordinates

We test whether functional correspondence can select intervention targets across models without shared coordinates. DeiT-3 (Touvron et al., 2022) and Swin (Liu et al., 2021) are independently trained vision transformers with different architectures. Using correct-class log-probability on the same 50 ImageNet classes (Deng et al., 2009; Russakovsky et al., 2015), we compare interventions

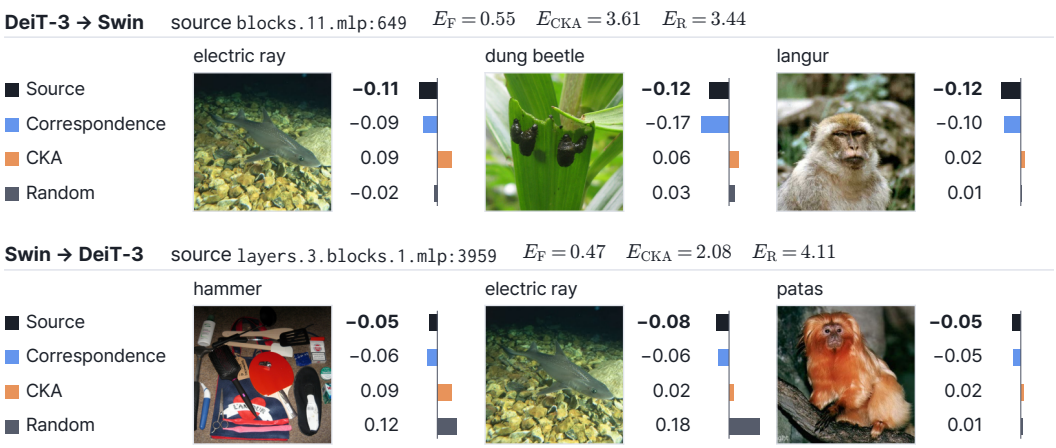


Figure 5: **Cross-architecture cases.** Rows: DeiT-3→Swin; Swin→DeiT-3. Each uses the source with $E_F < 1$ and lowest E_F/E_{CKA}, plus three images with the largest correspondence advantage over CKA. Bars show signed removal effects for source (black), correspondence (blue), CKA (orange), and random (gray); headers give reconstruction errors E. Outcome-blind examples appear in Figure 21.

on their MLP channels. For each source r, we compare the target support selected by correspondence (F) with supports selected by representation similarity (CKA) and random choice. We remove the source fully and each target support S_a at dose d_a; $C_a = \sum_{s \in S_a} M_{1,s}$ is its mass and $B_r = \sum_s \widehat{\Pi}_{rs}$ is the source's outflow. A common budget B_{eval} controls the amount removed by each arm:

$$\text{Arm dose: } d_a = B_{\mathrm{eval}}/C_a, \qquad \text{Common budget: } B_{\mathrm{eval}} = \min\big(B_r, C_F, C_{\mathrm{CKA}}\big). \tag{19}$$

Target ranking. On held-out images, correspondence-selected target interventions match the source's removal-effect shape better than CKA in 32/40 cases and random selection in 31/39 defined cases (Figure 19; protocol in Appendix E). Figure 5 illustrates signed per-image effects in outcome-selected cases. The aggregate result supports target ranking across these architectures, although absolute effect reconstruction remains weak.

Roles for single sources. To isolate the role of coupling, a control selects the same number of nearest-signed-role targets for each source without allocating capacity. It performs similarly to correspondence (14/40 wins, 19 losses, 7 ties for correspondence) and beats CKA in 31/40 cases (Table 20). Signed roles therefore account for the single-source ranking gain.

Capacity under competition. We also intervene on 32 packets of 32 sources chosen to compete for overlapping targets. Here, nearest-role targets raise effect-shape error by $+0.028$ $[0.009, 0.051]$ relative to the coupling, and 42% of their claimed mass exceeds target capacity. A capacity-aware greedy allocation matches the coupling, while the gap shrinks in packets formed without regard to competition (Appendix E). Capacity matters for joint allocation when sources compete for targets; LP optimality adds no measured advantage in this test.

6 Conclusion

We introduced functional correspondence as a capacity-constrained partial coupling of component masses based on signed intervention effects. It represents preservation, relocation, splitting, merging, appearance, and disappearance within a single relative account. Across two Base-to-SFT lineages, most matched mass remains at its original locus, although Tülu shows substantial relocation under the Instruction lens. Correspondence also predicts where independent weight removals take effect, selects intervention targets more accurately than representation matching, and uses unmatched mass to guide selective rollback. Between DeiT-3 and Swin, signed roles improve target ranking even without shared coordinates. Together, these results support correspondence as a tool for locating and testing behavioral changes across models. We discuss measurement dependence, tolerance selection, scalability, and the limits of effect reconstruction in Appendix B.

AI USE STATEMENT

In this work, we used generative AI tools for proofreading, retrieving related papers, and implement codes. We have not used generative AI tools for any task requiring disclosure: we did not use them to develop theoretical models or conceptual frameworks, formulate mathematical claims or proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, interpret results, or generate, clean, or reformat any dataset. Generating synthetic datasets, assisting with translation, and supporting qualitative and thematic analysis are not applicable to this work. We double-check whether the retrieved papers are real and related to ours work, and test AI-generated codes with a test snippet code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our code will be publicly released upon acceptance.

REFERENCES

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

David Applegate, Mateo Diaz Diaz, Oliver Hinder, Haihao Lu, Miles Lubin, Brendan O'Donoghue, and Warren Schudy. Practical large-scale linear programming using primal-dual hybrid gradient. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=_eXwwWOyqT_.

Aryaman Arora, Zhengxuan Wu, Jacob Steinhardt, and Sarah Schwettmann. Adag: Automatically describing attribution graphs, 2026. URL https://arxiv.org/abs/2604.07615.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=ak06J5jNR4.

Trenton Bricken, Siddharth Mishra-Sharma, Jonathan Marcus, Adam Jermyn, Christopher Olah, Kelley Rivoire, and Thomas Henighan. Stage-wise model diffing. Transformer Circuits Thread, December 2024. URL https://transformer-circuits.pub/2024/model-diffing/index.html.

Luis A. Caffarelli and Robert J. McCann. Free boundaries in optimal transport and Monge-Ampère obstacle problems. *Annals of Mathematics*, 171(2):673–730, 2010. doi: 10.4007/annals.2010.171.673.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldwosky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing, a method for rigorously testing interpretability hypotheses. *AI Alignment Forum*, 2022. https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

Jonathn Chang, Arya Datla, and Ziv Goldfeld. Plot: Progressive localization via optimal transport in neural causal abstraction, 2026. URL https://arxiv.org/abs/2605.06979.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulation, 2019. URL https://arxiv.org/abs/1508.05216.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dNigytemkL.

Alessio Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010. doi: 10.1007/s00205-008-0212-7.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL http://jmlr.org/papers/v26/23-0058.html.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching, 2023. URL https://arxiv.org/abs/2304.05969.

ML Nissen Gonzalez, Melwina Albuquerque, Laurence Wroe, Jacob Meyer Cohen, Logan Riggs Smith, and Thomas Dooms. When are two networks the same? tensor similarity for mechanistic interpretability, 2026. URL https://arxiv.org/abs/2605.15183.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie

Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie

Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Tim Rocktäschel, Edward Grefenstette, and David Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=A0HKeKl4Nl.

Aly Kassem, Thomas Jiralerspong, Negar Rostamzadeh, and Golnoosh Farnadi. Delta-crosscoder: Robust crosscoder model diffing in narrow fine-tuning regimes, 2026. URL https://arxiv.org/abs/2603.04426.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=i1uGbfHHpH.

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. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 26361–26378. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/lee24a.html.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence, 2015. URL https://arxiv.org/abs/1411.5908.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations?, 2016. URL https://arxiv.org/abs/1511.07543.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, October 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Los Alamitos, CA, USA, October 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00986. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00986.

Haihao Lu and Jinwen Yang. cupdlp.jl: A gpu implementation of restarted primal-dual hybrid gradient for linear programming in julia, 2024. URL https://arxiv.org/abs/2311.12180.

Haihao Lu, Zedong Peng, and Jinwen Yang. cupdlpx: A further enhanced gpu-based first-order solver for linear programming, 2025. URL https://arxiv.org/abs/2507.14051.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Julian Minder, Clément Dumas, Caden Juang, Bilal Chughtai, and Neel Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=yFdNygEryH.

Julian Minder, Clément Dumas, Stewart Slocum, Helena Casademunt, Cameron Holmes, Robert West, and Neel Nanda. Narrow finetuning leaves clearly readable traces in activation differences. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qyVzZsrsnS.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

Fırat Öncel, Matthias Bethge, Beyza Ermis, Mirco Ravanelli, Cem Subakan, and Çağatay Yıldız. Adaptation odyssey in LLMs: Why does additional pretraining sometimes fail to improve? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19834–19843, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1108. URL https://aclanthology.org/2024.emnlp-main.1108/.

Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020. URL https://arxiv.org/abs/1803.00567.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8sKcAWOf2D.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport, 2023. URL https://arxiv.org/abs/1910.05653.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=YW6edSufht.

Alan Sun and Mariya Toneva. Tracking equivalent mechanistic interpretations across neural networks, 2026. URL https://arxiv.org/abs/2603.30002.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen, editors, *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit, 2022. URL https://arxiv.org/abs/2204.07118.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Kevin Ro 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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

A Related Work

Alignment and representation comparison. Permutation alignment (Li et al., 2016; Entezari et al., 2022; Ainsworth et al., 2023), fusion transport (Singh and Jaggi, 2023), and symmetry removal (Gonzalez et al., 2026) align implementations under reordering; representation methods compare layer geometry (Raghu et al., 2017; Kornblith et al., 2019; Williams et al., 2021) or learn cross-representation maps (Lenc and Vedaldi, 2015; Bansal et al., 2021). We use CKA only as a control, since similar representations can differ functionally, as the signed distractors of Section 3.3 show, and instead assign counterparts from removal effects.

Model diffing and fine-tuning. Shared-dictionary methods expose features common to two models or specific to one (Lindsey et al., 2024; Minder et al., 2025; Kassem et al., 2026; Bricken et al., 2024; Minder et al., 2026); case studies find fine-tuning strengthens a circuit (Prakash et al., 2024), adds a wrapper (Jain et al., 2024), or steers from toxic directions (Lee et al., 2024). These describe representational or circuit change; we instead ask, across full inventories, how much matched function moves, stays, or goes unmatched.

Intervention and causal abstraction. Mediation analysis, causal tracing, and patching localize behavior via interventions (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023; Syed et al., 2024); we reuse the resulting signed profiles. Causal abstraction tests a high-level implementation via interchange (Geiger et al., 2021; 2025), subspaces (Geiger et al., 2024), or ablation (Chan et al., 2022); Sun and Toneva (2026) ask whether models share an interpretation unspecified in advance. PLOT (Chang et al., 2026) matches causal variables to sites but needs a predefined high-level model; our endpoints are native components with none. Path patching, attribution graphs, and ADAG trace particular routes instead (Goldowsky-Dill et al., 2023; Ameisen et al., 2025; Arora et al., 2026), which our roles could further condition on.

Partial transport. Partial and unbalanced transport relax the requirement that all mass be matched (Figalli, 2010; Caffarelli and McCann, 2010; Chizat et al., 2019; Peyré and Cuturi, 2020); ours follows directly, since components can lack a counterpart and capacity limits which sources claim a target.

Across these lines, the distinction is what gets matched: what components compute or encode, what a shared dictionary reconstructs, or neural sites against a predefined high-level model. We align behavioral dependence instead: components correspond when intervening on them changes behavior the same way, whatever they represent.

B Limitations

- **Measurement and model scope.** Correspondence depends on the behavioral lens, component inventory, and intervention path. We study two Base-to-SFT lineages and one vision-model pair; the language-model lenses use teacher-forced log-probability, so their allocations need not describe generated behavior or other readouts. First-order effects rank amounts well but recover roles less reliably, especially for Tülu weight groups; exact interventions validate selected predictions rather than every component (Appendix C.3).
- **Relative mass and tolerance.** Endpoint-normalized masses do not preserve absolute effect scale, and same-locus flow does not imply an unchanged role. The reporting tolerance $\kappa^\star$ is a convention that matches about half the mass, not an estimated optimum. Allocations and residual rankings can change with κ; at high tolerance, nearly all mass matches and the OLMo unmatched-mass rollback signal vanishes (Appendices D.1 and D.4).
- **Cross-architecture transfer.** On DeiT-3 and Swin, correspondence improves target ranking, but absolute effect reconstruction remains weak. Nearest signed roles perform similarly for single sources; capacity helps in joint tests designed to create target competition. The evaluation covers 50 ImageNet classes and selected MLP-channel sources, so it does not establish architectural or functional equivalence beyond those cases (Appendix E).
- **Computation and scale.** Forming intervention profiles and solving couplings becomes more costly as inventories or item sets grow. Nine tolerance-grid solutions outside the reporting points lack

tight verified optimality gaps, and capacity-aware greedy allocation performs comparably to the LP on the reported rollback and joint-target tests. Scaling to larger inventories and choosing an allocation rule for other settings remain open (Appendices D.1 and D.4).

C METHOD IMPLEMENTATION AND CONTROLLED VALIDATION

A *Coarse16* component is 16 contiguous slices of one projection, removed jointly. *Same-locus* means the same activation unit or, for weights, tied groups on the same channels of one layer: {gate, up, down}, {Q, K}, or {V, O} (gate, up, Q, K, and V slice output rows; down and O slice input columns). $G_s = M_{1,s} - \sum_r \widehat{\Pi}_{rs}$ is target-unmatched mass.

C.1 MODELS, LENSES, AND ITEMS

Table 1 lists the model pairs, item panels, and inventories. Forward passes use float32 with float64 accumulation, except the 256 mass-item columns of the OLMo Instruction SFT activation bank, which match a bfloat16 run and barely move the coupling (Appendix D.1.1); first-order effects are Jacobian-vector products along the removal path.

Reference-text provenance. General items come from `allenai/dolma3_mix-6T` (revision `689a3ea2`) and Instruction items from `allenai/Dolci-Think-SFT-7B` (revision `72ec0fe3`, six fixed shards). Both endpoints are teacher-forced on the same conversation and scored on assistant-content bytes only, keeping `<think>` tags as in the data. Dolci is also OLMo's SFT corpus, so disjoint panels do not separate evaluation from SFT training data.

Model pair (Base → SFT)	Lens	Readout	$\lvert D_{\text{func}}\rvert$ / $\lvert D_{\text{test}}\rvert$	Units per endpoint
OLMo-3-7B → OLMo-3-7B-Think-SFT	General, Instruction	Eq. (1)	768 / 128	W: ≈99k Coarse16; A: 353k
Llama-3.1-8B → Tülu-3-8B-SFT	General, Instruction	Eq. (1)	768 / 128	W: ≈106k Coarse16; A: 460k

Table 1: Model pairs, lenses, and inventories. Fitting uses 768 pooled amount and role items; exact evaluation uses separate items. W: weight groups; A: activation units.

C.2 INTERVENTIONS

Section 2.3 gives the WeightDiff and ActDiff removal operators (Equations (5)–(6)). In OLMo, one natural-pass RMSNorm denominator (gain $\gamma_{\text{RMS}} \in \mathbb{R}^{d_{\text{RMS}}}$, stabilizer $\epsilon_{\text{RMS}} > 0$) is shared over the raw output $z = \sum_r z_r$, so

$$a_r = \frac{\gamma_{\text{RMS}} \odot z_r}{\sqrt{d_{\text{RMS}}^{-1}\|z\|_2^2 + \epsilon_{\text{RMS}}}}, \qquad \sum_r a_r = \text{RMSNorm}(z). \tag{20}$$

and removal subtracts a_r after RMSNorm with the denominator fixed, not $\text{RMSNorm}(z - z_r)$; additivity checks use the fixed runtime and normalization denominator.

C.3 COUPLING COMPUTATION

The coupling is solved over components with positive amounts and defined roles by a first-order LP solver (Applegate et al., 2021; Lu and Yang, 2024; Lu et al., 2025) over strictly compatible pairs ($K_{rs} < \kappa$; ties excluded), on the grid $\kappa \in \{0.1, \ldots, 0.9\}$. The operating tolerance and integrated summaries are

$$\kappa^\star = \arg\min_{\kappa \in \mathcal{K}} \left|T(\kappa) - \tfrac{1}{2}\right|, \qquad \bar{x}_r = \sum_k \frac{T(\kappa_k) - T(\kappa_{k-1})}{T(\kappa_{\max})}\, x_r(\kappa_k). \tag{21}$$

$\kappa^\star$ takes T nearest $\frac{1}{2}$ per cell and level, a convention checked at the rungs with T nearest 0.25 and 0.75. The weight-to-activation analysis uses one rung per cell minimizing $\max(|T_W - \frac{1}{2}|, |T_A - \frac{1}{2}|)$.

Analysis scope and reproducibility. Couplings and targets are fit on fitting items only, and every replication fixed its test before evaluation. Undefined controls, zero-effect roles, and gap-unverified solutions are reported separately.

First-order fidelity. In 32-component panels stratified by operator, depth, and amount, first-order effects reproduce exact-removal amount rankings (Spearman ≥ 0.97; Table 2); role rankings agree less, most weakly for Tülu weight groups (0.69), so every trained-model claim in Sections 4.2–4.4 is tested by exact intervention.

Level	Lineage	Amount Spearman	Role-TV Spearman	Role-TV MAE
Activations	OLMo	0.997 [0.989, 0.999]	0.881 [0.843, 0.913]	0.015 [0.012, 0.018]
Activations	Tülu	0.997 [0.989, 0.998]	0.814 [0.744, 0.871]	0.022 [0.017, 0.027]
Weight groups	OLMo	0.990 [0.973, 0.994]	0.837 [0.758, 0.895]	0.018 [0.015, 0.022]
Weight groups	Tülu	0.983 [0.957, 0.991]	0.694 [0.601, 0.787]	0.032 [0.024, 0.040]

Table 2: **First-order effects closely track functional amounts.** Means over four endpoint–lens panels (32 components and 128 exact-evaluation items each), with 95% hierarchical bootstrap intervals.

C.4 Planted relations

We check the recovery claims of Section 3 against a known ground truth.

The controlled endpoints have two pre-LayerNorm Transformer blocks, residual width four, and three tokens; block b adds attention and MLP writes

$$h_t^{(b+1)} = h_t^{(b)} + \sum_j a_{b,j}^A(t) + \sum_j a_{b,j}^M(t), \qquad a_{b,j}^M(t) = \mathrm{GELU}\big((W_{\mathrm{up}}^{(b)} \tilde{h}_t^{(b)})_j\big)\, W_{\mathrm{down}}^{(b)}[:, j]. \quad (22)$$

First-block writes are rewritten as in Table 3; Figure 6 shows each allocation rule, and Sections 3.2 and 3.3 give the comparators.

Construction	Planted relation	What it tests
Allocation (5 relations)	Identity, split, merge, deletion, addition; $M^\star, K^\star \in \{0, 1\}$, $\kappa = 0.5$ given, $\Pi^\star$ withheld	Edge F1 of predicted edges ($\Pi_{rs} > 0$) against planted positive-flow edges
Signed: match	Target $(1, 0, 0, 0)$	Recovery of a preserved role
Signed: positive split	Target $(\frac{1}{3}, \frac{2}{3}, 0, 0)$	Recovery of split support
Signed: opposed distractor	Target $(1, 0, -1, 0)$	Opposed distractor beside a match
Signed: absence	Target $(0, 0, -\frac{1}{3}, -\frac{2}{3})$	Abstention with only opposed candidates

Table 3: Planted constructions with separate 32-item panels for amount, role, and exact evaluation. Signed supports use zero-based slot indices.

Construction	Signed	$\Delta\theta$	CKA
Match	0	0	0
Positive split	$< 10^{-30}$	$< 10^{-30}$	$< 10^{-30}$
Opposed distractor	$< 10^{-30}$	4	1
Absence	1	2.78	4

Table 4: Exact reconstruction error E of the signed-role constructions; the correct empty intervention has $E = 1$ in the absence case.

Each target set is scored by joint removal, normalized by source-effect energy:

$$v_0(x) = R_0(x) - R_0(x; \mathcal{I}_{0,r}), \quad (23)$$

$$v_{1,S}(x) = R_1(x) - R_1(x; \mathcal{I}_{1,S}), \quad (24)$$

$$E(S) = \frac{\sum_{x \in D_{\mathrm{test}}} \big(v_0(x) - v_{1,S}(x)\big)^2}{\sum_{x \in D_{\mathrm{test}}} v_0(x)^2}. \quad (25)$$

Signed coupling recovers the intended support in all four constructions, for MLP and attention writes, while $\Delta\theta$ and CKA fail both opposed-effect tests (Figure 7, Table 4). To summarize, only a signed measure resists an opposed distractor.

Relation	One-to-one	Balanced	Partial
Stay	1.00	1.00	1.00
Split / merge	0.67	1.00	1.00
Disappear / appear	1.00	0.67	1.00

Table 5: **Partial coupling recovers all five planted relations.** Edge F1 with known compatibility. Pairwise thresholding allocates no amounts and appears only in Figure 6; Table 4 gives the signed-role reconstruction errors.

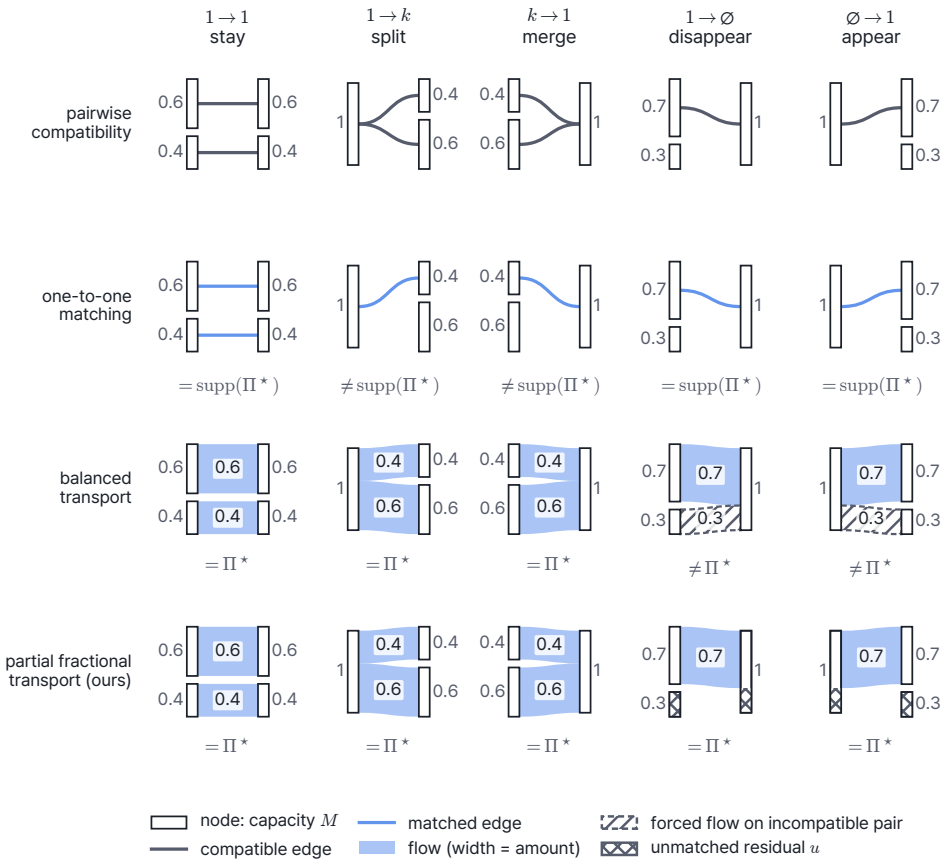


Figure 6: **Four allocation rules on the five planted relations.** Rows: pairwise compatibility, one-to-one matching, balanced transport, and partial fractional transport, given oracle masses and compatible pairs. Ribbon width is allocated mass and hatching unmatched residual; only partial transport recovers all five.

C.5 Planted relations at 7B scale

We check recovery at real scale under realistic role noise. A preregistered test plants relations into OLMo's SFT Instruction activation inventory (353,280 units), permuting units within (layer, class) blocks, splitting a random 10% in two ($\alpha \sim U(0.2, 0.8)$), deleting 5%, and adding 5% as Base copies. Target roles and amounts are perturbed by (i) item bootstrap reweighting (true-pair distance about 0.38) or (ii) per-unit Gaussian noise matching the real same-locus distance at each locus (quantiles

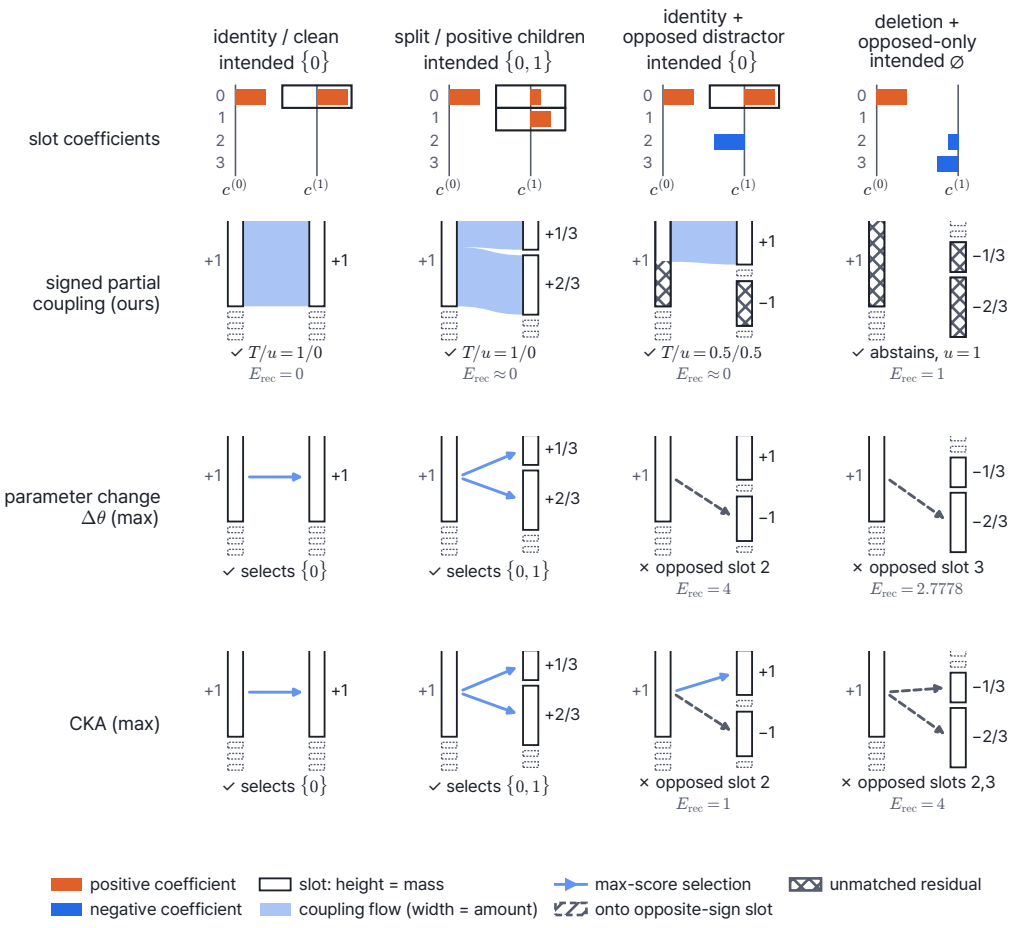


Figure 7: **Target selection as flows.** Columns: the four signed-role constructions (MLP writes; attention is identical); boxes mark the intended support. Only signed partial coupling recovers it, leaving half the mass unmatched beside an opposed distractor.

0.55, 0.61, 0.67), with added units copied from each source's nearest SFT neighbor so distractors sit near true targets.

To summarize, this yields an approximate relocation noise floor at real scale, which Section 4.1 and Appendix D.1 use as a reference.

D FINE-TUNING RESULTS

D.1 MASS ALLOCATION AFTER FINE-TUNING

We check whether Section 4.1's relocation numbers depend on the reporting tolerance, the item sample, or normalization; Figure 8 and Table 6 give the full matched-mass path.

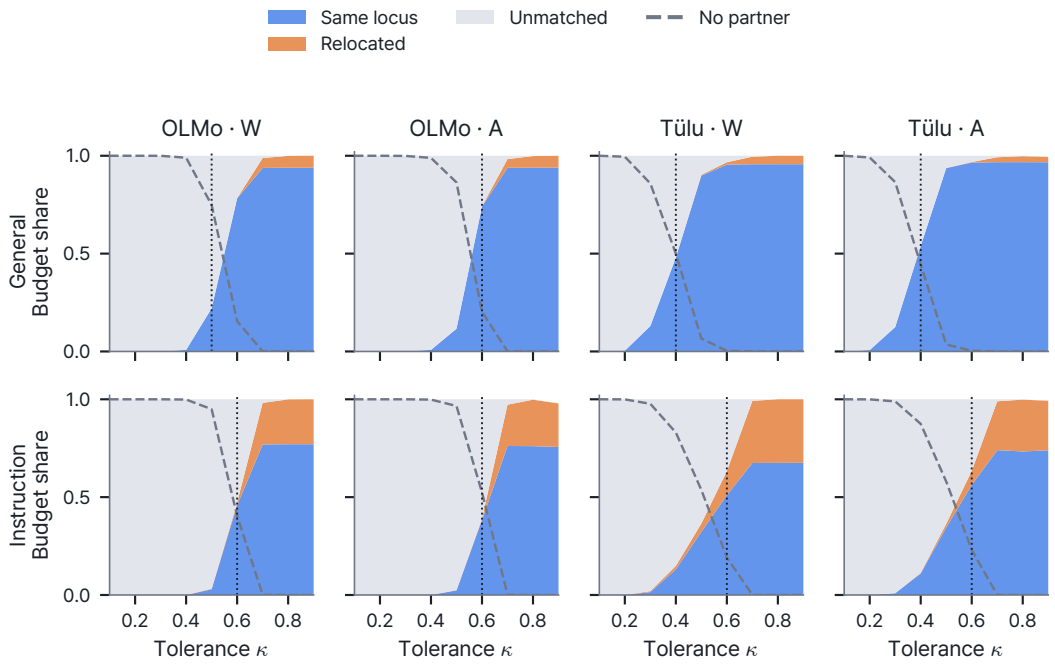


Figure 8: **Relaxing compatibility permits more matches and relocation.** Stacks show the endpoint budget; the line shows source mass without a compatible partner. Hollow points lack the verified gap bound (Appendix C.3).

Table 6: Matched fractions over the tolerance grid. Nine solutions lack the verified gap bound and are marked in Figure 8.

level	cell	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
W	OLMo/G	0	0	0	0.008	0.221	0.784	0.989	0.999	1.000
W	OLMo/I	0	0	0	0.001	0.032	0.472	0.982	0.999	1.000
W	Tülu/G	0	0.004	0.130	0.480	0.903	0.967	0.995	1.000	1.000
W	Tülu/I	0	0	0.020	0.149	0.365	0.633	0.992	1.000	1.000
A	OLMo/G	0	0	0.001	0.008	0.117	0.742	0.983	0.998	1.000
A	OLMo/I	0	0	0	0.001	0.025	0.390	0.973	0.998	0.979
A	Tülu/G	0	0.007	0.125	0.539	0.938	0.968	0.992	0.998	0.995
A	Tülu/I	0	0	0.009	0.113	0.364	0.634	0.990	0.998	0.993

Common tolerances. At a common $\kappa = 0.6$, Tülu Instruction relocates 19% and 11% against at most 3% elsewhere (Table 7); OLMo Instruction leads only at activation $\kappa = 0.5$, where it matches 2.5% of its mass, and relocates 22% at $\kappa = 0.7$.

			$\kappa = 0.5$		$\kappa = 0.6$		$\kappa = 0.7$	
Level	Lineage	Lens	R/T	T	R/T	T	R/T	T
Weight	OLMo	General	0.001	0.221	0.003	0.784	0.051	0.989
Weight	OLMo	Instruction	0.096	0.032	0.030	0.472	0.216	0.982
Weight	Tülu	General	0.005	0.903	0.013	0.967	0.039	0.995
Weight	Tülu	Instruction	0.116	0.365	0.194	0.633	0.320	0.992
Activation	OLMo	General	0.003	0.117	0.002	0.742	0.045	0.983
Activation	OLMo	Instruction	0.068	0.025	0.019	0.390	0.217	0.973
Activation	Tülu	General	0.001	0.938	0.003	0.968	0.025†	0.992†
Activation	Tülu	Instruction	0.051	0.364	0.114	0.634	0.252†	0.990†

Table 7: Relocated share of matched mass (R/T) and matched mass T at common tolerances. †No verified gap bound.

Why mass is unmatched. Unmatched mass splits into mass with no compatible partner,

$$N_0(\kappa) = \sum_{r \in \mathcal{S}_0^+} M_{0,r}\, \mathbf{1}\{\nexists s \in \mathcal{S}_1^+ : K_{rs} < \kappa\}\,,$$
$$N_1(\kappa) = \sum_{s \in \mathcal{S}_1^+} M_{1,s}\, \mathbf{1}\{\nexists r \in \mathcal{S}_0^+ : K_{rs} < \kappa\}\,. \tag{26}$$

and mass left unmatched by competition for capacity; a zero-effect component is an undefined-role state, separate from both. Table 8 splits $1 - T$ at $\kappa^\star$ this way. No-partner mass dominates (52–96% on the source side); competition is appreciable only under the Instruction lens, reaching 0.175 of the endpoint budget for Tülu weight groups.

						Source side		Target side	
Level	Lineage	Lens	$\kappa^\star$	T	$1-T$	N_0	$1-T-N_0$	N_1	$1-T-N_1$
Weight	OLMo	General	0.5	0.221	0.779	0.748	0.031	0.775	0.004
Weight	OLMo	Instruction	0.6	0.472	0.528	0.405	0.123	0.442	0.085
Weight	Tülu	General	0.4	0.480	0.520	0.500	0.020	0.509	0.010
Weight	Tülu	Instruction	0.6	0.633	0.367	0.192	0.175	0.213	0.154
Activation	OLMo	General	0.6	0.742	0.258	0.203	0.055	0.225	0.033
Activation	OLMo	Instruction	0.6	0.390	0.610	0.524	0.086	0.537	0.073
Activation	Tülu	General	0.4	0.539	0.461	0.444	0.017	0.448	0.013
Activation	Tülu	Instruction	0.6	0.634	0.366	0.230	0.135	0.221	0.144

Table 8: Decomposition of unmatched mass at each setting's reporting tolerance.

Computation. Every reported coupling has a verified optimality gap below 10^{-5} except nine rungs outside the reporting tolerances (hollow in Figure 8), which stopped at relative gaps of 2.9×10^{-5} to 0.095.

Operation anatomy. OLMo activation flows show all five motifs (Figure 9), and matching can depend on the lens (Figure 10).

Role distances and their reliability. Same-locus pairs are far closer than others (Table 9): median 0.36–0.61 against 0.73–0.75 for 1,000,000 random off-locus and 1,000,000 same-layer different-unit pairs, under 0.3% of which fall below 0.6. "Stay" thus picks the closest counterpart, not an unchanged role: OLMo Instruction same-locus activation matches lie at 0.51–0.60 (largest-amount match 0.59). Split-half agreement is lowest for the lightest components (0.34–0.84); three item-bootstrap replicates change distances by a median 0.011–0.014 (Spearman 0.90–0.99). Item halves give no noise floor, as disjoint items give disjoint support ($K = 1$).

Grouped-query attention. Llama-3.1-8B shares each key/value head among four query heads, so tying query and key groups by raw row index mispairs heads. Re-tying by head changes Tülu Instruction weight relocation from 19.4% to 19.0%; two thirds of relocated mass comes from attention

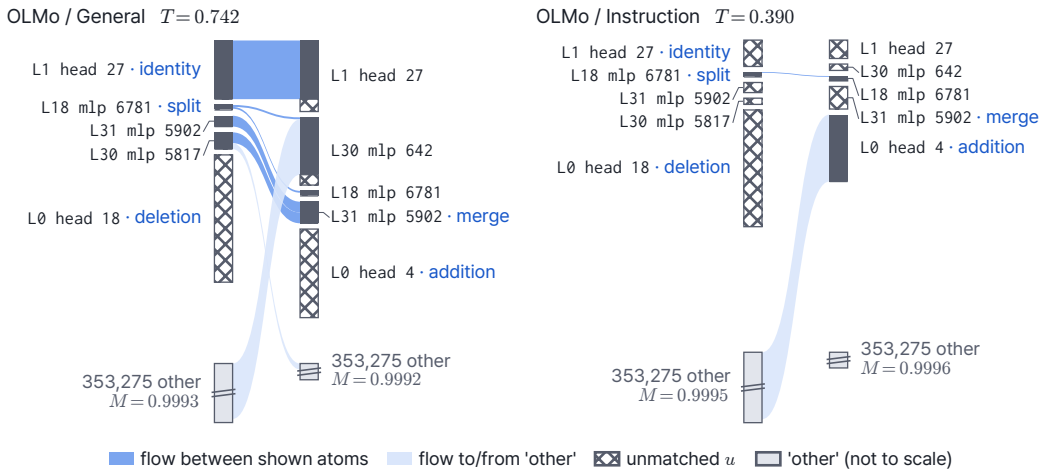


Figure 9: **Operation anatomy for OLMo activations at $\kappa = 0.6$.** The clearest General-lens example of each motif, chosen before plotting, under both lenses. The one-to-one pair (L1 head 27) is almost entirely unmatched under Instruction. Flows come from one pinned LP solution and illustrate motifs rather than certify edges.

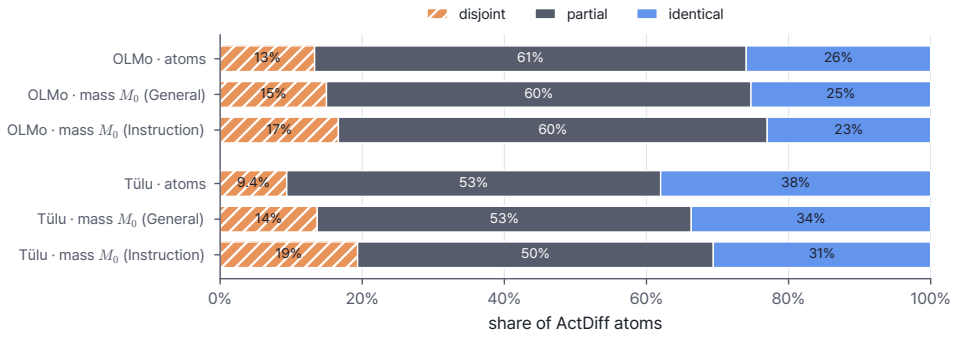


Figure 10: **Lens dependence of activation correspondence at $\kappa = 0.7$.** Among components transported under both lenses, 26% (OLMo) and 38% (Tülu) keep identical targets, 61% and 53% share some, and 13% and 9% have disjoint targets.

groups, about half from query/key groups, and over a third crosses between attention and MLP. OLMo (multi-head attention) is unaffected.

Total amount and normalization. Normalization hides changes in total amount, which rises by up to 80% (OLMo General) and falls by 20% (OLMo Instruction; Table 9), so allocation shares are relative claims. Unnormalized capacities at $\kappa = 0.6$ barely change the Instruction rollback cell (relocation 0.018 versus 0.019; residual top-1% Jaccard 0.875, Spearman 0.958; recovery 0.853 versus 0.896, with less agreeing-item drift). Under OLMo General the binding side switches (79% of Base mass saturates capacity) and the ranking diverges (Jaccard 0.526, Spearman 0.788).

Cell	Level	SFT/Base amount	Same-locus K (mass-wtd)	Off-locus	Same-locus < 0.6	Split-half ρ	$\lvert K_A - K_B \rvert$
OLMo General	A	1.80	0.557 (0.559)	0.750	0.81	0.69	0.031
OLMo General	W	1.75	0.546 (0.543)	0.739	0.82	0.81	0.029
OLMo Instruction	A	0.80	0.608 (0.617)	0.744	0.44	0.74	0.030
OLMo Instruction	W	0.78	0.601 (0.607)	0.734	0.50	0.85	0.027
Tülu General	A	1.29	0.361 (0.380)	0.743	1.00	0.88	0.028
Tülu General	W	1.28	0.361 (0.400)	0.734	1.00	0.90	0.027
Tülu Instruction	A	1.15	0.486 (0.566)	0.734	0.75	0.94	0.027
Tülu Instruction	W	1.09	0.491 (0.577)	0.734	0.72	0.95	0.026

Table 9: Normalization and role-distance reliability. Off-locus uses 1,000,000 random pairs; split-half ρ is the Spearman correlation of same-locus distances computed on two disjoint halves of the fitting items. A: activation units; W: weight groups.

Relocation on item halves. Refitting both Instruction weight couplings at $\kappa = 0.6$ on each seeded half of the fitting items gives closely agreeing halves (Table 10), with higher relocation from noisier 384-item roles, yet Tülu still relocates about five times OLMo's share.

Setting	Items	T	R/T	R/T (GQA ties)
Tülu Instruction, W	all 768	0.633	0.194	0.190
	half A / half B	0.665 / 0.669	0.253 / 0.251	0.248 / 0.246
OLMo Instruction, W	all 768	0.472	0.030	0.030
	half A / half B	0.475 / 0.488	0.046 / 0.050	–

Table 10: Weight-group relocation at $\kappa = 0.6$ refit on each half of the fitting items.

Spurious relocation on planted relations. On the 7B planted inventories (Appendix C.5) at matched mass near one half, realistic role noise sends 3.5–12% of matched flow to a wrong target. Tülu Instruction's weight relocation (19%) exceeds this floor, its activation relocation (11%) lies within it, and all other settings (at most 3%) lie below; the real floor may be higher, as planted inventories have fewer near-ties. To summarize, Tülu's weight-level relocation survives the checks run on it (item halves, grouped-query tying) and clears the planted floor, though the normalization and low-mass checks cover OLMo cells only; OLMo's relocation stays below the floor and is best read as staying in place.

D.1.1 Reorganization during SFT

We follow the OLMo SFT run itself: Instruction-lens activation effects on all 768 items at the public Olmo-3-7B-Think-SFT checkpoints (steps 1k, 5k, 10k, 21k, 32k) and the final model, each coupled to Base at $\kappa = 0.6$, under a protocol fixed before any outcome.

Numerics. A 16-item parity check matched the reported bank on role items (relative $L_1 \leq 6.5 \times 10^{-5}$) but found its mass items in bfloat16; the trajectory recomputes Base and Final in float32, and the reported bank gives the same result ($T = 0.391$ and identity flow 0.383, against 0.389 and 0.382).

Result. OLMo reorganizes in place throughout SFT (Table 11, Figure 11): relocated flow stays below 1%, mass shifts mostly in the first 1k steps, and same-locus role distance rises steadily (median 0.42 to 0.61; off-locus 0.745), so matched mass falls from 0.91 to 0.39 as roles cross the tolerance. Each later segment matches 96–98% of its mass in place, with relocation below 0.05%.

Forecasts. Both prespecified forecasts hold. The top 1% of target-unmatched units at 5k make up 0.51 of the final top 1% (mass-matched random: 0.18, $p < 0.001$, 1000 draws; 5k mass alone: 0.47, so the gain over mass is small). Chaining Base→5k with 5k→21k reproduces the direct Base→21k targets (mass-weighted row TV 0.008; row-permutation null 1.00) and keeps more mass matched ($T = 0.74$ against 0.51), since each segment's role change stays within tolerance. To summarize, SFT refines OLMo in place, and the coupling forecasts and composes along the run.

Checkpoint	$\sum_r a_r$	TV(M_{Base}, M_c)	T	Relocated	Same-locus K (median / mass-weighted)	Off-locus K
Base	7.96	–	–	–	–	–
1k	6.46	0.071	0.907	0.003	0.416 / 0.433	0.744
5k	5.73	0.091	0.741	0.002	0.524 / 0.538	0.746
10k	5.51	0.095	0.613	0.006	0.558 / 0.568	0.746
21k	5.33	0.099	0.506	0.009	0.584 / 0.591	0.745
32k	6.46	0.094	0.391	0.006	0.607 / 0.613	0.747
Final	6.35	0.096	0.389	0.008	0.608 / 0.612	0.744

Table 11: Base→checkpoint couplings along OLMo SFT (Instruction lens, activations, $\kappa = 0.6$). T is matched mass; relocated is matched mass outside the same unit.

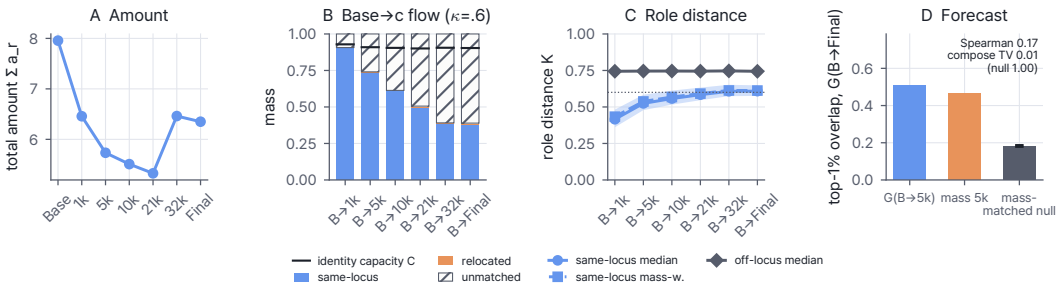


Figure 11: **OLMo SFT retunes roles in place.** (A) Total amount per checkpoint. (B) Base→checkpoint flow at $\kappa = 0.6$ (same-locus, relocated, unmatched); the line marks the identity capacity $\sum_r \min(M_{0,r}, M_{1,r})$. (C) Same-locus role distance rises toward the tolerance (dotted); off-locus stays flat. (D) Top-1% overlap of target-unmatched mass at 5k and final, against the 5k-mass baseline and a mass-matched null.

D.2 Cross-level consistency

We check whether the activation coupling predicts where weight packets act, at the joint tolerance of Appendix C.3. Eight disjoint packets per setting cover the highest-flow 1% of transported weight mass; their predicted activation footprints are compared with those observed after exact SFT-weight removal, and the predicted support is restored in the same pass. Specificity uses the exact 8! permutation test; recovery is compared with matched random support and an outcome-informed oracle.

Cell	$\bar{D}_{\text{diag}}$	$\bar{D}_{\text{off}}$	ΔD	top-1	p	γ_F	γ_R	γ_O
OLMo General	0.239	0.440	0.201	8/8	2.5×10^{-5}	0.994	0.714	1.000
OLMo Instruction	0.503	0.628	0.125	5/8	2.5×10^{-5}	0.925	0.722	0.999
Tülu General	0.252	0.604	0.352	6/8	2.5×10^{-5}	0.983	0.754	1.000
Tülu Instruction	0.368	0.517	0.149	4/8	5.2×10^{-4}	1.000	−0.168	1.000
Prospective replication: new packets (1–2% flow band), unseen items								
OLMo General	–	–	0.192	8/8	2.5×10^{-5}	0.981	0.715	0.999
OLMo Instruction	–	–	0.152	7/8	2.5×10^{-5}	0.966	0.765	0.999
Tülu General	–	–	0.321	8/8	2.5×10^{-5}	0.988	0.743	1.000
Tülu Instruction	–	–	0.269	5/8	5.5×10^{-4}	0.998	0.642	1.000

Table 12: Cross-level prediction and restoration, including the prospective replication. Recovery is averaged over eight packets per setting; dashes mark distances omitted from this summary.

Restoration is within 0.08 of the oracle in every cell (Table 12, Figure 12); the high-flow packet rule defines the scope of this test.

Support size and placement. The random support (Table 13) matches the predicted set's MLP-channel and head counts, drawn from all layers as the closest of 100 draws in SFT activation mass; it is not layer-matched. Units upstream of the earliest removed layer have no observed footprint. The last column is a coverage proxy (predicted, random, and 20 layer-histogram-matched draws) that understated the prediction's measured advantage (Table 14).

Cell (original packets)	$\lvert F\rvert/N$ median [max]	Upstream units, F / random (median)	Footprint coverage F / random / layer-matched
OLMo General	0.33 [0.43]	0 / 0.05	0.83 / 0.47 / 0.86
OLMo Instruction	0.11 [0.23]	0 / 0	0.54 / 0.25 / 0.52
Tülu General	0.33 [0.43]	0 / 0.36	0.88 / 0.35 / 0.87
Tülu Instruction	0.01 [0.46]	0.04 / 0.78	0.82 / 0.18 / 0.60

Table 13: Restored-set size and layer placement in the cross-level assay. N is the number of SFT activation units (353,280 for OLMo and 459,776 for Tülu). About 99% of restored units are MLP channels.

Layer-matched random restoration. A preregistered follow-up restored, for all 64 packets, three random supports closest to F in SFT activation mass among 100 hash-seeded draws: (i) F's (layer, MLP/head) histogram, overlap allowed; (ii) the same histogram outside F (undefined when the pool is too small); (iii) F's operator histogram from layers at or after the earliest removed one. F beats the informative disjoint arm wherever it exists (18/18 original, 16/16 replication), including 12 packets with at most 25% nearest-layer fill. The overlap-allowed arm shares a median 74% of units with F and restores nearly as much; the disjoint arm is undefined for all OLMo General packets and the largest Tülu packets, where F covers a third or more of all units.

Packets	F	Layer-matched	Layer-matched disjoint	Downstream	Original random	Oracle
Original (32)	0.976	0.920 [21/32]	0.428 [18/18]	0.809 [25/29]	0.505	1.000
Replication (32)	0.983	0.957 [20/32]	0.288 [16/16]	0.881 [25/32]	0.716	1.000

Table 14: Mean restoration recovery with layer-matched random supports. Brackets: packets where F beats the arm, over packets where it is defined. Preregistered primary contrast (F versus the disjoint arm, else the overlap-allowed arm): 23/32 in both sets (sign test $p = 0.02$ each).

Packets beyond the highest-flow band. A preregistered extension drew eight equal-mass packets per setting from a mid band (10–20% flow) and a random band (outside the top 2%) (Table 15). Restoration is as complete as in the top band and beats random in every packet, and specificity holds everywhere (exact 8! test, $p = 2.5 \times 10^{-5}$). The disjoint layer-matched control, defined only for the 8 OLMo Instruction packets, loses all of them (0.96–0.98 versus 0.25–0.34).

Band (16 packets each)	γ_F	γ_R	γ_O	$F > R$
Top (original / replication)	0.976 / 0.983	0.505 / 0.716	1.000	32/32
Mid (10–20% flow)	0.985	0.776	1.000	16/16
Random (outside top 2%)	0.980	0.746	1.000	16/16

Table 15: Cross-level restoration by flow band, pooled over the four settings. Sign tests: $p = 3.1 \times 10^{-5}$ for each new band.

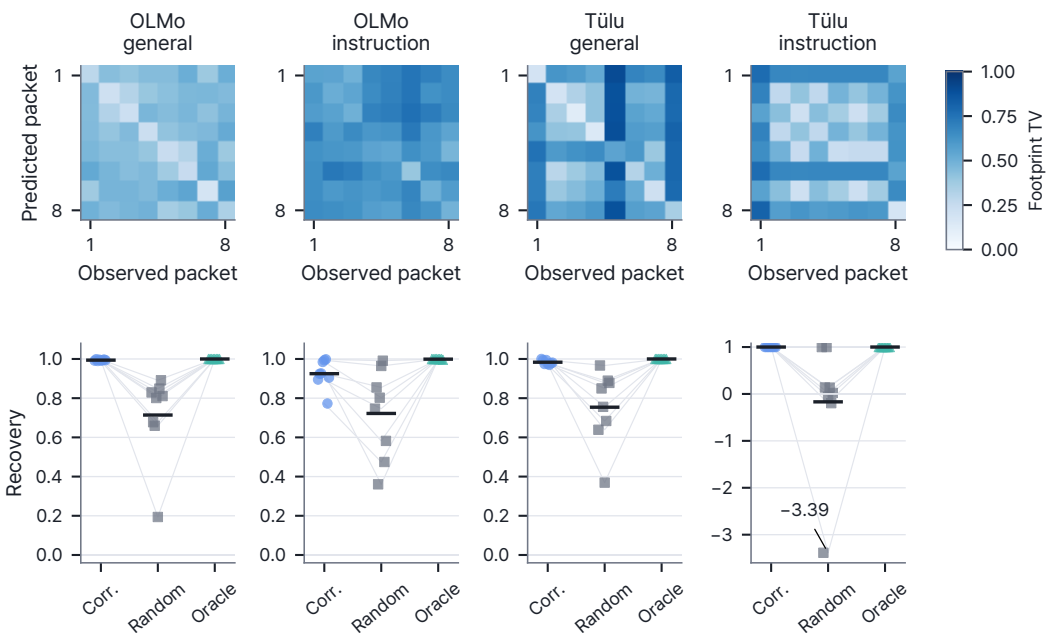


Figure 12: **Predicted footprints identify weight packets, and restoration recovers the removal effect.** Top: predicted–observed distances; bottom: recovery per packet and control. Median random recovery is 0.81, 0.78, 0.80, and 0.08 (OLMo General, OLMo Instruction, Tülu General, Tülu Instruction); the Tülu Instruction mean of -0.17 is driven by packet 7 (-3.39), without which it is 0.29.

The prospective replication fixes coupling, tolerance, footprints, and controls, and uses the next 1–2% flow band and new items (the Instruction source-profile panel is shared with the rollback replication). Both preregistered criteria hold: specificity ($\Delta D > 0$, exact $p < 0.05$), required in three of four settings, holds in all four; recovery beats random in all 32 packets (two-sided sign test $p = 4.7 \times 10^{-10}$; 28/32 ranked first). To summarize, the prediction extends to new packets, items, and lower flow bands, and beats cardinality-matched and disjoint layer-matched random supports wherever they are defined; it wins fewer packets against the overlap-allowed and downstream arms (Table 14).

D.3 Causal target selection and capacity-aware allocation

We give the protocol behind Section 4.3's target-selection results: multi-unit supports, singleton placement, and allocation under competition.

Metric hierarchy and descriptive scope. Multi-unit reconstruction prespecified normalized squared error E (Equation (14)); effect-shape error E_{shape}, added post hoc to separate pattern from dose, is the reconstruction error left after the best non-negative rescaling of the target effect (Equation (17)), so it is 1 for an empty, orthogonal, or opposed effect and 0 for a positively proportional one, as in Appendix E. Singleton placement uses signed-role TV (primary) and normalized-L_1 error (secondary). Win counts are descriptive, since cases within a setting share model pair, lens, and panel.

D.3.1 Multi-unit supports

We check whether correspondence finds counterparts spanning several units. For a positive-flow source r, let $B_r = \sum_s \widehat{\Pi}_{rs}$, $p_r(s) = \widehat{\Pi}_{rs}/B_r$, and $S_{\text{split}}(r) = 1 - \max_s p_r(s)$; $U_F(r)$ is the shortest descending-flow prefix with mass at least $0.95B_r$. Excluding identity-calibration sources, each setting takes the four with largest split score and $|U_F| \geq 2$. Representation matching uses centered linear CKA (activations) or signed weight-slice cosine (weights). All arms share budget

$B = \min(B_{r,95}, \min_m C_m)$, where $B_{r,95} = \sum_{s \in U_F(r)} \widehat{\Pi}_{rs}$ is the flow into the 95% support and $C_m = \sum_{s \in S_m} M_{1,s}$ is the target amount of arm m's support S_m, and use source and target doses $B/M_{0,r}$ and B/C_m.

Across the 32 sources, post-hoc E_{shape} wins are 28/32 against representation, 31/32 against random, and 23/32 against same-locus identity supports (22, 25, and 17 among the 26 supports that leave native channels); the prespecified E gives 29, 30, and 26. To summarize, correspondence finds multi-unit counterparts, and the prespecified metric agrees with the post-hoc one.

D.3.2 Singleton placement

We isolate target choice with single-unit targets: eight Coarse16 sources per setting (16 for activations), each arm placing the same transported amount with the same capacity-fill rule, scored by signed-role TV and normalized-L_1 error

$$E_{L_1}(S) = \frac{\sum_{x \in D_{\text{test}}} |v_0(x) - v_{1,S}(x)|}{\sum_{x \in D_{\text{test}}} |v_0(x)|}, \tag{27}$$

(undefined for a zero source denominator). Role-distance placement beats parameter change and random controls in all 32 weight and all 64 activation placements; normalized-L_1 agrees for weights (31/32 and 32/32). Excluding the 13 weight placements on the source's own group, wins remain 19/19. To summarize, signed role alone solves singleton placement.

D.3.3 Joint allocation under shared capacity

We ask whether capacity matters when many sources compete for targets. Independent nearest-compatible assignment over-claims 0.02 to 0.46 of the endpoint budget across the eight settings, with as many as 10,585 sources selecting one target (Figure 13a).

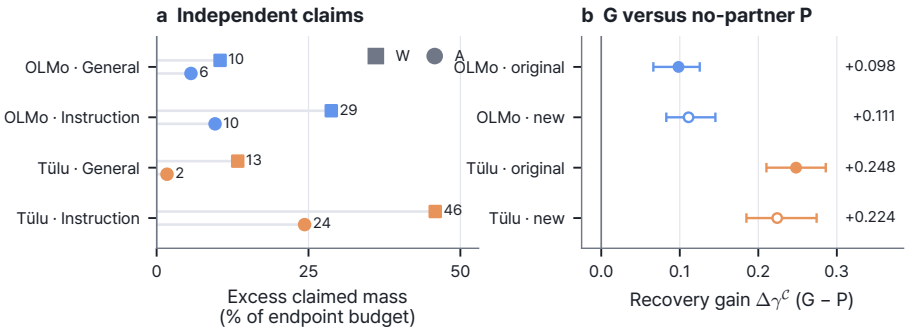


Figure 13: **Capacity matters for a coherent allocation and for the rollback ranking.** (a) Excess target mass claimed by independent nearest-compatible assignment at $\kappa^\star$, in % of the endpoint budget. (b) Paired recovery gain of G over no-partner mass P at the 1% budget, with 95% item-bootstrap intervals.

Stability of the residual ranking across LP optima. The pipeline pins the first certified feasible primal without claiming uniqueness. For the rollback cell (OLMo Instruction activations, $\kappa = 0.6$; 508,395 compatible pairs), 23 preregistered re-solves (cold start, 10 relabelings, 10 cost jitters with $\epsilon = 10^{-9}$, warm starts from $\kappa = 0.5$ and 0.7) are optimal and identical (top-1% Jaccard 1.0, matched mass 0.3906), and match the pinned ranking at Spearman 0.9994 (top-1% and top-5% Jaccard 0.997 and 0.998); the pinned objective is 1.2×10^{-4} below the best, within tolerance. Three re-executed activation cells gave bitwise-identical couplings.

Low-mass components. The lightest components have the least reliable roles, but the lightest Base-mass decile holds only 2–6% of Base mass (Figure 14). Dropping the lightest 10, 25, or 50% of units changes rollback-cell matched mass by -2%, -7%, and -12% and relocation by at most

+0.011; the residual ranking keeps Spearman 0.993, 0.986, and 0.966 (top-1% Jaccard 0.996, 0.989, 0.970), and no reported top-1% rollback unit is dropped. To summarize, in the rollback cell the residual ranking survives LP non-uniqueness and low-mass noise.

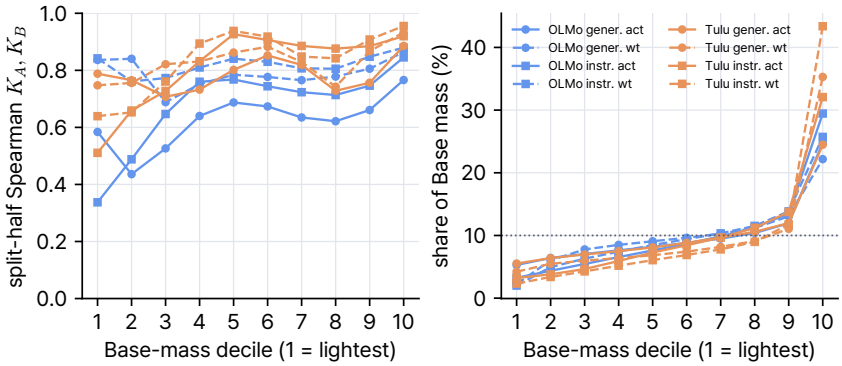


Figure 14: **Role reliability rises with component mass, and light components carry little mass.** Left: split-half Spearman correlation of same-locus role distances by Base-mass decile. Right: share of Base mass per decile.

D.4 OLMo unmatched-mass rollback

We check Section 4.4's claim that, in OLMo, target-unmatched mass G ranks restoration targets better than SFT amount M_1. Rollback restores Base activation writes at the top 1% of SFT units (Instruction lens, $\kappa^\star = 0.6$). Changed items have Base–SFT differences at least the panel median; $\gamma^{\mathcal{C}}$ is the fraction reversed, and $\delta^{\mathcal{A}}$ is agreeing-item movement away from SFT relative to the Base–SFT gap.

Prospective replication. With coupling, rankings, budget, and changed-item rule fixed, G recovers 0.889 versus 0.827 for M_1 on 128 new OLMo Instruction items (+0.062 $[0.026, 0.097]$), with agreeing-item displacement 1.011 versus 1.307. No-partner mass $P_s = M_{1,s}\mathbf{1}[\min_r K_{rs} \geq \kappa^\star]$ ignores capacity competition; at 1%, $G - P$ recovery is +0.098 $[0.066, 0.125]$ (OLMo) and +0.248 $[0.210, 0.286]$ (Tülu) on held-out items and +0.111 $[0.083, 0.145]$ and +0.224 $[0.185, 0.274]$ on new items.

Which units the rankings select. At 1% (3,533 units), G and M_1 share 1,679. G takes 737 heads and 2,796 MLP channels, mostly early and middle (913, 1,537, 459, and 624 in layers 0–7, 8–15, 16–23, 24–31); M_1 takes 998 heads and 2,535 channels, 2,177 in layers 16–31 (496, 860, 1,033, 1,144); raw change D is almost all late MLP channels (53 heads; 2,699 units in layers 24–31). The design does not isolate what causes G's smaller General-lens loss.

Tolerance dependence of the rollback ranking. Across $\kappa \in \{0.5, 0.6, 0.7\}$ and path-integrated scores (Figure 15), G needs intermediate matched mass: at $\kappa = 0.7$ matched mass reaches 0.97–0.99, G approaches zero, and OLMo recovery falls to 0.12. Relocated incoming plus unmatched mass ranks stably (0.81–0.89), and integrated G needs no choice of $\kappa^\star$ (Table 16).

Greedy and capacity-free residuals across tolerances. A preregistered check recomputed the residual per tolerance with greedy allocation and with capacity-free nearest-role assignment, whose residual is unclaimed target mass (Table 16). Greedy reproduces the LP up to $\kappa = 0.6$; at $\kappa = 0.7$ both capacity-respecting residuals vanish, while nearest-role sends 31% of Base mass to already-filled targets and its residual recovers 0.902.

Greedy instead of LP allocation. A frozen comparison computes $G^{\mathrm{g}}_s = M_{1,s} - \sum_r \widehat{\Pi}^{\mathrm{g}}_{rs}$ by a greedy pass over all compatible edges ($K_{rs} < \kappa^\star$; 0.5M for OLMo, 129M for Tülu) in ascending cost, each assigned the smaller of remaining source mass and target capacity. It matches nearly as much mass as the LP (OLMo 0.3903 versus 0.3903; Tülu 0.628 versus 0.634) and ranks targets

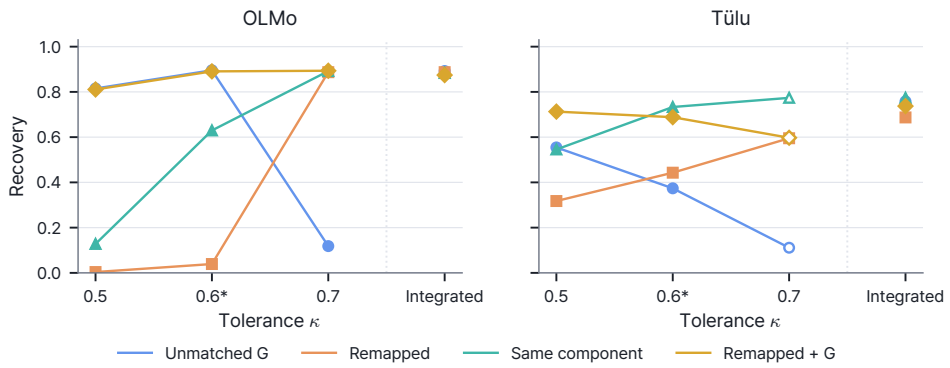


Figure 15: **Rollback recovery across tolerances.** Changed-item recovery at 1% for G, relocated incoming ("Remapped" in the legend), same-component, and relocated plus unmatched mass, per tolerance and path-integrated. Hollow markers lack a verified gap bound.

κ	Method	Top-1% Jaccard with LP	γ^C	LP minus method [95%]
0.5	LP / greedy / nearest-role	1 / 0.999 / 0.988	0.815 / 0.816 / 0.812	– / −0.001 / +0.003 [0.002, 0.004]
0.6	LP / greedy / nearest-role	1 / 0.996 / 0.884	0.896 / 0.896 / 0.886	– / 0.000 / +0.010 [−0.005, 0.026]
0.7	LP / greedy / nearest-role	1 / 0.439 / 0.042	0.118 / 0.108 / 0.902	– / +0.011 / −0.783
integrated	LP / greedy	1 / 0.845	0.892 / 0.893	–

Table 16: Rollback ranking across tolerances by allocation rule (OLMo Instruction activations, 1% budget, held-out items).

almost identically (Spearman 0.999 and 0.974; top-1% Jaccard 0.996 and 0.948). OLMo rollbacks are equivalent; in Tülu greedy recovers slightly more, with more agreeing-item drift (Table 17). Capacity, not LP optimality, is what matters.

Lineage	Items	Budget	$\gamma^C(G)$	$\gamma^C(G^g)$	$G - G^g$ [95%]	$\delta^{\mathcal{A}}$: $G - G^g$ [95%]
OLMo	held-out	1%	0.896	0.896	−0.000 [−0.002, 0.002]	+0.006 [0.004, 0.008]
OLMo	new	1%	0.889	0.891	−0.002 [−0.004, 0.000]	+0.003 [0.001, 0.005]
OLMo	held-out	5% / 10%	0.871 / 0.861	0.876 / 0.866	−0.005 / −0.005	+0.012 / +0.013
Tülu	held-out	1%	0.374	0.406	−0.032 [−0.045, −0.019]	−0.053 [−0.081, −0.022]
Tülu	new	1%	0.319	0.350	−0.031 [−0.045, −0.018]	−0.021 [−0.058, 0.021]
Tülu	held-out	5% / 10%	0.437 / 0.489	0.468 / 0.527	−0.031 / −0.038	−0.035 / −0.038

Table 17: Rollback with the LP residual G versus the greedy residual G^g, with paired item-bootstrap intervals (10,000 resamples); negative $\delta^{\mathcal{A}}$ differences mean G drifts less. A 1% re-run of G reproduced outcomes bitwise.

Behavioral marker and General-lens change. Greedy generation on the 128 original OLMo Instruction prompts restores Base writes at every position; a separate 128-item General panel measures log-probability change relative to SFT (Table 18; 10,000 paired item-bootstrap resamples). Removing the opening tag need not remove the reasoning register or imply broad capability reversal.

Ranking	Tags / 128	γ^C	$\delta^{\mathcal{A}}$	General change [95% interval]
G	0	0.896	1.009	−0.047 [−0.054, −0.041]
M_1	2	0.813	1.291	−0.093 [−0.103, −0.083]
Raw change	31	0.381	0.440	−0.066 [−0.073, −0.060]
Random (5 draws)	108.8	0.158	0.169	0.004 [0.002, 0.006]

Table 18: OLMo activation rollback at the primary 1% budget. General-lens change is in nats per scored byte relative to SFT; SFT has 126 opening tags and Base has zero.

To summarize, in OLMo the G ranking survives prospective replication, matches or slightly exceeds a capacity-free residual up to $\kappa = 0.6$, keeps its recovery when integrated over tolerances, and its gain corresponds to a behavioral rollback with less General-lens loss than amount ranking.

D.4.1 BEHAVIOR IN TÜLU

We check whether rollback generalizes to Tülu, where amount, not G, wins on the readout. With Tülu's sealed 1% selections, restoring nothing reproduces SFT generations exactly and restoring everything moves the first-token distribution to Base. We use a fixed subset of 256 IFEval prompts and 200 seeded GSM8K problems (exact match), with greedy generation up to 1,024 tokens; for OLMo we did not run GSM8K, the same-component arm, or permanent weight reversion.

Every structured ranking removes most instruction following, but only G keeps generation intact: under M_1 and same-component restoration most responses degenerate, as under Base, and GSM8K accuracy collapses, whereas under G it stays near SFT (Table 19). To summarize, in Tülu G removes instruction following selectively, whereas amount ranking reverts the model broadly.

	IFEval (256)			GSM8K (200)		
Condition	strict acc.	degenerate	length	accuracy	degenerate	length
SFT	0.691	0.035	270	0.700	0.005	214
Base	0.160	0.961	997	0.090	1.000	1024
G (unmatched)	0.195	0.066	146	0.655	0.030	219
M_1 (amount)	0.176	0.887	931	0.210	0.940	976
I (same component)	0.176	0.703	766	0.295	0.825	774
Random	0.586	0.055	224	0.760	0.005	200

Table 19: Tülu activation rollback at the 1% budget. Paired prompt-bootstrap contrasts: IFEval accuracy $G - M_1 = +0.020$ $[-0.035, 0.074]$, degeneration $G - M_1 = -0.82$ $[-0.87, -0.77]$; GSM8K accuracy $G - M_1 = +0.445$ $[0.370, 0.520]$, $G - I = +0.360$ $[0.280, 0.440]$, $G - \text{SFT} = -0.045$ $[-0.120, 0.030]$, $G - \text{random} = -0.105$ $[-0.175, -0.035]$.

D.5 ITEM-WISE ROLES: QUALITATIVE EXAMPLES

A role is a signed profile over items, so a correspondence can be read item by item. The examples come from OLMo Instruction activations at $\kappa^\star = 0.6$; values are signed roles ($\times 100$).

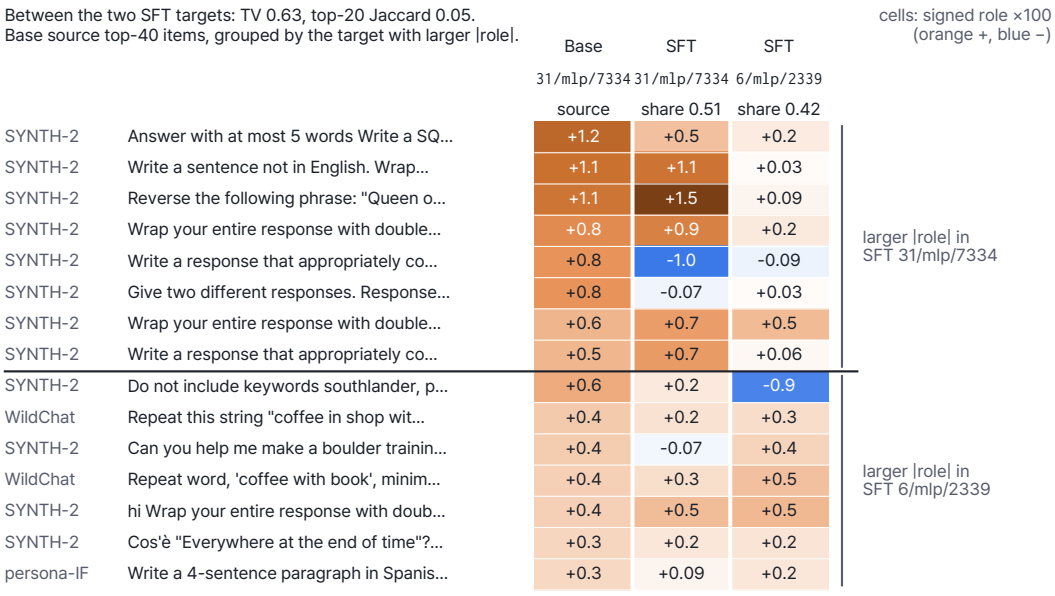


Figure 16: **One Base unit split across two SFT units.** The coupling sends 0.51 of Base channel 31/7334's mass to its SFT copy and 0.42 to SFT channel 6/2339. The targets differ (TV 0.63; top-20 item Jaccard 0.05) and carry different subsets of the source's items, mostly constraint-following prompts in both.

To summarize, the examples show a split that respects item structure, an added function that rollback independently flags, and a same-locus match whose role drifts.

Largest target-unmatched SFT unit (G rank 0; inflow 3% of its mass).
TV(SFT role, Base role of the same head) 0.72. Empty <think></think> in 4/15 rows (3% of panel).

cells: signed role ×100
(orange +, blue −)

		0/head/18 SFT	0/head/18 Base	
SYNTH-2	Do not include keywords southlander, p...	+3.7	+0.7	
persona-IF	PROVIDE 3 LANDMARKS IN AMERICA THAT FE...	+1.9	-0.3	
SYNTH-2	Wrap your entire response with double...	-1.4	+0.2	
persona-IF	In ENGLISH and all CAPITAL letters, li...	-0.8	+0.02	
SYNTH-2	give 10 random examples of sentences s...	-0.8	+0.4	
SYNTH-2	Answer the following question: Remy ha...	-0.7	-0.2	
WildChat	What's your knowledge cutoff date? [as...	-0.7	-0.1	empty <think>
persona-IF	Describe how to make homemade tortilla...	+0.7	+0.03	
WildChat	Hello, ChatGPT. In 2014, Ukraine was c...	-0.6	-0.2	empty <think>
persona-IF	Compose a two-sentence social media po...	-0.6	-0.09	
WildChat	Who is Tucker Carlson? [assistant] <th...	-0.5	+0.02	empty <think>
persona-IF	Repeat the following prompt exactly as...	-0.5	+0.1	
WildChat	Make a vividly detailed and romantic f...	-0.5	+0.03	empty <think>
WildChat	How is "Tatneft" written in Russian? [...	+0.5	-0.00	
persona-IF	In no more than 100 words, explain the...	-0.5	-0.08	

Figure 17: **What fine-tuning added.** SFT head 0/18 has the largest target-unmatched mass: 3% of its mass is matched, and its role differs from its Base role (TV 0.72). Its top items are enriched for replies with an empty <think> block (4 of the top 15, against 3% of the panel), consistent with Section 4.4.

Same head matched to itself (flow rank 1). TV 0.60, of which sign-flipped items 0.35.
Positive role mass 0.71 → 0.34; persona-IF signed mass +0.21 → -0.20. Rows: 15 largest |Δrole|.

cells: signed role ×100
(orange +, blue −)

		16/head/14 Base	16/head/14 SFT	
persona-IF	Write 2 possible cheers in support of...	+0.6	-0.3	sign flip
persona-IF	List 3 PORTLAND TIMBERS OR TRAIL BLAZE...	+0.6	-0.2	sign flip
WildChat	Etsy Digital Marketing, minimalist ton...	+0.4	-0.3	sign flip
SYNTH-2	Translate to Romanian: Emotions have s...	+0.3	-0.3	sign flip
WildChat	How is "Tatneft" written in Russian? [...	+0.4	-0.1	sign flip
persona-IF	Name three classic jazz albums suitabl...	+0.1	-0.4	sign flip
SYNTH-2	Translate to Romanian: Emotions have s...	+0.3	-0.3	sign flip
persona-IF	List three essential safety steps for...	+0.5	-0.04	sign flip
SYNTH-2	At the end of your response, please ex...	-0.03	-0.6	
SYNTH-2	Your entire response should be in Engl...	+0.1	-0.4	sign flip
WildChat	Generate a popular random social media...	+0.6	+0.1	
persona-IF	Describe how using different amounts o...	+0.5	-0.07	sign flip
SYNTH-2	Do not include keywords southlander, p...	-0.4	+0.1	sign flip
persona-IF	Name up to three native bird species o...	+0.3	-0.2	sign flip
SYNTH-2	In your entire response, refrain from...	-0.09	+0.4	sign flip

Figure 18: **Same-locus, changed role.** Head 16/14 is matched to itself with the second-largest flow in the coupling, yet its role moves by TV 0.60, of which sign-flipped items contribute 0.35. The flips span item families; the largest family-level shift is on persona instruction-following items ($+0.21 \rightarrow -0.20$).

E CROSS-ARCHITECTURE TARGET RANKING

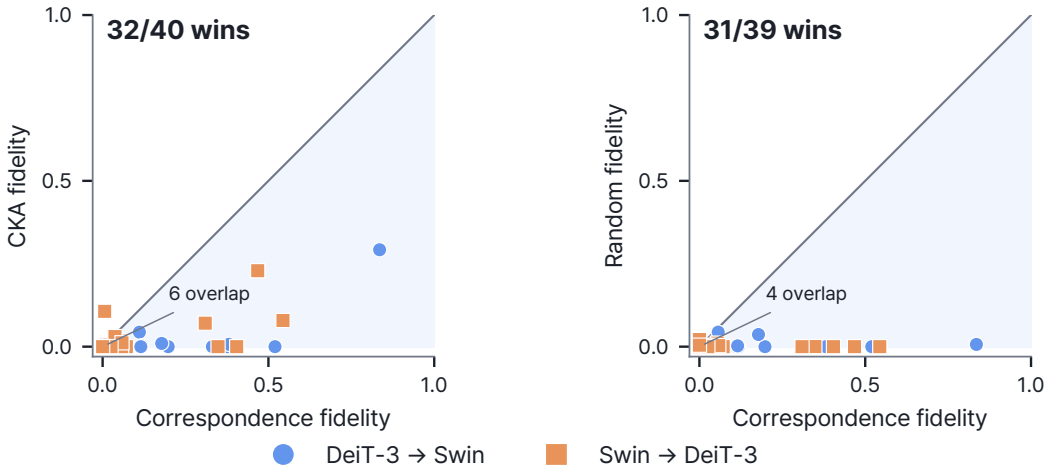


Figure 19: **Correspondence improves cross-architecture target ranking, while absolute fidelity remains low.** DeiT-3–Swin shape fidelity $1 - E_{\text{shape}}$ on held-out images; points below the diagonal favor correspondence.

We give the protocol behind Section 5 and the controls that attribute the transfer to signed roles and, under competition, to capacity. The MLP channels of DeiT-3 (`timm/deit3_base_patch16_224.fb_in22k_ft_in1k`; 36,864) and Swin (`timm/swin_base_patch4_window7_224.ms_in22k_ft_in1k`; 48,128) share no coordinates. Both classify 50 ImageNet classes in 10 families (28 fitting and 8 held-out images per class) under correct-class log-probability. At $\kappa^\star = 0.3$, the 20 sources with most allocated mass per direction are compared across the 95%-flow correspondence support, equally many targets ranked by centered linear CKA on fitting images, and random targets from the 50 components nearest each selected target in endpoint mass, removed exactly at a common budget on 400 held-out images.

Correspondence has lower effect-shape error than CKA in 32/40 cases and random in 31/39 defined cases (a random support is undefined when B_{eval} exceeds its capacity).

Coupling-free nearest-signed-role control. A frozen additional arm takes, for source r, the $N = |U_F(r)|$ defined, positive-amount targets with smallest K_{rs} (ties by unit ID), per source and without capacity. It uses the same budget B_{eval} (Equation (19)) with target dose $\min(B_{\text{eval}}/C_{\text{NN}}, 1)$, where C_{NN} is its support's target amount; six cases reach the cap. Sharing 35% of units with correspondence on average, it matches correspondence and beats CKA (Table 20): the gain comes from signed roles, not the coupling.

Direction	Corr. vs NN (win/tie/loss)	Median E_{shape} corr. / NN	Corr. vs CKA wins	NN vs CKA wins
DeiT-3→Swin	6/2/12	0.887 / 0.873	18/20	19/20
Swin→DeiT-3	8/5/7	0.969 / 0.994	14/20	12/20
Pooled	14/7/19	0.943 / 0.935	32/40	31/40

Table 20: Cross-architecture target ranking against a coupling-free nearest-signed-role (NN) control, E_{shape} on 400 held-out images; wins mean lower error. On the uncapped cases correspondence wins 13 of 34. NN beats random in 31 of 39 defined cases, the same count as correspondence.

Competition for targets. Single-source placement never tests capacity. A preregistered joint assay removes 16 packets of 32 source channels per direction, grouped by overlapping nearest-role targets (fit side only), against each arm's jointly removed targets (Table 21). Capacity-free arms over-claim: 42% (role only) and 36% (role and amount) of claimed mass exceeds capacity, and 27% of claimed targets serve several sources. Respecting capacity removes the penalty, with no LP gain over greedy;

in 32 packets formed without regard to competition the role-and-amount penalty shrinks to $+0.013$ $[-0.005, 0.033]$. Absolute shape error stays high (coupling 0.896).

Arm	Role	Amount	Capacity	DeiT-3→Swin	Pooled (32 packets)
Random (mass-matched)	–	–	–	$+0.134$	$+0.103$ $[0.047, 0.172]$
CKA	–	–	–	$+0.071$ $[0.003, 0.157]$	$+0.066$ $[0.017, 0.127]$
Role only	✓	–	–	$+0.035$ $[0.005, 0.072]$	$+0.028$ $[0.009, 0.051]$
Role and amount	✓	✓	–	$+0.037$ $[0.006, 0.074]$	$+0.028$ $[0.009, 0.050]$
Greedy	✓	✓	✓	-0.005 $[-0.017, 0.009]$	-0.006 $[-0.013, 0.003]$

Table 21: Joint removal of competing 32-source packets across architectures: arm minus coupling effect-shape error, with 95% packet-bootstrap intervals. Positive values favor the coupling.

Figure 20 gives per-case contrasts; Figure 5 shows outcome-selected cases and Figure 21 an outcome-blind one.

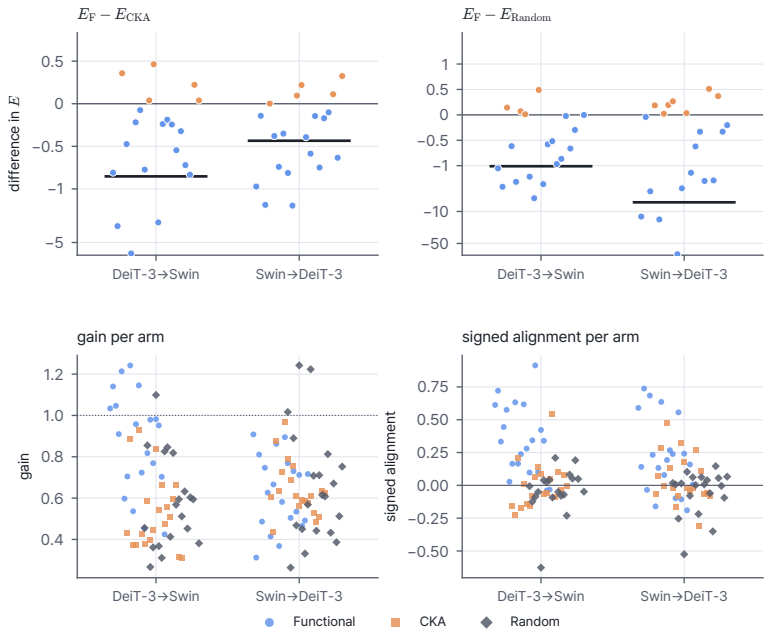


Figure 20: **Cross-architecture results vary across source cases.** Per-case held-out-error contrasts, $E_F - E_{\mathrm{CKA}}$ and $E_F - E_R$ (F: correspondence, "Functional" in the legend), with each arm's gain and signed alignment.

To summarize, the transfer rests on signed roles: nearest-role targets match correspondence for single sources, capacity matters only when sources compete, and outcome-blind examples match outcome-selected ones; absolute reconstruction stays weak throughout.

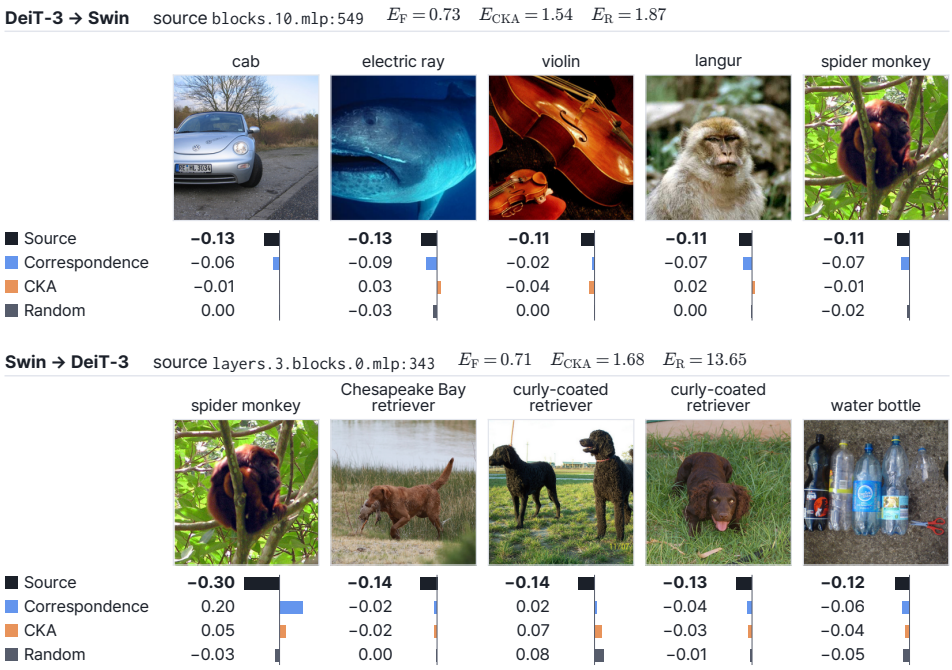


Figure 21: **Outcome-blind per-image intervention examples.** The first preselected case per direction under the frozen largest-allocated-outflow ordering, on the five images the source affects most; all arms share a mass budget.