CAUSAL CORE–PERIPHERY ORGANIZATION AND REMODELING IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Small subsets of neurons can contribute disproportionately to the behavior of large language models, raising questions about how task-related causal support is organized and remodeled during post-training. We introduce Causal Contribution Profiling, a framework that ranks neurons in the multilayer perceptron (MLP) of Transformer blocks by attribution and measures contribution density by patching activations in progressively larger sets of top-ranked neurons. We study 31 models (0.27B–72B parameters) from five families using a 36-task suite and conduct analyses along three training trajectories. Within the scanned range, we find task-selective core–periphery organization, characterized by a small core with high contribution density and broader support with lower density along the neuron ranking. Core membership is more stable than peripheral membership in 92.0% of 224 eligible task–transition comparisons. Retrospective interventions reveal contributions from future core members before models achieve full accuracy on the evaluated sample groups. These contributions tend to strengthen with behavioral gains. After existing members are excluded, source-stage attribution outperforms layer-matched random baselines in predicting new core and support members in 90.5% and 99.1% of eligible cases, respectively. Together, these findings characterize task support as a contribution-defined organization with differential stability and partially predictable remodeling.

1 INTRODUCTION

Task behavior in LLMs can depend strongly on small subsets of internal components. Prior work has localized task-relevant neurons and circuits, linking individual neurons, attention heads, and small subnetworks to linguistic patterns and computations (Geva et al., 2021; 2022; Wang et al., 2023; Han et al., 2026; Ji et al., 2026). However, localization alone does not explain how task-related causal support is distributed across neurons. Many neuron-level analyses focus on highly ranked components or fixed-size subsets, leaving open how contributions are distributed beyond the most prominent units and how this structure changes during post-training.

Task support may vary continuously across neurons rather than form sharply separated modules. A small set of high-contribution neurons may coexist with broader, lower-density support, and the membership of these regions may change across training stages. Previous studies have examined mechanism formation before behavioral gains, circuit persistence and reconfiguration, and changes induced by fine-tuning (Nanda et al., 2023; Tigges et al., 2024; Prakash et al., 2024; Wang et al., 2025). It remains unclear whether higher-contribution regions are more stable, whether neurons that become important at later stages already contribute before full behavioral success, and whether future membership can be predicted from information available at an earlier checkpoint.

We introduce Causal Contribution Profiling to characterize the structure and remodeling of task-related causal support in MLP sublayers (Figure 1). For each task, we rank neurons using attribution patching (Syed et al., 2024) and jointly patch progressively larger top-ranked sets (Zhang & Nanda, 2024). After subtracting size- and layer-matched random-control effects, we estimate contribution density from cumulative effects over successive rank intervals. These profiles define coverage-based reference sets for the core and periphery without assuming a universal sharp boundary, allowing us to track membership stability and change across checkpoints.

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.

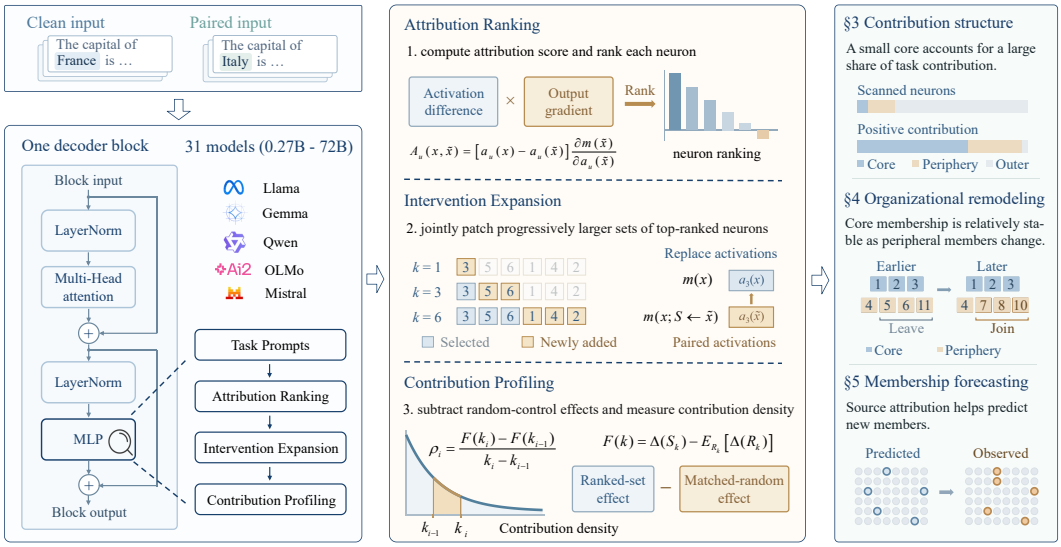


Figure 1: Overview of Causal Contribution Profiling. Paired inputs are used to rank MLP neurons by attribution. Interventions on progressively larger top-ranked sets, compared with matched random controls, yield contribution-density profiles. The framework examines contribution structure (Section 3), organizational remodeling (Section 4), and prediction of new members from earlier-checkpoint attribution (Section 5). Curves and membership diagrams are schematic.

Across models and tasks, we find a task-selective organization comprising a small, high-density core and a broader, lower-density periphery. Within the scanned range, the top-ranked 0.1% of all MLP neurons account for at least half of the positive, random-corrected contribution in 83.2% of model–task pairs. During post-training, core members are more consistently retained than peripheral members, while earlier-checkpoint attribution helps predict new core and support members. Together, these findings extend the analysis from identifying important neurons to understanding how task-related causal support is organized and reshaped during training.

Our contributions are fourfold:

- We introduce **Causal Contribution Profiling**, a framework that quantifies **contribution density** along attribution-ranked Transformer MLP neurons through progressively expanded set interventions and size- and layer-matched random controls.
- Using this framework, we find that task-related causal contributions concentrate in a small subset of neurons, with lower-density contributions extending across a broader population. We characterize this task-selective pattern as **core–periphery organization**.
- By comparing adjacent training checkpoints, we find that high-contribution core neurons **retain their core membership more consistently** than peripheral neurons retain theirs. We further use attribution scores from the earlier checkpoint to **predict new organizational members** at the later checkpoint, outperforming matched random baselines.
- We conduct extensive experiments covering **31 models from five families (0.27B–72B parameters)**, **a 36-task suite spanning four domains**, and **three training trajectories**. These experiments provide converging evidence for the identified organization and its remodeling, and validate source-stage attribution for forecasting new members.

2 Causal Contribution Profiling

We measure task-related causal contributions through activation patching (Vig et al., 2020; Zhang & Nanda, 2024), applied to progressively larger neuron sets under a fixed attribution ranking. Comparisons with size- and layer-matched random controls yield contribution profiles, from which we define reference sets for membership analyses.

2.1 ATTRIBUTION-GUIDED RANKING

We analyze MLP neurons in Transformer blocks, defined as intermediate channels before the output projection. Let $u = (\ell, j)$ denote channel j in layer ℓ, and let $a_u(x)$ be its activation at the final prompt token of input x.

We rank neurons using attribution patching, a first-order approximation to activation patching (Syed et al., 2024; Han et al., 2026). Each example consists of paired inputs x and $\tilde{x}$ that differ in task-relevant information, with respective correct answers y and $\tilde{y}$. Let $m(x) = \log p(y \mid x) - \log p(\tilde{y} \mid x)$ denote the target-minus-alternative margin. This equals the logit difference for single-token answers and uses summed token log probabilities for multi-token answers. With the answer pair held fixed, we compute attribution using the gradient evaluated on $\tilde{x}$:

$$A_u(x, \tilde{x}) = \big[a_u(x) - a_u(\tilde{x})\big] \frac{\partial m(\tilde{x})}{\partial a_u(\tilde{x})}. \tag{1}$$

Within each task, we first average attribution scores over related variants of the same source example, which form a source group, and then equally across groups. Neurons are ranked in descending order of the resulting signed scores. This ranking remains fixed for the set interventions in Section 2.2.

2.2 FROM INTERVENTIONS TO CONTRIBUTION PROFILES

Let S_k be the top-k neuron set in the fixed ranking, and let R_k be a random set drawn from outside the largest scanned set, with the same neuron count in each layer as S_k. For any set S, we jointly replace its activations during the run on x with those from $\tilde{x}$. The mean intervention effect and random-corrected cumulative contribution are

$$\Delta(S) = \mathbb{E}_{(x,\tilde{x})}\left[m(x) - m(x; S \leftarrow \tilde{x})\right], \quad F(k) \quad = \Delta(S_k) - \mathbb{E}_{R_k}[\Delta(R_k)]\,. \tag{2}$$

Here, $m(x; S \leftarrow \tilde{x})$ is the answer margin after replacement. The first expectation averages pairwise effects within each source group and then equally across groups. The second averages over matched random sets.

Let $k_1 < \cdots < k_J$ denote the J neuron-set sizes evaluated along the ranking, with $k_0 = 0$ and $F(0) = 0$. For interval $i = 1, \ldots, J$, the contribution density is

$$\rho_i = \frac{F(k_i) - F(k_{i-1})}{k_i - k_{i-1}}. \tag{3}$$

This measures the incremental random-corrected effect per added neuron. The densities retain their signs and need not decline monotonically. To quantify the concentration of positive contributions, we use the positive part of each density to define the interval mass w_i and cumulative coverage threshold K_η:

$$w_i = (k_i - k_{i-1}) \max(\rho_i, 0), \qquad K_\eta = \min\left\{ k_j : \sum_{i=1}^{j} w_i \geq \eta \sum_{i=1}^{J} w_i \right\}. \tag{4}$$

Here, K_η is the smallest evaluated set size covering at least a fraction $\eta \in (0, 1]$ of positive mass within the scanned range. It is defined only when $\sum_{i=1}^{J} w_i > 0$, without interpolation. For membership analyses, we use $S_{K_{0.5}}$ as the core and $S_{K_{0.9}} \setminus S_{K_{0.5}}$ as the periphery. Together, they form the support set $S_{K_{0.9}}$. These are coverage-based reference sets. Density regimes and their boundaries are tested separately in Section 3.2.

Within each task, let D_1, D_2, and D_3 be disjoint sample sets for attribution ranking, profile estimation, and independent confirmation. Related variants from the same source example are assigned together. Rankings from D_1 remain fixed for profiling on D_2, and membership sets selected on D_2 are evaluated unchanged on D_3. Sampling and statistical procedures are detailed in Appendix A.

3 ANATOMY OF CAUSAL ORGANIZATION

In this section, we examine how causal contributions are distributed across models and tasks. Section 3.1 tests whether contributions concentrate in a small number of neurons. Section 3.2 characterizes their density along the attribution ranking, and Section 3.3 examines whether similar distributional patterns coexist with task-selective causal support.

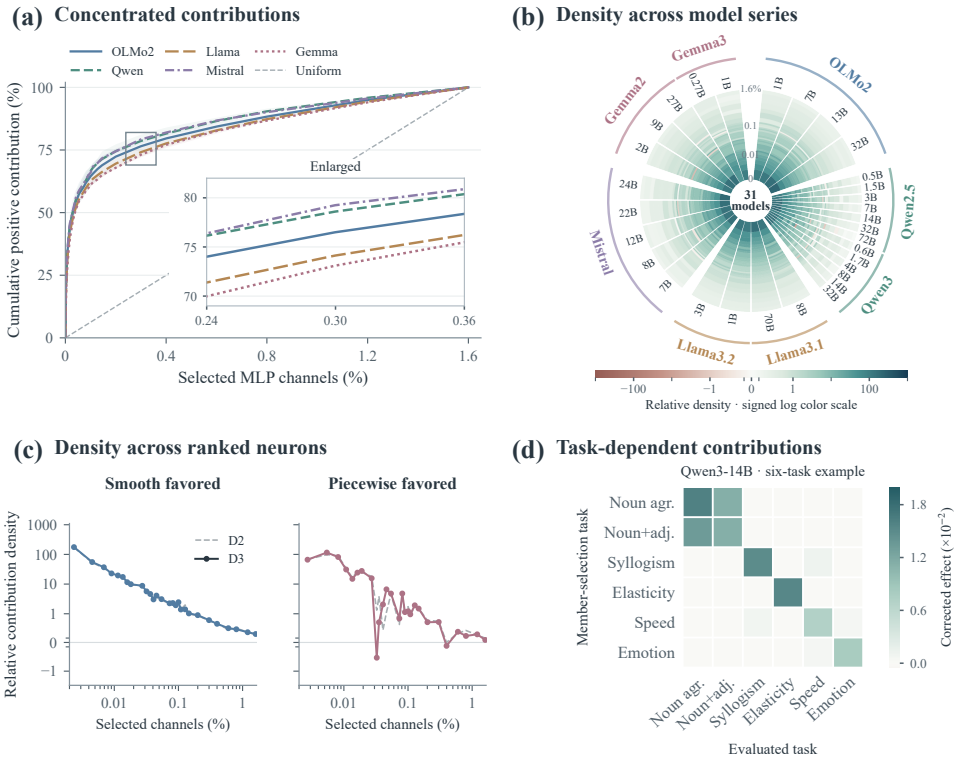


Figure 2: Task-related contribution structure across models and tasks. **(a)** Cumulative positive random-corrected contribution within the scanned range. Lines and shading show family medians and interquartile ranges of model-level task averages. **(b)** Task-averaged relative contribution density across 31 models, with one radial strip per model. **(c)** Measured D_2/D_3 density profiles illustrating cases favoring smooth or piecewise models. **(d)** A six-task cross-intervention example in Qwen3-14B. Rows indicate member-selection tasks and columns evaluation tasks. Colors show random-corrected effects.

3.1 Concentration of Causal Contributions

We analyze 31 post-trained models (0.27B–72B) from five families, covering 894 model–task pairs from the 36-task suite. Following Section 2, we measure coverage relative to the positive, random-corrected contribution within the scanned range of approximately 1.6% of MLP neurons. In Figure 2(a), all five family curves rise steeply near the top of the ranking and then more gradually, departing from the uniform reference. Figure 2(b) complements these family summaries with individual-model profiles, showing generally higher relative contribution density near the top of the ranking.

To assess the consistency of this pattern across models and tasks, we examine the contribution captured by the top-ranked 0.1% of MLP neurons. On independent D_3 data, this set reaches 50% coverage in 744 of 894 model–task pairs (83.2%, Table 1). The proportion ranges from 76.2% to 87.1% across families. **These results show that task-related causal contributions are commonly concentrated in a small fraction of highly ranked neurons across model families.**

3.2 Contribution Density Along the Neuron Ranking

To characterize how contribution density changes along the ranking, we compare smooth and piecewise-constant curves across the same 894 model–task pairs. Both curves are constrained to be non-increasing, with complexity and segment boundaries selected on D_2 and held fixed before evaluation on independent D_3 data.

Table 1: Contribution concentration and density-profile comparisons across model families.

Experimental coverage		Concentration (top 0.1%)	Favored density curve		
Family (models)	Pairs	$\geq 50\%$ coverage	Smooth	Piecewise	Inconclusive.
OLMo (4)	119	94 (79.0%)	92	8	19
Qwen (13)	394	343 (87.1%)	277	20	97
Llama (4)	112	87 (77.7%)	83	6	23
Mistral (5)	168	143 (85.1%)	119	11	38
Gemma (5)	101	77 (76.2%)	58	11	32
Total (31)	**894**	**744 (83.2%)**	**629 (70.4%)**	**56 (6.3%)**	**209 (23.4%)**

Smooth curves are favored in 629 pairs (70.4%), whereas piecewise curves are favored in 56 (6.3%, Table 1). In the remaining 209 pairs (23.4%), the paired 95% confidence intervals for the prediction-error differences include zero, leaving the comparison inconclusive. Figure 2(c) illustrates profiles favoring each description. Together with the concentration results in Section 3.1, these comparisons support a small, high-density core with a broader, lower-density periphery. **This core–periphery organization reflects differences in contribution density without requiring a sharp boundary between the two regions.**

3.3 Task Selectivity of Causal Contributions

Similar contribution profiles need not involve the same members. We examine 30 models, with 13–36 tasks per model meeting prespecified behavioral and coverage requirements. For each model, we intervene on each task's core set (50% coverage), selected on D_2, using each target task's own independent D_3 examples and activations. Effects are corrected against size- and layer-matched random controls and normalized by neuron count and target output scale.

In the six-task Qwen3-14B example (Figure 2(d)), corrected effects are concentrated along the diagonal, whereas most cross-task effects are weaker. The two noun-agreement tasks also exhibit substantial off-diagonal effects, indicating that a core selected for one task can contribute to another. **These patterns support task-selective causal contributions without requiring functionally isolated core sets.** Experimental details and complete cross-task results are provided in Appendix B.

4 Organizational Remodeling Across Training Stages

In this section, we trace causal organization along three post-training trajectories: OLMo2-1B, OLMo2-7B, and Llama–Tulu3-8B. Section 4.1 examines membership stability and turnover. Section 4.2 tests whether future core members contribute before complete behavioral success, and Section 4.3 investigates whether their contributions strengthen as capabilities improve.

4.1 Core Retention and Membership Turnover

We track core and peripheral membership across ten primary training transitions, adjusting overlap against size- and layer-matched random controls. In Figure 3(a), task-level overlaps are averaged across transitions within each trajectory, with most averages lying above the equal-stability line. Mean core overlap exceeds peripheral overlap in OLMo2-1B (0.817 versus 0.706), OLMo2-7B (0.714 versus 0.559), and Tulu3-8B (0.762 versus 0.614). **Core membership is therefore generally more stable than peripheral membership across training checkpoints.**

We also examine membership exchanges among the core, periphery, and region outside the 90%-coverage support set. In addition to the ten primary transitions, Figure 3(b) includes two comparisons involving the Instruct checkpoints of OLMo2-1B and OLMo2-7B. Exchanges between core and periphery or between periphery and the outside region account for 95.0–99.6% of changing members, averaged across tasks within each transition. **Membership turnover is therefore concentrated between neighboring contribution-defined regions.**

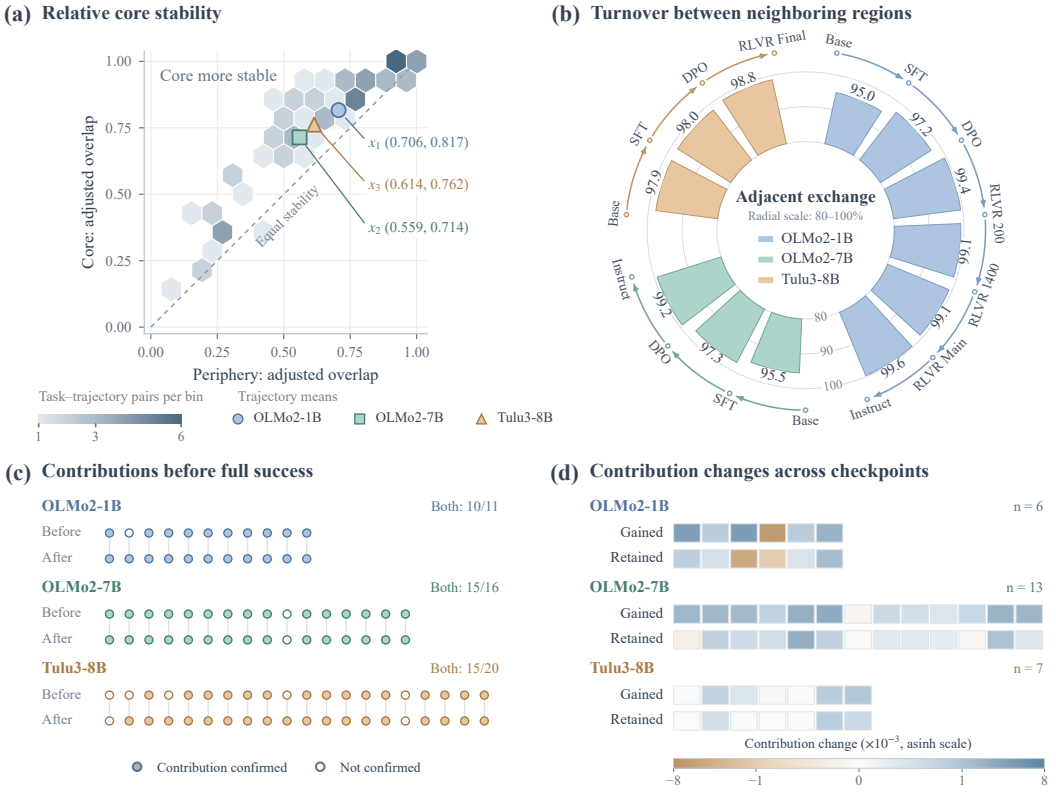


Figure 3: Organizational remodeling across training. **(a)** Core membership is generally more stable than peripheral membership, as measured by overlap across checkpoints. **(b)** Membership turnover mainly occurs between neighboring contribution-defined regions. The radial scale spans 80–100%. **(c)** Later-stage cores often contribute before full behavioral success. **(d)** Normalized contribution changes in matched cohorts that become fully correct (Gained) or remain fully correct (Retained).

Table 2: Contributions before full success.

Trajectory	n	Both	Later only	Neither
OLMo2-1B	11	10 (90.9)	1	0
OLMo2-7B	16	15 (93.8)	0	1
Tulu3-8B	20	15 (75.0)	3	2
Total	**47**	**40 (85.1)**	**4**	**3**

Table 3: Causal contribution changes.

Analysis	n	Positive	Negative	Inconclusive
Gained	47	42	0	5
Retained	140	41	30	69
All cohorts	187	83	30	74
Matched Γ	26	12	0	14

4.2 Causal Contributions Before Full Behavioral Success

To test whether later-stage cores contribute before full behavioral success, we examine sample groups that become fully correct between checkpoints. Across 47 task–transition cases from the ten primary transitions, we identify the later-stage core using D_1 and D_2 and test the same neuron set at both checkpoints on the same independent D_3 groups. We measure mean per-neuron effects relative to size- and layer-matched random sets. At each checkpoint, confirmation requires both a 95% group-bootstrap confidence interval for this corrected effect entirely above zero and a Benjamini–Hochberg-adjusted $q \leq 0.05$ (Benjamini & Hochberg, 1995).

Figure 3(c) shows the paired confirmation outcomes before and after full behavioral success. Contributions are confirmed at both checkpoints in 40 of 47 cases (85.1%, Table 2), with confirmation rates ranging from 75.0% to 93.8% across trajectories. Four cases are confirmed only at the later checkpoint, while three remain unconfirmed at both. **Later-stage cores therefore often contribute before full behavioral success on the evaluated sample groups.**

4.3 Behavioral Gains and Stronger Causal Contributions

We examine the association between contribution strengthening and behavioral improvement by comparing all 47 gained cases from Section 4.2 with 140 retained cases, whose sample groups remain fully correct at both checkpoints. Each cohort uses its own later-stage core, held fixed across checkpoints. Let $\widetilde{V}_c$ denote the random-corrected per-neuron contribution for cohort c, normalized by the checkpoint-specific root-mean-square clean-input answer margin. We define

$$\Delta_c = \widetilde{V}_c^{\,\mathrm{dst}} - \widetilde{V}_c^{\,\mathrm{src}}, \qquad \Gamma = \Delta_{\mathrm{GAINED}} - \Delta_{\mathrm{RETAINED}}, \tag{5}$$

where Δ_c measures within-cohort contribution change and Γ compares gained and retained cohorts matched on task, trajectory, and adjacent-checkpoint transition.

In Table 3, the individual 95% group-bootstrap confidence intervals for Δ_c lie entirely above zero in 42 of 47 gained cases and include zero in the remaining five. Among 140 retained cases, 41 intervals lie above zero and 30 below zero, indicating changes in both directions despite continued behavioral success.

Figure 3(d) displays 26 matched pairs from these cases, with gained cohorts generally showing greater contribution strengthening than their retained counterparts. Individual 95% confidence intervals for Γ lie entirely above zero in 12 comparisons and include zero in 14 (Table 3). **Together, these analyses associate behavioral gains with stronger relative causal contributions from later-stage cores.** Experimental details and supplementary results for Section 4, including the Instruct comparisons, are provided in Appendix C.

5 Forecasting Organizational Membership

In this section, we test whether source-checkpoint task-related signals predict subsequent organizational membership along the same three training trajectories. Section 5.1 examines next-stage core prediction. Section 5.2 tests whether predictive relationships transfer across tasks, and Section 5.3 examines new members after existing members are excluded.

Table 4: Forecasting future membership from source-stage information. New-member targets exclude the corresponding source-stage members. Results cover the ten primary training transitions.

Section	Prediction setting	OLMo2-1B		OLMo2-7B		Tulu3-8B	
		Acc@K (%)	Lift > 1	Acc@K (%)	Lift > 1	Acc@K (%)	Lift > 1
5.1	Next-stage core (all members)	80.8	80/80	67.4	51/51	69.3	93/93
5.2	Transferred prediction rule	35.2	80/80	35.8	51/51	37.4	93/93
5.3	New core members	40.7	50/57	34.6	46/49	29.2	76/84
	New support members	39.0	74/75	26.5	51/51	26.4	91/92

5.1 Forecasting Next-Stage Core Membership

Across the 36-task suite, we evaluate 224 task–transition cases from ten primary training transitions. We predict the next-stage core directly from raw source-checkpoint attribution rankings. Let C_t be the core at the later checkpoint t and $\widehat{C}_s(K)$ the top-K neurons in the ranking at the earlier checkpoint s, with $K = |C_t|$. We measure recovery and enrichment over random prediction as

$$\mathrm{Acc@}K = \frac{|\widehat{C}_s(K) \cap C_t|}{K}, \qquad \mathrm{Lift@}K = \frac{\mathrm{Acc@}K_{\mathrm{obs}}}{\mathrm{median}(\mathrm{Acc@}K_{\mathrm{null}})}. \tag{6}$$

The random baseline permutes future-core labels within layers, preserving core size and layer counts. Lift@K expresses observed recovery as a multiple of the random-baseline median, with values above one indicating higher recovery than this reference. An above-baseline prediction requires recovery above the 97.5th percentile of this distribution and Lift@$K > 1$. Prediction uses the observed future-core size.

Task-balanced recovery reaches 80.8% in OLMo2-1B, 67.4% in OLMo2-7B, and 69.3% in Tulu3-8B (Table 4). In all 224 cases, recovery exceeds the median of the layer-matched random baseline. **Thus, attribution at the earlier checkpoint already contains information about next-stage core membership.**

Because the current core contains the highest-ranked source neurons, selecting the top K also prioritizes existing core members. This connects next-stage prediction with core persistence. Section 5.3 examines whether attribution also predicts new members after existing ones are excluded.

5.2 Cross-Task Transfer of Membership Prediction

We examine cross-task prediction in 1,950 directed comparisons within prespecified task panels, holding the trajectory and checkpoint transition fixed. For each $A \to B$, we fit a logistic regression with fixed regularization using within-layer attribution percentiles at the earlier checkpoint and next-stage core labels from task A. We apply the model unchanged to the corresponding features from task B and select the top-K neurons, using its future-core labels only to determine K and evaluate recovery.

Table 4 averages recovery and lift across training tasks within each of 224 target task–transition cases. Task-balanced recovery reaches 35.2% in OLMo2-1B, 35.8% in OLMo2-7B, and 37.4% in Tulu3-8B, with enrichment over the random baseline in all cases. **The relationship between attribution and future core membership therefore transfers broadly across tasks.**

Each transferred model uses attribution features from the target task rather than neuron identities selected for the training task. A shared prediction rule can therefore identify different members across tasks. Because a single-feature model may preserve the ordering of attribution percentiles, successful transfer does not necessarily imply an improvement over directly ranking the same features. Instead, it shows that a relationship learned on one task remains informative about future core membership in another.

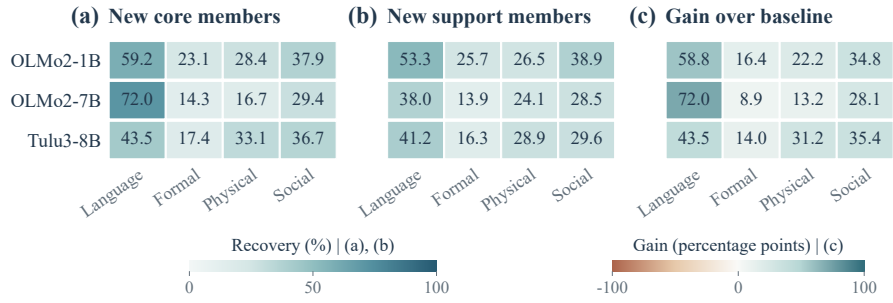


Figure 4: **Forecasting new members beyond persistence. (a)** Recovery of new core members after excluding the source core. **(b)** Recovery of new support members after excluding the source support. **(c)** New-core recovery minus the median recovery of a baseline prioritizing source-periphery members.

5.3 Forecasting New Organizational Members

We revisit the 224 task–transition cases from Section 5.1 to examine whether attribution at an earlier checkpoint predicts new members beyond those already present. Let C_t and P_t be the core and periphery at checkpoint t, with $S_t = C_t \cup P_t$ their support set. For consecutive checkpoints s and t, new core and support members are defined as

$$E_C = C_t \setminus C_s, \qquad E_S = S_t \setminus S_s. \tag{7}$$

We exclude C_s and S_s from the respective candidate sets. Retaining only cases with new members leaves 190 new-core and 218 new-support cases. Predictions follow the raw attribution ranking at checkpoint s, with $\mathrm{Acc}@K$ and $\mathrm{Lift}@K$ evaluated as in Section 5.1 and K equal to the observed number of new members. Label permutations remain within the candidate sets and preserve new-member counts in each layer.

Recovery exceeds the median of the layer-matched random baseline in 172 of 190 new-core cases (90.5%) and 216 of 218 new-support cases (99.1%, Table 4). Across all three trajectories, domain-average recovery is highest for language and lowest for formal reasoning (Figure 4(a) and (b)). Although recovery varies across domains, the above-baseline results extend to both new core and new support members. **Attribution at the earlier checkpoint therefore predicts future membership beyond the persistence of existing members.**

We further compare new-core prediction with a baseline that prioritizes neurons in the earlier periphery over those outside the earlier support, with random ordering within each region. Figure 4(c) shows positive mean recovery gains over the baseline median across all four domains and three trajectories, ranging from 8.9 to 72.0 percentage points. **Attribution therefore provides information about future core membership beyond current peripheral membership.** Experimental details and supplementary results for Section 5 are provided in Appendix D.

6 Related Work

Task-related neurons and contribution structure. Prior work localizes components associated with factual knowledge (Dai et al., 2022; Meng et al., 2022), identifies language-specific units from activation statistics (Tang et al., 2024), and tests task relevance through inhibition and amplification (Song et al., 2024). Other studies examine functional differentiation (Han et al., 2026), sparse neuronal contributions (Ji et al., 2026; Arora et al., 2026), and circuit discovery and evaluation (Conmy et al., 2023; Bhaskar et al., 2024; Hanna et al., 2024; Mueller et al., 2025). We complement these perspectives by profiling intervention-derived contributions along task-conditioned rankings, including lower-ranked support. Unlike connectivity-based core–periphery models (Borgatti & Everett, 2000) or temporal brain-network cores defined by stable community affiliation (Bassett et al., 2013), our regions are defined by positive, random-corrected contribution coverage within the scanned range. Relative stability is therefore an empirical finding, not a defining criterion, and strong contributions need not imply task exclusivity.

Remodeling and membership prediction. Studies of learning dynamics trace the training signals underlying feature formation (Rofin et al., 2026) and characterize changes in model predictions during fine-tuning (Ren & Sutherland, 2025). Related work examines the strengthening of existing computations (Prakash et al., 2024), circuit reconfiguration (Wang et al., 2025), and recovery after corruption (Chhabra et al., 2025). Circuit-guided approaches select task-relevant subgraphs for fine-tuning (Li et al., 2025), while prospective localization uses probing fine-tuning and extrapolation to anticipate post-tuning importance for targeted optimization (Chen et al., 2026). We instead track retention and exchange of contribution-defined members along observed training trajectories and test whether earlier-checkpoint signals predict new members after excluding existing ones.

7 Discussion and Conclusion

Using Causal Contribution Profiling, we characterize task support as a task-selective, contribution-defined core–periphery organization with concentrated contributions and generally declining density along neuron rankings, without requiring a universal sharp boundary. Across checkpoints, core membership is generally more stable than peripheral membership, and turnover occurs mainly between neighboring regions. Contributions from future core members can precede full behavioral success and strengthen with behavioral gains. Earlier-checkpoint attribution predicts new members after existing ones are excluded, showing that remodeling is partially predictable.

Our interventions characterize causal contributions within the examined MLP channels and scan range, while how training produces and reshapes these contributions remains a question for further study. Forecasting assesses membership given the observed size of the future target set. By linking contribution profiles with membership changes, the framework offers a basis for studying how task support is maintained and adapted, and for distinguishing changes in the contributions of existing members from the recruitment of new ones. Future work could test how support reuse and changing neuronal roles relate to retention, forgetting, and cross-task transfer.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and to obtain feedback on methodological descriptions and interpretations of the experimental results. Additionally, we used these tools for manuscript drafting and revision, literature search and reference formatting, and the preparation of figures and tables. The authors made the final decisions on the manuscript and take responsibility for the final content of this work, including text, claims, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We describe Causal Contribution Profiling in Section 2, with model and task coverage, paired-input construction, data splits, and intervention settings detailed in Appendix A. Appendices B and C provide additional analysis procedures, statistical tests, and results for contribution structure and organizational remodeling, respectively. Appendix D specifies the membership prediction methods, cross-task training procedure, evaluation metrics, and random baselines. We will release the code on GitHub upon acceptance.

REFERENCES

Aryaman Arora, Zhengxuan Wu, Jacob Steinhardt, and Sarah Schwettmann. Language model circuits are sparse in the neuron basis. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.22594.

Danielle S. Bassett, Nicholas F. Wymbs, M. Puck Rombach, Mason A. Porter, Peter J. Mucha, and Scott T. Grafton. Task-based core-periphery organization of human brain dynamics. *PLOS Computational Biology*, 9(9):e1003171, 2013. doi: 10.1371/journal.pcbi.1003171. URL https://doi.org/10.1371/journal.pcbi.1003171.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/20fdaf67581e6d7157376d1ed584040a-Abstract-Conference.html.

Stephen P. Borgatti and Martin G. Everett. Models of core/periphery structures. *Social Networks*, 21(4):375–395, 2000. doi: 10.1016/S0378-8733(99)00019-2. URL https://doi.org/10.1016/S0378-8733(99)00019-2.

Hang Chen, Jiaying Zhu, and Wenya Wang. Beyond static interpretability: Anticipating post-SFT mechanisms from pre-SFT parameters for better tuning. *arXiv preprint arXiv:2608.24482*, 2026. URL https://arxiv.org/abs/2608.24482.

Vishnu Kabir Chhabra, Ding Zhu, and Mohammad Mahdi Khalili. Neuroplasticity and corruption in model mechanisms: A case study of indirect object identification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3099–3122. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.170. URL https://aclanthology.org/2025.findings-naacl.170/.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502. Association

for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Pengrui Han, Jacob Andreas, Evelina Fedorenko, and Andrea Gregor de Varda. Modular cognitive architecture emerges in large language models, 2026. URL https://arxiv.org/abs/2608.13567.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*, 2024. URL https://belinkov.com/assets/pdf/colm2024.pdf.

Xiangtian Ji, Yuxin Chen, Zhengzhou Cai, Xiang Wang, An Zhang, and Tat-Seng Chua. Tiny brains, giant impact: Uncovering the keystone neurons of LLM with just a few prompts, 2026. URL https://arxiv.org/abs/2605.24846.

Yueyan Li, Wenhao Gao, Caixia Yuan, and Xiaojie Wang. Fine-tuning is subgraph search: A new lens on learning dynamics. *arXiv preprint arXiv:2502.06106*, 2025. URL https://arxiv.org/abs/2502.06106.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108. PMLR, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2301.05217.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8082273791021571c410f41d565d0b45-Abstract-Conference.html.

Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/afe1aa79e5eea7955f553c61a307273e-Abstract-Conference.html.

Mark Rofin, Jalal Naghiyev, and Michael Hahn. Understanding the emergence of seemingly useless features in next-token predictors. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/60845c4cfc33932c6b2edcc1171f7657-Abstract-Conference.html.

Ran Song, Shizhu He, Shuting Jiang, Yantuan Xian, Shengxiang Gao, Kang Liu, and Zhengtao Yu. Does large language model contain task-specific neurons? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7101–7113. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.403. URL https://aclanthology.org/2024.emnlp-main.403/.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309/.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. LLM circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/47c7edadfee365b394b2a3bd416048da-Abstract-Conference.html.

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 *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63088–63112. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25ak.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

APPENDIX

1. **Appendix A: Experimental Settings and Contribution Measurement.** This section provides additional details on the models, task data, input construction, data splits, and contribution measurements used in Section 2.
2. **Appendix B: Contribution Structure.** This section provides additional methods and results for Section 3, covering contribution concentration, changes in contribution density along neuron rankings, and task selectivity across models and tasks.
3. **Appendix C: Organizational Remodeling.** This section provides additional analyses for Section 4 on membership stability and exchange, contributions before full behavioral success, and changes in causal support across training checkpoints.
4. **Appendix D: Membership Forecasting.** This section describes the prediction settings used in Section 5 and presents task-level results, cross-task transfer analyses, and baseline comparisons for predicting new members.

A EXPERIMENTAL SETTINGS AND CONTRIBUTION MEASUREMENT

This section provides additional details on the experimental settings and contribution measurements used in Section 2. We describe the models, task data, paired inputs, and data splits, followed by attribution ranking and set interventions. We also outline an adaptation to neurons within individual experts in OLMoE, a mixture-of-experts (MoE) model, with the corresponding experiments presented in Appendix B.

A.1 MODELS, TASKS, AND DATA

Table 5 summarizes the model families and parameter sizes used in our experiments. Cross-model analyses cover five model families, while longitudinal analyses follow three training trajectories. We additionally examine contribution structure within individual experts in OLMoE as a separate extension. Appendix C provides additional details on the longitudinal analyses.

Table 5: Model families and parameter sizes used in the experiments.

Model family	Model series and nominal sizes
Cross-model analyses: 31 models	
OLMo2	OLMo2: 1B, 7B, 13B, 32B
Qwen	Qwen2.5: 0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B
	Qwen3: 0.6B, 1.7B, 4B, 8B, 14B, 32B
Llama	Llama-3.1: 8B, 70B
	Llama-3.2: 1B, 3B
Mistral	Mistral: 7B, Ministral: 8B, Mistral NeMo: 12B, Mistral Small: 22B, 24B
Gemma	Gemma2: 2B, 9B, 27B
	Gemma3: 0.27B, 1B
Longitudinal analyses: three trajectories	
OLMo2	OLMo2-1B, OLMo2-7B
Llama–Tulu3	Tulu3-8B
MoE extension	
OLMoE	OLMoE: 1B active / 7B total

We use the 36-task suite drawn from Han et al. (2026), covering language, formal reasoning, physical reasoning, and social reasoning. Table 6 lists the tasks in each domain. Paired inputs are constructed using the task-specific formats, with changes to task-relevant information and corresponding answer alternatives. Related variants derived from the same source example are treated as a source group.

Table 7 gives one paired example from each task family, with the correct answers y and $\tilde{y}$ for the original input x and alternative input $\tilde{x}$, respectively. For the physical and social reasoning examples, the shared prefix is prepended to both continuations shown in the table.

Table 6: Tasks included in the four-domain evaluation suite.

Domain	Tasks	Number
Language	Anaphor gender agreement, determiner–noun agreement (regular, irregular, with adjective), subject–verb agreement, negative polarity items, hypernymy, Wug	8
Formal reasoning	Syllogistic reasoning (verbal, symbolic), propositional reasoning, two-operation addition/subtraction (verbal, symbolic), three-operation addition/subtraction (verbal, symbolic), two-operation multiplication/division (verbal, symbolic), three-operation multiplication/division (symbolic), number sorting, code conditionals, code lists, Code A, Code B	15
Physical reasoning	Newtonian reasoning, PROST, buoyancy, brightness, elasticity, solubility, speed, stability, temperature	9
Social reasoning	Agent, emotion (few-shot), norm appropriateness, morality	4
Total		36

Table 7: Examples of paired inputs from the four task families. Answers are shown separately from the input strings.

Input	Raw prompt text	Answer
Language —— det_noun_agreement_irregular		
Original	Some customers know **these**	men
Alternative	Some customers know **this**	man
Formal reasoning —— add_sub_2op_symbolic		
Original	`862 - 158 =`	704
Alternative	`862 - 128 =`	734
Physical reasoning —— physics_speed		
Shared prefix	What is the physical outcome in the following contexts? Context: A sailboat has a strong tailwind on calm open water. Answer: fast Context: A disk is sliding across a table covered in sticky tree sap. Answer: slow	
Original	Context: A block is moving on a surface covered in a layer of tar. Answer:	slow
Alternative	Context: A rock is tumbling down a steep bare cliff with no trees. Answer:	fast
Social reasoning —— emotion_fewshot		
Shared prefix	What are the inferred emotions in the following contexts? Context: Omar was rejected from every single school. Answer: sadness Context: Omar adopted a cute puppy from a shelter. Answer: joy	
Original	Context: Mei hosted a party but nobody came over. Answer:	sadness
Alternative	Context: Mei saw the northern lights for a night. Answer:	joy

Within each task, source groups are assigned to the three disjoint splits defined in Section 2.2, keeping all variants of a source example in the same split. Samples for each analysis are selected within these partitions. Task coverage and sample counts are reported with the corresponding experiments in Appendices B–D.

A.2 Contribution Measurement

Activation replacement. With model parameters fixed in evaluation mode, we patch MLP activations at the final prompt token, after native activation and gating and before the output projection, using token-length-aligned input pairs. Selected coordinates are replaced jointly across layers and identically in both answer-scoring passes. Answers are scored by summed token log probabilities under teacher forcing, without length normalization, using the fixed answer pair and attribution gradient defined in Section 2.1.

Scan grid. We use the D_1 ranking from Section 2.1, breaking ties by a fixed layer–channel index. Dense-model scans use 28 fractions $q_i = b_i/352256$, where

$$\mathbf{b} = \big(8, 16, 24, 32, 40, 48, 56, 64, 96, 112, 128, 144, 160, 192,$$
$$256, 288, 320, 352, 384, 448, 512, 704, 1056, 1409, 2114, 2818, 4227, 5636\big).$$

For N MLP channels, we round $q_i N$ to the nearest integer with a minimum of one, then remove duplicate counts. This leaves 27 scan points for Gemma3-0.27B. The requested fractions span approximately 0.0023%–1.6%; scans require enough positively attributed neurons to reach the endpoint.

Random controls and uncertainty. Random expansion sequences sample without replacement from outside S_{kJ}, matching the newly added neuron count in each layer at every interval. We use three independent sequences on D_2 and five fresh sequences on D_3, with target and control sets fixed across source groups and evaluated on the same pairs. The D_1 ranking and scan grid are reused on D_3. Coverage is computed from the aggregated density profile, as in Section 2.2. For dense-model profiles, 95% intervals are the 2.5th and 97.5th percentiles of 1,000 source-group bootstrap resamples, using the same resampled groups across all scan points and controls. Rankings and neuron sets remain fixed during resampling.

OLMoE experts. Experts are selected by sample coverage within early, middle, and late layer ranges before attribution. We retain correctly answered pairs for which both inputs' final tokens route to the selected expert in both answer-scoring passes. Patching occurs after SwiGLU and before the output projection, holding routing choices, routing weights, and other experts fixed at the intervened layer. Attribution retains native routing weights through the gradient, without additional frequency weighting. Within each 1,024-channel expert, we scan $k \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$. Controls come from the same expert outside its largest target set, using the three/five-sequence scheme above. We use 10,000 source-group bootstrap resamples with neuron sets fixed. Sample requirements, contribution contrasts, and results are reported in Appendix B.1.

B Contribution Structure

This section provides additional results and analyses of the contribution structure examined in Section 3. We first examine contribution concentration across models and tasks, including within individual OLMoE experts. We then compare descriptions of how contribution density changes along neuron rankings and assess the task selectivity of the selected neuron sets.

B.1 Contribution Concentration Across Models and Tasks

Using the paired inputs and data splits described in Appendix A, we measure contribution profiles separately for each model and task. Each model–task pair therefore represents one model evaluated on one task. The available tasks differ across models, so the analysis does not assume that every model contributes measurements for all 36 tasks. In total, the 31 models provide 894 contribution profiles evaluated on the independent split D_3. Table 8 reports the results for each of the 31 models, grouped by model family.

We examine the contribution of the 0.1% of MLP neurons ranked highest by attribution in each model–task pair. In 744 of the 894 pairs (83.2%), these neurons account for at least half of the positive, random-corrected contribution measured within the scanned range. This pattern appears in a majority of pairs in each of the five model families (Table 8), showing that concentrated task support is a recurring pattern across model families rather than a result driven by a single family.

We further examine whether contribution concentration can arise within an individual expert in OLMoE-1B-7B. Because an expert processes only a subset of inputs, we retain correctly answered input pairs that both activate the expert being studied. We require at least 32, 20, and 20 independent sample groups in D_1, D_2, and D_3, respectively. Of the 36 tasks, 27 provide sufficient samples under these requirements. For each task, we select one expert from each of the early, middle, and late layer ranges based on sample coverage, before measuring neuron contributions. This yields 81 expert–task pairs. Within each selected expert, we progressively expand the intervention set to the top-ranked 256 of its 1,024 neurons, covering one quarter of the expert.

Table 8: Contribution concentration for all 31 models, grouped by model family.

Family	Model	Model–task pairs	$\geq 50\%$ coverage (top 0.1%)
OLMo	OLMo2-1B	21	12
	OLMo2-7B	30	21
	OLMo2-13B	32	28
	OLMo2-32B	36	33
Qwen	Qwen2.5-0.5B	22	15
	Qwen2.5-1.5B	27	20
	Qwen2.5-3B	32	28
	Qwen2.5-7B	34	32
	Qwen2.5-14B	36	32
	Qwen2.5-32B	36	34
	Qwen2.5-72B	35	33
	Qwen3-0.6B	16	12
	Qwen3-1.7B	24	23
	Qwen3-4B	34	29
	Qwen3-8B	31	27
	Qwen3-14B	33	28
	Qwen3-32B	34	30
Llama	Llama-3.2-1B	20	10
	Llama-3.2-3B	25	19
	Llama-3.1-8B	33	30
	Llama-3.1-70B	34	28
Mistral	Mistral-7B	33	29
	Ministral-8B	34	24
	Mistral-NeMo-12B	34	29
	Mistral-Small-22B	32	29
	Mistral-Small-24B	35	32
Gemma	Gemma2-2B	17	11
	Gemma2-9B	26	21
	Gemma2-27B	31	28
	Gemma3-0.27B	13	7
	Gemma3-1B	14	10
Total	**31 models**	**894**	**744**

Within each selected expert, we compare the mean random-corrected contribution per neuron between ranks 1–16 and ranks 129–256, with both intervals specified before evaluating the contribution profiles. On the independent evaluation split D_3, the higher-ranked neurons show positive contributions that exceed those of the lower-ranked neurons in 15 of the 81 expert–task pairs, with both findings supported after Holm correction across all 81 comparisons. These 15 pairs span 15 distinct tasks and all involve the selected late-layer experts. Thus, task support can be concentrated within individual experts, among neurons that share the same expert-level routing decision.

B.2 Density Profiles Along the Neuron Ranking

This section provides additional details and a model-level breakdown of the comparison reported in Section 3.2. The analysis uses signed, random-corrected contribution densities, including negative values. Both descriptions use the logarithm of the selected neuron fraction as the horizontal coordinate. The piecewise-constant description allows one to six segments, each spanning at least two consecutive intervals of the original scan. The smooth description uses a cubic B-spline with five basis functions. Both descriptions are constrained to be non-increasing along the ranking.

We choose the complexity of each description on D_2 by leaving out one sample group at a time, fitting on the remaining groups, and evaluating the omitted group. For the smooth description, we compare six smoothing strengths from 10^{-4} to 10 in tenfold steps. Within each description, we select the simplest candidate whose mean prediction error is within one standard error of the minimum. This favors fewer segments for the piecewise curve and stronger smoothing for the spline

when their predictive performance is similar. We then fit the selected descriptions using all D_2 groups and keep them unchanged for evaluation on D_3.

For each model–task pair, we evaluate both curves on the same D_3 sample groups and calculate their mean squared prediction errors across the measured rank intervals. We subtract the piecewise error from the smooth error and estimate a 95% confidence interval for this difference using 1,000 sample-group bootstrap resamples. An interval entirely below zero favors the smooth description, whereas an interval entirely above zero favors the piecewise description. Intervals containing zero are recorded as showing no clear difference. Table 9 reports these outcomes for each model, grouped by family, providing the breakdown underlying the pooled comparison in Section 3.2.

Table 9: Comparison of smooth and piecewise descriptions of contribution density for all 31 models, grouped by model family.

Family	Model	Pairs	Smooth favored	Piecewise favored	No clear difference
OLMo	OLMo2-1B	21	14	3	4
	OLMo2-7B	30	22	2	6
	OLMo2-13B	32	27	0	5
	OLMo2-32B	36	29	3	4
Qwen	Qwen2.5-0.5B	22	10	1	11
	Qwen2.5-1.5B	27	16	1	10
	Qwen2.5-3B	32	22	4	6
	Qwen2.5-7B	34	28	1	5
	Qwen2.5-14B	36	27	3	6
	Qwen2.5-32B	36	31	0	5
	Qwen2.5-72B	35	25	3	7
	Qwen3-0.6B	16	8	2	6
	Qwen3-1.7B	24	12	3	9
	Qwen3-4B	34	25	0	9
	Qwen3-8B	31	22	1	8
	Qwen3-14B	33	26	1	6
	Qwen3-32B	34	25	0	9
Llama	Llama-3.2-1B	20	11	2	7
	Llama-3.2-3B	25	16	2	7
	Llama-3.1-8B	33	24	2	7
	Llama-3.1-70B	34	32	0	2
Mistral	Mistral-7B	33	23	1	9
	Ministral-8B	34	23	1	10
	Mistral-NeMo-12B	34	24	3	7
	Mistral-Small-22B	32	23	3	6
	Mistral-Small-24B	35	26	3	6
Gemma	Gemma2-2B	17	10	1	6
	Gemma2-9B	26	19	3	4
	Gemma2-27B	31	23	0	8
	Gemma3-0.27B	13	2	4	7
	Gemma3-1B	14	4	3	7
Total	**31 models**	**894**	**629**	**56**	**209**

B.3 Task Selectivity of Causal Contributions

Figure 5 shows the complete 36×36 matrix for Qwen2.5-14B, selected as an illustrative example from the models evaluated on all 36 tasks. The stronger effects along much of the diagonal indicate task selectivity, while off-diagonal effects among related tasks, such as the three noun-agreement variants, show that selected neuron sets can also contribute to other tasks.

For each task-selected neuron set in Section 3.3, we reuse the five random controls constructed for its D_3 evaluation following Appendix A.2. Specifically, we take the matching-size set from each of the five random expansion sequences. Random neurons are drawn from outside that selection task's largest scanned neuron set, with the number drawn from each layer matching the number of selected

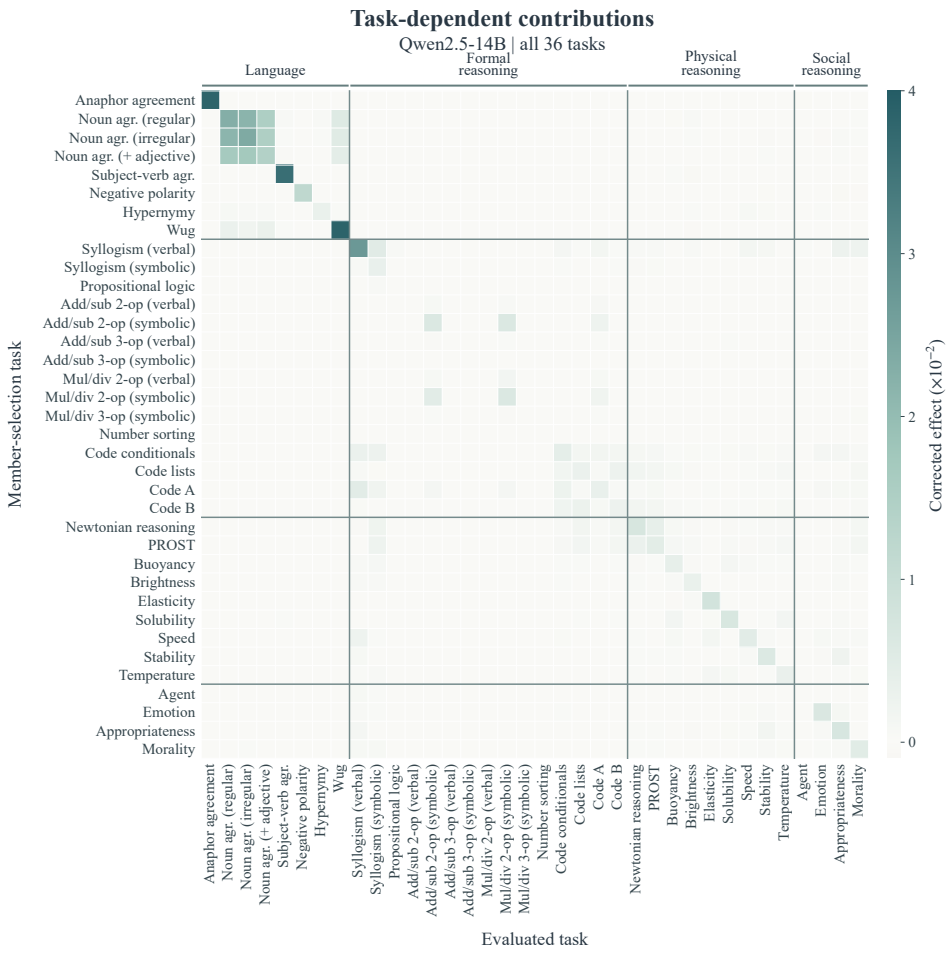


Figure 5: Complete cross-task contribution matrix for Qwen2.5-14B. Rows identify the task used to select the 50%-coverage neuron set on D_2; columns identify the task on which that fixed set is evaluated using D_3. Colors show the random-corrected, per-neuron effect divided by the evaluation task's root mean square answer margin, as described below. All 1,296 cells are measured; near-white cells indicate near-zero effects. Separators mark the four task families, ordered identically on both axes.

neurons in that layer. Each control therefore matches both the total size and the layer composition of the original set. These controls are determined before observing the D_3 intervention results and remain unchanged when the original set is evaluated on different tasks.

For each evaluation task, we first average intervention effects within each sample group and weight the groups equally. Effects for the task-selected set and its controls are divided by their common neuron count, after which the mean effect of the five controls is subtracted from that of the selected set. The corrected effect is then divided by the root mean square of the evaluation task's unperturbed, group-averaged answer margins. This denominator is shared by all neuron sets evaluated on the same task. To estimate uncertainty, we perform 1,000 bootstrap resamples of the D_3 groups within each evaluation task, using the same resampled groups for all selected sets and their controls. Each resample recomputes the normalization and the difference between the mean diagonal and off-diagonal effects, yielding a model-specific 95% confidence interval.

Table 10 summarizes the same matrix by the relationship between the selection and evaluation tasks. The mean effect is largest for the same task; among different tasks, it is larger within a task family than across families. This pattern illustrates that task selectivity can coexist with contributions shared among related tasks.

Table 10: Mean contribution effects within and across task families in Qwen2.5-14B.

Selection–evaluation relationship	Matrix cells	Mean effect ($\times 10^{-2}$)
Same task	36	0.8598
Different task, same family	350	0.0660
Different task, different family	910	0.0100

C Organizational Remodeling

To supplement Section 4.1, we distinguish task-level comparisons from comparisons restricted to sample groups that remain fully correct at both checkpoints. Task-level profiles use the sample groups for each task without separating them according to changes in correctness. The restricted analysis instead constructs profiles using only groups answered fully correctly at both checkpoints. Across the ten primary training transitions, these analyses respectively comprise 190 comparisons covering 31 tasks and 34 comparisons covering five additional tasks. Each comparison concerns one task and one checkpoint pair within a training trajectory. The two Instruct comparisons provide a further 45 task-level comparisons, reported separately in Table 11.

C.1 Membership Stability and Turnover Across Checkpoints

This section supplements Section 4.1 by distinguishing two analyses of membership stability. Task-level analyses construct contribution profiles from each task's sample groups without separating them according to changes in correctness. The other analysis constructs profiles using only groups answered fully correctly at both checkpoints. These analyses cover 31 tasks and five additional tasks, respectively, across ten primary training transitions. Each comparison concerns one task and one checkpoint pair within a training trajectory. Table 11 reports the two analyses separately and lists results for the two additional Instruct comparisons.

For each region, let A and B denote its neuron sets at the earlier and later checkpoints. We measure their overlap as $|A \cap B| / \min(|A|, |B|)$. To account for overlap expected from set size and layer composition, we independently sample a random counterpart for each set, preserving its neuron count in every layer. We subtract the median overlap across 1,000 such pairs from the observed overlap. This adjusted measure describes shared membership relative to the smaller set, rather than the fraction of earlier members retained. Both analyses show higher adjusted overlap for the core than for the periphery in most comparisons. The Instruct comparisons show the same overall tendency.

We further examine exchanges among the core, periphery, and region outside the 90%-coverage support. Among neurons whose regional membership changes, we calculate the proportion exchanged between core and periphery or between periphery and the outside region. To assess whether this proportion exceeds what region sizes alone would predict, we keep the earlier assignments fixed and randomly reassign the later region labels within each layer, preserving the number of neurons in each region. The baseline is the median proportion from 1,000 such permutations. Comparisons with no regional changes have undefined turnover proportions but remain included in the stability analysis. These exchanges describe membership differences between checkpoints, not an observed sequence of intermediate movements.

Table 11: Summary of membership stability and regional turnover.

Analysis	Total	Core more stable	No turnover	With turnover
Task-level	190	175 (92.1)	2	188
Consistently correct groups	34	31 (91.2)	0	34
Primary total	**224**	**206 (92.0)**	**2**	**222**
Instruct supplement	45	42 (93.3)	0	45

C.2 Earlier Contributions of Later-Stage Core Members

This section supplements Section 4.2 with details of the retrospective tests of later-stage cores. For each task and checkpoint pair, we select sample groups with at least one incorrect response at the earlier checkpoint but all responses correct at the later checkpoint, considering both clean and paired inputs. Each task–transition case requires at least 32 groups in D_1 and 20 groups each in D_2 and D_3. We identify the later-stage core using its D_1 attribution ranking and D_2 contribution profile, then evaluate the same neuron set at both checkpoints on the same independent D_3 groups. This core may include members already present in the earlier core, rather than only newly recruited neurons.

At each checkpoint, we compare the core with five random neuron sets, each sampled to match its size and neuron count in every layer. We subtract their mean intervention effect from the core effect, divide by the core size, and average across sample groups. The same random sets are used at both checkpoints. We obtain 95% confidence intervals by resampling D_3 groups 1,000 times, keeping the group selections paired across checkpoints. Positive contributions are confirmed only when the interval for this corrected per-neuron effect lies entirely above zero and the Benjamini–Hochberg-adjusted q-value is at most 0.05 (Benjamini & Hochberg, 1995). Multiple-testing correction is performed separately for the earlier and later checkpoints within each prespecified task set. The Instruct tests are corrected separately from the primary analyses.

The primary analyses cover 21 tasks. Table 12 separates cases with contributions confirmed at both checkpoints, only at the later checkpoint, or at neither checkpoint. Confirmation at both checkpoints occurs across all three trajectories, showing that later-stage cores can contribute while some responses remain incorrect. Cases without confirmation remain included in the reported totals. For the Instruct supplement, only the emotion and syllogism tasks on OLMo2-1B provide enough newly fully correct groups to meet the same sample requirements. Both cases show confirmed contributions at the earlier and later checkpoints.

Table 12: Confirmation of later-stage core contributions before and after full behavioral success.

Training trajectory	Cases	Both	Later only	Neither
Primary analyses				
OLMo2-1B	11	10 (90.9)	1	0
OLMo2-7B	16	15 (93.8)	0	1
Tulu3-8B	20	15 (75.0)	3	2
Primary total	**47**	**40 (85.1)**	**4**	**3**
Instruct supplement				
OLMo2-1B	2	2 (100.0)	0	0

C.3 Contribution Changes with Behavioral Improvement

To supplement Section 4.3, we retain all 47 gained cases from Appendix C.2, regardless of whether their earlier contributions were confirmed, and examine 140 retained cases whose sample groups remain fully correct at both checkpoints. Each case corresponds to one task, training trajectory, and checkpoint transition for a given cohort, subject to the sample-count requirements described in Appendix C.2. Together, these cases cover 27 tasks. The direct gained–retained comparison uses 26 prespecified pairs across 15 tasks, matched on task, trajectory, and transition. These pairs reuse the preceding cases rather than constituting additional observations.

For each cohort, its later-stage core is selected using D_1 and D_2, and the same neurons are evaluated on the same independent D_3 sample groups at both checkpoints. Random-corrected per-neuron contributions are normalized by the checkpoint-specific root-mean-square clean-input answer margin to account for changes in output scale. We estimate 95% confidence intervals for Δ_c and Γ, defined in Equation (5), using 1,000 group-bootstrap resamples. Within each cohort, sample groups are resampled jointly across checkpoints, with normalization recomputed in each resample. Gained and retained cohorts contain disjoint sample groups and are resampled independently when estimating Γ.

Table 13 classifies cohort changes and matched contrasts as positive, negative, or inconclusive according to whether their individual 95% confidence intervals lie entirely above zero, entirely below zero, or include zero, respectively. Among retained cases, contributions increase in 41 cases and decrease in 30 according to their individual intervals, showing that continued behavioral success can coexist with changes in causal support. The matched comparisons therefore assess strengthening relative to this varying background. The Instruct supplement contains two gained cases and 29 retained cases, with two matched comparisons. Both matched differences have confidence intervals above zero, while retained cases again exhibit changes in both directions.

Table 13: Contribution changes and matched contrasts.

Analysis	n	Positive	Negative	Inconclusive
Primary analyses				
Gained cohorts	47	42	0	5
Retained cohorts	140	41	30	69
Matched contrast Γ	26	12	0	14
Instruct supplement				
Gained cohorts	2	2	0	0
Retained cohorts	29	6	7	16
Matched contrast Γ	2	2	0	0

D MEMBERSHIP FORECASTING

This section provides the prediction methods and supplementary comparisons for Section 5. We distinguish direct prediction from source-stage attribution rankings from cross-task prediction using a fitted attribution-based model, and report the two Instruct comparisons separately from the ten primary training transitions.

D.1 PREDICTION FROM SOURCE-STAGE ATTRIBUTION

Let $\mathcal{U}$ denote all MLP neurons in a model, and let $A_{T,s}(u)$ be the signed, group-averaged attribution score of neuron u for task T at the source checkpoint s, computed on D_1 as described in Section 2.1. To predict the core at checkpoint t, we select

$$\widehat{C}_{T,s}(K) = \mathrm{TopK}(\mathcal{U}; A_{T,s}), \qquad K = |C_{T,t}|, \tag{8}$$

where $\mathrm{TopK}(\mathcal{V}; q)$ selects the K neurons with the highest scores q from candidate set $\mathcal{V}$. Equal attribution scores are ordered by a fixed neuron index. This prediction uses the original attribution ranking without fitting an additional model.

The 36-task suite gives 360 possible task–transition comparisons across the ten primary transitions and 72 across the two Instruct comparisons. Evaluation requires defined core memberships at both checkpoints and available source attribution scores. Table 14 reports the resulting comparison counts.

We evaluate recovery and lift using Equation (6). For each comparison, 1,000 random permutations of future core labels within layers preserve core size and layer counts while leaving the prediction fixed. An above-baseline case requires recovery above the 97.5th percentile of this distribution and $\mathrm{Lift@}\,K > 1$. The supplied target size K makes this a test of membership prediction rather than size prediction.

Table 14: Coverage of next-stage core prediction.

Comparisons	Evaluated / possible	Above baseline
Primary	224 / 360	224
Instruct	45 / 72	45

D.2 Cross-Task Prediction Model

For cross-task prediction, we express attribution as a within-layer percentile. Let $d_{\ell(u)}$ be the number of MLP neurons in the layer containing u, and let $r_{T,s}(u)$ be its ascending attribution rank within that layer, with average ranks assigned to ties. The predictor input and training label are

$$z_{T,s}(u) = \frac{r_{T,s}(u) - 1}{d_{\ell(u)} - 1}, \qquad y_{A,t}(u) = \mathbf{1}\{u \in C_{A,t}\}. \tag{9}$$

Thus, larger $z_{T,s}(u)$ indicates higher attribution relative to other neurons in the same layer. Each training observation is a neuron, with one source-stage feature and a binary label indicating future core membership for task A.

For each task A and checkpoint transition within a trajectory, we fit a single-feature logistic regression with an intercept. Training minimizes class-weighted binary cross-entropy with an L2 penalty on the coefficient. The two classes receive equal total weight, and the inverse regularization strength is fixed at $C_{\text{reg}} = 1$. Let $\widehat{\beta}_A$ and $\widehat{b}_A$ denote the fitted coefficient and intercept. We apply them unchanged to task B:

$$q_{A\to B}(u) = \sigma\left(\widehat{\beta}_A z_{B,s}(u) + \widehat{b}_A\right), \qquad \sigma(v) = \frac{1}{1+\exp(-v)}. \tag{10}$$

The predicted set is $\mathrm{TopK}(\mathcal{U}; q_{A\to B})$, with $K = |C_{B,t}|$. Future labels from B are not used for fitting. They determine the target size and evaluate the resulting prediction. For tied prediction scores at the selection boundary, recovery is averaged over the equally ranked candidates.

We compare distinct tasks within prespecified task panels, holding the trajectory and checkpoint transition fixed. A panel with m available tasks contributes $m(m-1)$ directed comparisons. These yield 1,950 primary and 378 Instruct comparisons. Rule transfer requires AUPRC, computed as average precision, above the 97.5th percentile of 1,000 within-layer label permutations, together with recovery above the target prevalence $K/|\mathcal{U}|$.

To distinguish transfer of the attribution–membership relationship from transfer of neuron identities, we also predict task B's future core directly from task A's raw attribution ranking. This comparison does not use task B's source features and is evaluated by the recovery criteria in Appendix D.1. Table 15 reports both comparisons.

Table 15: Cross-task predictions meeting their above-baseline criteria.

Prediction	Primary	Instruct
Attribution-based rule	1,950 / 1,950	378 / 378
Direct neuron ranking	921 / 1,950	119 / 378

D.3 Prediction of New Members

To evaluate new-member prediction, we retain the previously defined scores and remove existing members from the candidate set. Let $E_C = C_t \setminus C_s$ and $E_S = S_t \setminus S_s$ be the new core and support members defined in Equation (7). For a prediction score q, we select

$$\begin{aligned} \widehat{E}_C &= \mathrm{Top}_{|E_C|}(\mathcal{U} \setminus C_s; q)\,, \\ \widehat{E}_S &= \mathrm{Top}_{|E_S|}(\mathcal{U} \setminus S_s; q)\,. \end{aligned} \tag{11}$$

Within-task prediction uses raw source attribution. Cross-task prediction reuses the model in Equation (10) without refitting, with memberships and exclusions defined for task B.

Among the 224 primary comparisons, 34 contain no new core members and six contain no new support members, leaving 190 and 218 comparisons. The corresponding exclusions among the 45 Instruct comparisons are 13 and two, leaving 32 and 43. Cross-task comparisons are likewise omitted when the target task has no new members. These cases are not counted as prediction failures. For both prediction methods, we use the recovery criteria from Appendix D.1, with 1,000 within-layer label permutations restricted to the remaining candidates. Tables 16 and 17 report the results separately.

Table 16: Prediction of new core members.

Prediction	Primary	Instruct
Within-task attribution	172 / 190	27 / 32
Cross-task attribution rule	1,035 / 1,740	134 / 298

Table 17: Prediction of new support members.

Prediction	Primary	Instruct
Within-task attribution	216 / 218	41 / 43
Cross-task attribution rule	1,710 / 1,909	248 / 359

For new-core prediction, we additionally ask whether attribution provides information beyond source-periphery membership. The periphery-first baseline places all source-periphery neurons ahead of neurons outside the source support, with random ordering within each region. Raw attribution exceeds the 97.5th percentile of recovery from 1,000 such rankings in 159 of 190 primary comparisons and 24 of 32 Instruct comparisons. The gain shown in Figure 4(c) is instead measured relative to the baseline median.