TRACING THE FORMATION OF ALTERNATIVE INTERNAL SOLUTIONS IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

How do transformers arrive at alternative internal solutions to the same state-tracking task? We study cyclic state tracking, where distinct Fourier modes can each encode the complete state. Under standard training on C_5, runs with balanced two-mode representations converge rapidly, whereas single-mode-dominant runs converge slowly. Encouraging balanced mode energy yields balanced endpoints across runs; most converge rapidly, but a minority remain slow. Tracing the modes through the final MLP reveals a difference in how these representations are formed: in slow runs, one mode depends directionally on another, as shown by removal and rotation interventions. This intervention signature, identified in discovery runs, also separates the convergence regimes in held-out runs. Extending the analysis to the three-mode setting of C_7, we find that faster convergence is likewise associated with energy distributed across multiple modes, while differences in their causal contributions are associated with cross-mode dependence. Together, these findings show that coexisting representations need not reflect independent computational pathways, highlighting the importance of tracing how they arise within the network when interpreting a model's internal solution.

1 INTRODUCTION

Transformers can learn different internal solutions to the same task. Fourier analyses of modular addition have made some of these solutions concrete: trigonometric circuits implement arithmetic, and related circular representations can participate in different computations (Nanda et al., 2023; Zhong et al., 2023). When several complete state codes coexist, what does their energy allocation tell us about the learned solution?

We study cyclic state tracking: a causal transformer predicts the cumulative state at every prefix of an action sequence. For prime cycle size, each nonzero real Fourier harmonic identifies every state; C_5 admits two such two-dimensional codes and C_7 admits three. This gives an exactly known family of alternative complete state codes for comparing internal solutions. The task is commutative: the final state depends on the cumulative sum, while supervision requires that sum at each prefix.

Our investigation starts with a convergence pattern. Under ordinary C_5 training, balanced two-mode endpoints accompany earlier convergence than single-mode-dominant endpoints. Does encouraging energy balance therefore produce consistently rapid convergence? Regularization makes all discovery endpoints balanced, yet a minority remains slow. To understand this mismatch, we examine how these similar representations are formed inside the trained network. Functional assays reveal weaker isolated usability of one code in the slow group; final-MLP removal and rotation interventions then reveal directional dependence on another code. Cross-mode dependence thus emerges from investigating the regularization result.

Our contribution is to characterize the causal relations between coexisting complete state codes. First, energy regularization separates endpoint balance from convergence regime. Second, complementary directional interventions identify final-MLP cross-mode synthesis in the slow balanced runs, and its signature aligns with convergence in 70 held-out runs. Third, the three-mode C_7 setting links multiplicity to convergence and cross-mode dependence to unequal causal allocation. Throughout, "formation" means construction during forward computation, principally across the final MLP, rather than reconstruction of the optimization history. The novelty is the causal organization of codes within a solution, not the existence of multiple internal solutions.

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.

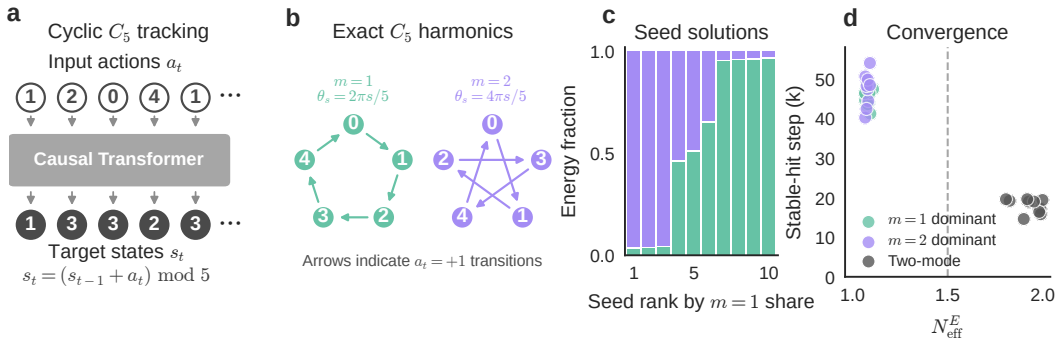


Figure 1: **Alternative complete state codes in C_5.** (a) A causal transformer predicts the cumulative state after each action. (b) The two exact real harmonics encode all five states; arrows show action $+1$. (c) Endpoint energy fractions for baseline seeds 1–10, sorted by mode-1 share. (d) All 30 baseline runs: energy effective modes versus stable-hit step. A two-mode endpoint is associated with earlier convergence, but this observation does not determine how either code is used or formed.

2 Task, Representations, and Experimental Logic

Task and model. Actions a_t are sampled uniformly from $\{0, \ldots, n-1\}$ and the target is $s_t = (s_{t-1} + a_t) \bmod n$, with $s_0 = 0$. As in Li et al. (2025), we supervise the state after every action prefix of a length-100 sequence; here, the actions are cyclic increments in C_5 or C_7. Models are causal transformers (Vaswani et al., 2017) with eight pre-normalization blocks, width 256, eight attention heads, MLP width 1024, and rotary positional embeddings (Su et al., 2021). The endpoint is the saved final training model, not a checkpoint chosen retrospectively for an intervention effect. Stable-hit step is the recorded step of the third consecutive evaluation with last-position accuracy at least 0.97; evaluations are 200 steps apart. The separate learning-trajectory summaries use 1k-step sampling. Appendix A gives the stopping, optimization, and checkpoint protocols.

Complete harmonic codes. For a state s, mode m has ideal coordinates

$$\phi_m(s) = \big(\cos(2\pi ms/n),\, \sin(2\pi ms/n)\big). \tag{1}$$

Because $n = 5$ or 7 is prime, multiplication by any $m \neq 0$ permutes states, so each ϕ_m is injective. We decompose state-conditioned residual centroids into conjugate Fourier pairs. Their energies E_m give shares $p_m = E_m / \sum_j E_j$ and the effective number $N_{\text{eff}}^{E} = (\sum_m p_m^2)^{-1}$. This quantity ranges from one for a single dominant code to $M = (n-1)/2$ for equal energy. Endpoint measurements use the residual after the final block and before the final LayerNorm. Different assays use either positionwise energy fractions or pooled calibration geometry; Appendix B specifies these distinctions.

From presence to use and dependence. We use three complementary questions. Energy measures how much state-conditioned variation lies in each harmonic. Functional assays ask whether a code affects predictions and remains usable when other harmonic components are removed. Cross-mode assays ask whether changing one code before the final MLP changes another code produced downstream. All interventions retain the non-harmonic residual unless explicitly specified, and the output head includes the model's final LayerNorm. We use joint harmonic coordinates rather than assume that separately fitted planes are mutually orthogonal.

Cohorts and interpretation. The discovery comparison contains baseline and energy-regularized C_5 models for seeds 1–30. Regularized seeds 31–100 form the held-out cohort. C_7 uses baseline and regularized seeds 1–30. Within C_5, "fast" denotes stable-hit step below 30k and "slow" denotes at least 30k, consistently across figures. These labels describe convergence; they do not define intervention directions. We do not apply this boundary to C_7. A seed/run is the experimental unit. Mode observations within a run are dependent.

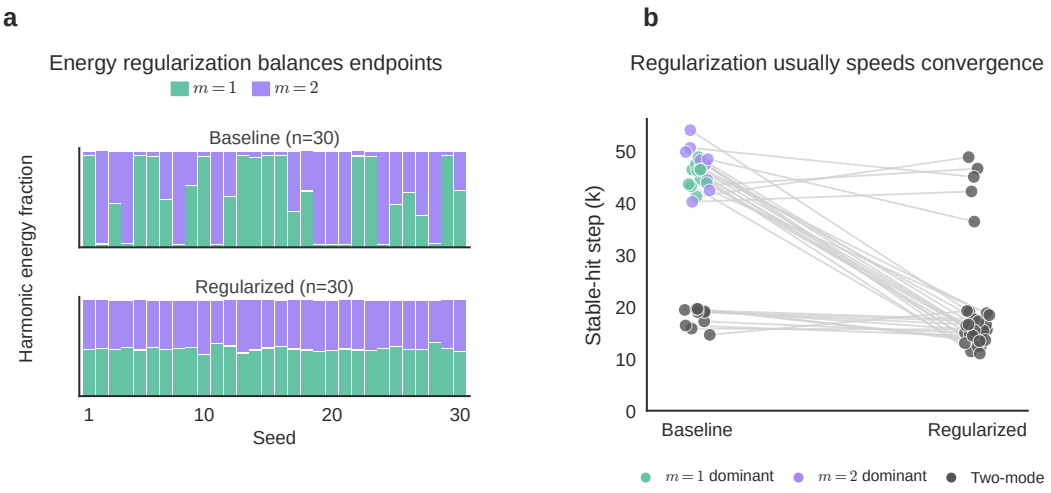


Figure 2: **Energy regularization makes endpoints similar but leaves slow runs.** Thirty seed-matched baseline/regularized pairs. The regularized condition has 30 two-mode endpoints, compared with 10 at baseline ($N_{\text{eff}}^{E} \geq 1.5$). Stable-hit steps decrease in 26 pairs, and the cohort median decreases from 43.7k to 15.5k. Nevertheless, five regularized runs remain above the 30k boundary.

3 Results

3.1 Endpoint multiplicity accompanies earlier baseline convergence

Does the existence of several complete state codes distinguish learned solutions? Figure 1 first examines this observational question. The baseline C_5 cohort contains 10 two-mode endpoints ($N_{\text{eff}}^{E} \geq 1.5$), 11 mode-1-dominant endpoints, and nine mode-2-dominant endpoints. The two-mode runs reach the stable target at 14.6–19.6k steps (median 19.0k), whereas single-mode-dominant runs reach it at 40.2–54.0k. All 30 converge.

Thus, an endpoint statistic separates these baseline convergence outcomes. To investigate this association, we next directly encourage balanced energy during training and ask whether similar endpoints retain different behavior.

3.2 Balanced endpoints do not eliminate the convergence difference

The regularized objective adds $\lambda \sum_m (p_m - 1/M)^2$, averaged over valid positions, to token-level cross-entropy, with $\lambda = 0.5$. Fractions are computed from minibatch state centroids at the final residual location. This objective favors distributing harmonic energy; it imposes no constraint that each mode be independently sufficient or causally independent (Appendix A).

Across the 30 seed-matched pairs, regularization yields two-mode endpoints in all 30 runs (Figure 2). It also accompanies earlier convergence in most pairs: 26 of 30 improve, and median stable-hit step changes from 43.7k to 15.5k. However, the regularized cohort contains 25 fast and five slow runs. The latest fast run reaches the target at 19.2k, whereas the earliest slow run reaches it at 36.4k; the 30k division lies in this empty interval. The slow seeds are 10, 11, 13, 21, and 28.

This dissociation changes the explanatory question. The slow regularized models have both complete harmonic codes at the endpoint, so their difference from fast models cannot be summarized as the absence of a second represented mode. We therefore test the functional and computational organization of the resulting codes.

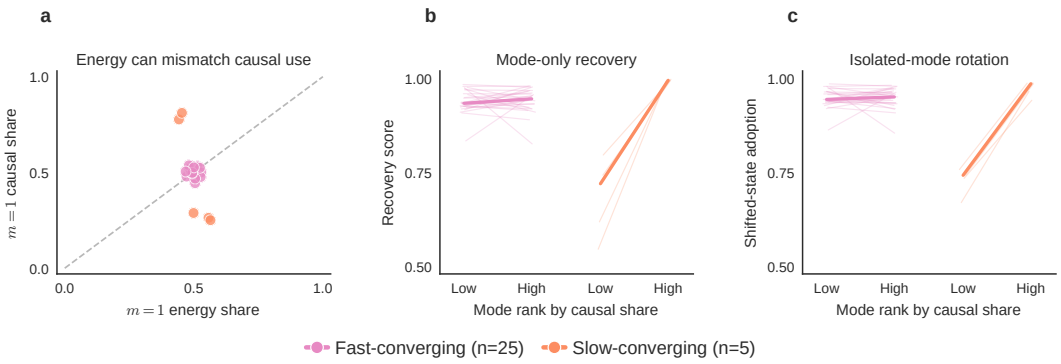


Figure 3: **Similar energy can conceal unequal independent usability.** Regularized discovery runs (25 fast, 5 slow). (a) Energy share versus rotation-derived causal share. (b) Mode-only recovery after removing all harmonic components and restoring one. (c) Target adoption when rotating that isolated mode. Low/high rank is fixed by causal share in (a), before inspecting (b–c). Lines pair modes within a run. Lower-share modes in the slow group retain substantial information and function, but are less independently usable under these assays.

3.3 The slow balanced runs have unequal independently usable codes

Do similarly represented modes contribute similarly to task output? We quantify a mode's *causal share* using the positive, control-adjusted gain in target-state log probability after a semantic rotation, normalized across modes. The control is matched to the mode's perturbation magnitude in residual space. Causal share is an assay-specific allocation of response, not a decomposition of the model's prediction into additive pathways.

Figure 3a shows that near-equal energy need not imply equal causal allocation. To determine whether this is merely a difference in one response metric, we rank modes by causal share and apply two further assays. In mode-only recovery, all harmonic contributions are removed and one is restored. If r_0 and r_m are accuracies on originally correct tokens after removing all modes and restoring m, respectively, recovery is $(r_m - r_0)/(1 - r_0)$. In isolated rotation, only mode m is restored after its coordinates are rotated toward another state; we measure adoption of that state. These assays test usability by the existing output head, without retraining it. Their geometry and responses use the same fixed probe set, so recovery is a within-sample diagnostic, not independent test performance.

For the slow group, median recovery is 0.720 for the lower-share mode and 0.995 for the higher-share mode; corresponding isolated-rotation adoption is 0.742 and 0.986. In the fast group, recovery medians are 0.934 and 0.946, and adoption medians are 0.944 and 0.951. Thus, the slow group's lower-share mode is not absent or wholly unused. Its weakness is relative and intervention-dependent: it is less able to support the existing readout in isolation. These assays reveal a difference in isolated usability, but do not explain how it arises. We therefore trace where each code becomes available across layers.

3.4 Final-MLP interventions reveal directional cross-mode synthesis

We first fit harmonic state decoders at the residual output of each block using calibration sequences independent of evaluation. In fast runs, both causal-share ranks become decodable near the end of the network. In slow runs, the higher-share code becomes decodable one block earlier, while the lower-share code remains near chance until the last block (Figure 4a–b). The lower-share code thus becomes decodable only in the final block, but decoding does not identify the computation that produces it. We next intervene at the final MLP input and measure how each mode responds.

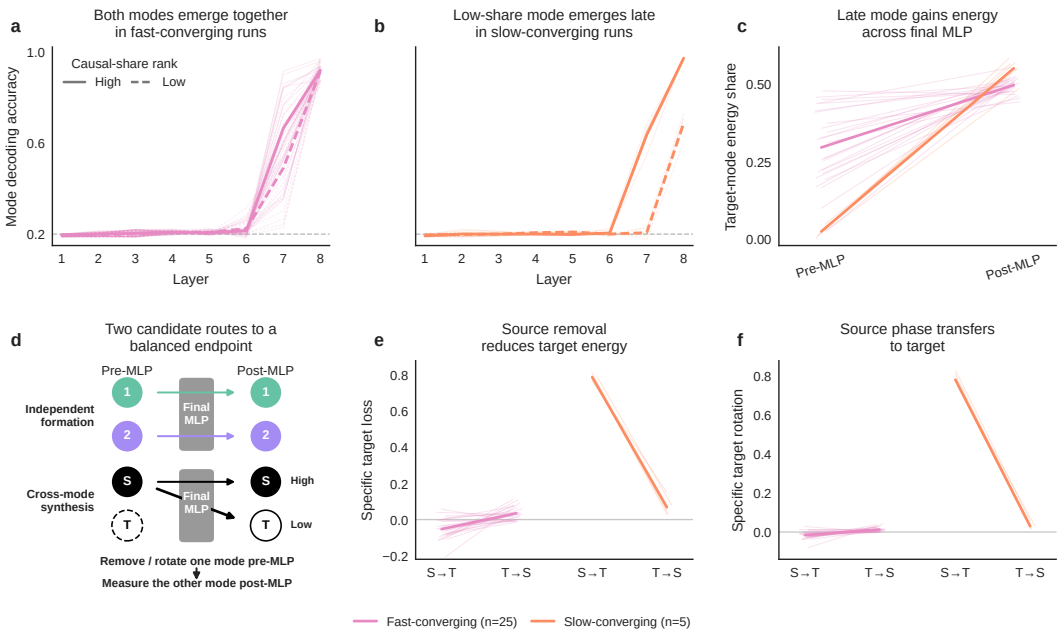


Figure 4: **Directional dependence across the final MLP in the discovery cohort.** All 30 regularized runs are included. (a–b) State decoding after each block, ordered by endpoint causal share; thin curves are runs, thick curves group means, and the horizontal reference is chance $(1/5)$. (c) Target energy fraction before and after the final MLP at position 100. (d) Candidate local routes to a similar endpoint; the upper route is a hypothesis, not a claim of globally independent pathways. (e) Source removal reduces target energy specifically through the recomputed MLP. (f) Source rotation transfers state phase to the target. Panels (c,e,f) use the pre-MLP energy-dominant mode as source, and the other as target; faint lines are runs and thick lines are means. Both directions and norm-matched residual controls are evaluated.

For these interventions, let u be the residual input to the final MLP, $w(u) = \mathrm{MLP}(\mathrm{LN}_2(u))$, and $h = u + w(u)$. We designate the mode with larger pre-MLP energy as source S, and the other as target T, independently of convergence labels. In all five slow runs, source/target also coincide with higher/lower endpoint causal share; this correspondence holds in 23 of 25 fast runs. In slow runs, the median target energy fraction rises from 0.027 before the MLP to 0.566 after it. Endpoint energy therefore obscures a large asymmetry immediately upstream.

A rise in target energy across the MLP does not by itself identify how that energy was produced. Fast seed 23 and slow seed 21 both have a pre-MLP target share near 0.027, rising to 0.519 and 0.486, respectively. We therefore test whether changing the source before the MLP affects the target.

To test source dependence, remove or rotate the source by δ and compare the recomputed output $u + \delta + w(u + \delta)$ with the bypass output $u + \delta + w(u)$. Only the former lets the MLP respond. Subtracting norm-matched residual-control effects isolates specific target-energy loss under removal and progress toward the shifted-state target centroid under rotation. Appendix F defines both normalizations and their aggregation.

Despite their similar energy profiles, the two runs have source-to-target removal effects of -0.112 and 0.776, and rotation effects of -0.030 and 0.755, respectively (Table 4).

The slow group shows strong forward effects and weak reverse effects. Median source-to-target specific energy loss is 0.794, versus 0.057 in reverse; median specific phase progress is 0.783, versus 0.030 in reverse. Fast-group forward medians are close to zero (-0.055 for energy loss and -0.017 for phase progress). The combination of target creation across the MLP, selective loss under source removal, and content-specific transfer under source rotation supports *cross-mode synthesis* in the slow balanced runs. Removal alone could reflect generic disruption, and an energy increase alone could reflect independent formation; together with directionality and phase transfer, the evidence is more discriminating.

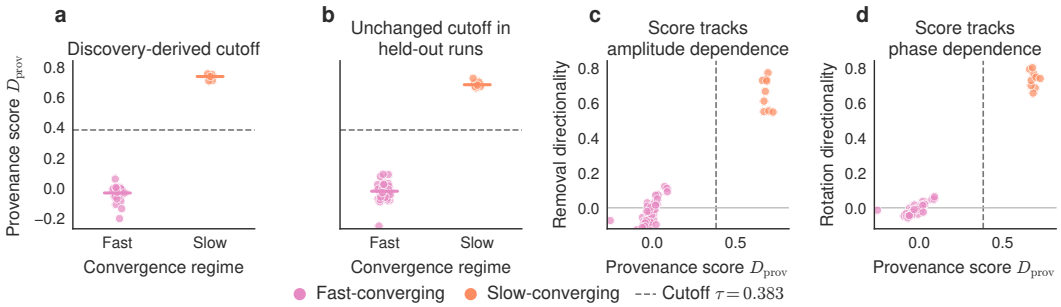


Figure 5: **A discovery-derived directional rule aligns with held-out convergence.** (a) Discovery runs, colored by convergence. The threshold $\tau = 0.383$ is the midpoint between the maximum fast and minimum slow discovery scores. (b) Without changing the score or threshold, the rule aligns with all 61 fast and nine slow held-out runs. (c–d) Held-out score versus related removal and rotation directionality measurements, characterizing the intervention signature.

3.5 A DISCOVERY-DERIVED PROVENANCE RULE ALIGNS WITH HELD-OUT CONVERGENCE

Does the directional signature recur beyond the discovery runs? We compute a provenance score from state-matched donor interventions using the same fixed assay in discovery and held-out runs. A donor supplies the source-mode component for a shifted state; we then measure the target phase progress in the final MLP's write, relative to norm-matched residual controls. If $D_{S\to T}$ is the resulting directed effect, the score is

$$D_{\text{prov}} = D_{S\to T} - D_{T\to S}. \tag{2}$$

It summarizes directional asymmetry, rather than representation amount. The donor-based score and the direct rotation assay in Figure 4 use distinct interventions and measurement locations (Appendix H).

We calculate the cutoff using only discovery convergence labels, with the same 30k boundary as above. The maximum fast score is 0.0586208 and the minimum slow score is 0.7072251; their midpoint gives $\tau \approx 0.383$. We apply the score definition, orientation, and cutoff without adjustment to held-out runs.

Applied unchanged to regularized seeds 31–100, the rule aligns with the convergence grouping: all 61 fast runs fall below the cutoff and all nine slow runs above it (Figure 5). The maximum fast stable-hit step is 21.0k and the minimum slow step is 40.0k. Lower-mode causal share also separates the discovery groups (Figure 3a). The provenance score's contribution is therefore mechanistic characterization of directional dependence, not superiority as a classifier; the held-out result extends this signature–regime correspondence to additional runs.

3.6 MODE MULTIPLICITY AND CAUSAL ALLOCATION IN C_7

Does the association between mode multiplicity and convergence extend beyond two modes? In C_7, runs with energy distributed over multiple modes converge earlier than single-mode-dominant runs (Figure 6c). Baseline endpoints span nearly one to almost three effective modes; 29 of 30 reach the target, with median stable-hit step 25.2k among those that converge, and one is censored at the 200k budget. Regularization concentrates energy across three modes (median $N_{\text{eff}}^{E} = 2.990$) and lowers the median stable-hit step to 18.7k. All 30 regularized runs converge, but their range is 13.2–43.6k, with seed 7 reaching the target at 43.6k. The pattern associates multiplicity with earlier convergence; it does not require perfectly equal energy in all three modes or imply uniformly rapid regularized convergence.

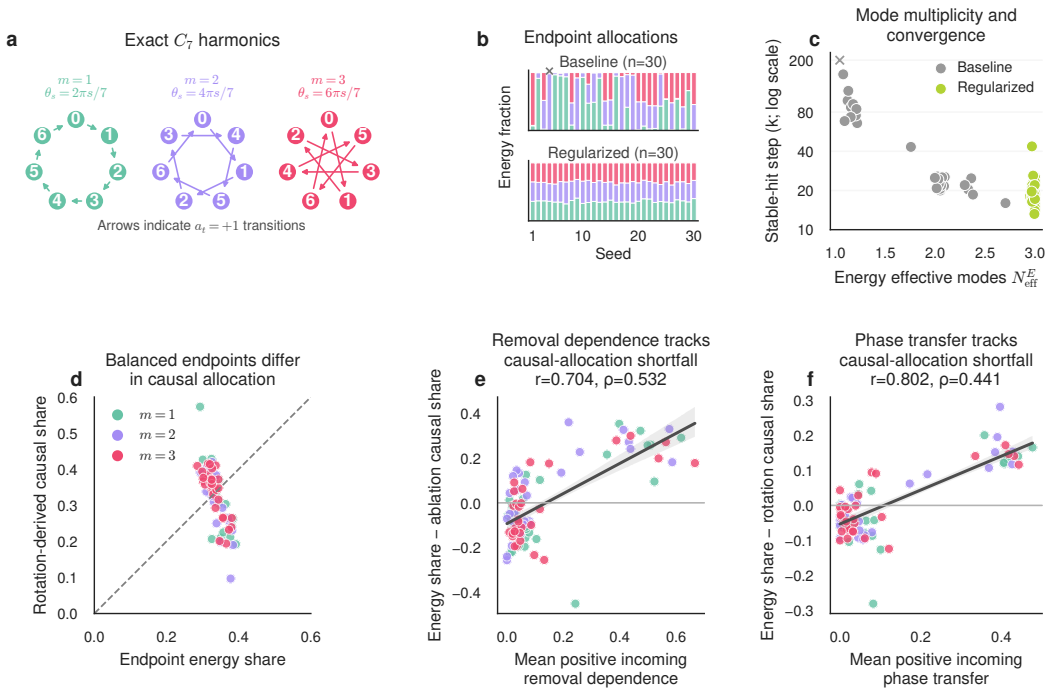


Figure 6: **Representation and dependence in the three-mode C_7 setting.** (a) Exact harmonic codes. (b) Endpoint energy fractions for baseline and regularized cohorts, each with 30 runs. (c) Effective modes and stable-hit step; the small $\times$ marks the baseline run censored at the training budget. All regularized runs converge, including one at 43.6k. (d) Energy share versus rotation-derived causal share in regularized runs. (e–f) Causal-allocation shortfall versus incoming removal or phase-transfer dependence. Each incoming value averages the positive parts of both directed edges into that mode. The 90 mode observations belong to 30 runs; correlations are descriptive and bands use whole-run bootstrap resampling.

With three complete codes, a target can depend on either of two other modes. We therefore retain all six directed edges and ask whether the energy balance reached by regularization also equalizes causal allocation.

Despite tightly clustered regularized energy shares, causal shares vary across modes and runs (Figure 6d). For a target mode m, define incoming dependence as the mean positive directed effect from the two other modes. We compare it with allocation shortfall $p_m - q_m$, where q_m is the normalized removal-derived or rotation-derived causal response, matched to the assay. A positive shortfall means that the mode accounts for less causal response than its energy share would suggest; it does not mean that the mode has no function.

Incoming removal dependence is positively associated with removal-allocation shortfall ($r = 0.704$, $\rho = 0.532$). Incoming phase transfer is associated with rotation-allocation shortfall ($r = 0.802$, $\rho = 0.441$). These correlations summarize 90 mode observations nested within 30 regularized models, not 90 independent models. Figure 6e–f uses whole-run bootstrap bands, and Appendix I supplies the numerical definitions and sensitivity to using the maximum incoming edge. Thus, the three-mode setting supports the broader distinction between representation amount and causal organization, without requiring the same fast/slow separation observed in C_5.

4 Related work

Fourier circuits and algebraic solutions. Nanda et al. (2023) identify a trigonometric circuit for modular addition, while clock/pizza mechanisms show that related circular geometry can support different computations (Zhong et al., 2023). Chughtai et al. (2023) characterize group-representation circuits with shared algorithmic structure but variable features and circuits across networks. Fourier analyses also extend beyond addition to modular polynomials (Furuta et al., 2024). Margin-maximization theory explains Fourier and group-representation feature selection in stylized algebraic models (Morwani et al., 2024), without directly explaining our eight-layer transformer's regimes. An approximate Chinese Remainder Theorem abstraction unifies diverse modular-addition implementations (McCracken et al., 2025). We build on these accounts by examining directed dependence among complete codes within one trained model, not by establishing distinct abstract algorithms.

State-tracking mechanisms. Automata studies analyze how transformers learn parallel shortcuts (Liu et al., 2023). Most closely, Li et al. (2025) identify associative-scan and parity-assisted mechanisms for permutation composition, relating them to convergence and generalization and steering their emergence with intermediate tasks. Our question concerns dependence between coexisting codes that each identify the complete state. Recent quotient/coset tracking results instead characterize partial state information (Zhang & Li, 2026). Prime cycles make this contrast explicit: each nonzero harmonic here is injective, and each prefix supplies a state-tracking target.

Internal solutions, trajectories, and representation strength. Mechanistic mode connectivity relates parameter-space connectivity to predictive mechanisms (Lubana et al., 2023); its optimization "modes" are distinct from our Fourier modes. Latent-state models summarize differences among training trajectories (Hu et al., 2023). Grokking describes delayed generalization after overfitting on small algorithmic datasets (Power et al., 2022); our convergence regimes are not evidence of that phenomenon. Alternating Gradient Flows characterizes staged Fourier-feature learning in two-layer quadratic modular-addition networks (Kunin et al., 2025), rather than the training mechanism of our deep transformer. Our interventions instead examine code construction inside trained models. Feature strength also depends on complexity, learning order, and position (Lampinen et al., 2024), while probe accuracy need not imply task relevance (Ravichander et al., 2021). Our target is the directional dependence between coexisting complete codes, beyond whether each can be decoded or used.

Interventions and their interpretation. Interventions on Othello state representations (Li et al., 2023) and multidimensional circular features (Engels et al., 2025) connect internal geometry to behavior. Causal abstraction tests whether aligned neural and high-level variables respond consistently to interchange interventions (Geiger et al., 2021). Patching conclusions depend on metrics, perturbations, and the scope of the tested behavior (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). Subspace edits can activate dormant pathways (Makelov et al., 2024); even successful causal alignments can admit multiple explanations (Méloux et al., 2025). With unrestricted nonlinear alignment maps, causal abstraction can become uninformative (Sutter et al., 2025). Here, task-defined two-dimensional harmonic coordinates are fitted on calibration data and held fixed during intervention evaluation, without optimizing a nonlinear alignment to intervention success. Combining bidirectional removal and rotation with bypass contrasts and norm-matched controls characterizes dependence in those coordinates, but does not identify a unique complete algorithm.

5 Discussion and limitations

Our results distinguish geometric presence, usability by the existing readout, and cross-mode dependence. The lower-share code in slow balanced C_5 runs is neither absent nor wholly unused; its formation across the final MLP depends directionally on another code. The held-out signature and three-mode allocation–dependence relationships extend this distinction beyond illustrative runs.

What endpoint regularization constrains. The regularizer acts on the final residual, so it permits a solution in which the immediately preceding MLP produces an additional mode. This placement

may contribute to the occurrence of synthesis, but we have not varied the regularization location or objective to test that possibility. Requiring endpoint energy balance does not guarantee separately formed codes. The frequency of these solutions cannot be generalized to other regularization schemes, and the data do not identify a particular polynomial implemented by the MLP or an optimization sequence in which a single-mode solution precedes synthesis.

Endpoint structure versus learning history. Runs stop at stable hit or the budget, so their endpoints have different training durations. Learning-rate histories are also not disentangled from convergence regime. Regularization changes the optimization objective, and the comparison does not isolate energy balance as a mediator. Consequently, neither balanced energy nor cross-mode synthesis is identified as the cause of faster or slower learning. These limitations concern why a solution was learned; they do not negate the directional responses measured within the trained models. Weak directed effects in fast runs likewise do not establish globally independent computational pathways.

Scope of the evidence. We study small prime cycles and one architecture family, with five slow discovery runs and nine slow held-out runs. Synthetic tasks expose exact state geometry but do not establish prevalence in language models. Recovery is a within-sample diagnostic, and edits may produce off-distribution states. Directionality, bypass comparisons, and matched controls constrain alternatives without resolving every subspace-identification ambiguity. Figure 5c–d relates the score to closely related intervention measurements, not independent external validation. In C_7, normalized allocation shortfalls are dependent by construction, and mode observations are clustered within runs; their correlations identify an association, not its causal direction. The supported conclusion is that directional interventions reveal causal organization left unspecified by endpoint energy alone.

AI USE STATEMENT

Generative AI tools (ChatGPT/Codex) assisted research planning and discussion, development and debugging of experiment and analysis code, interpretation of results, literature checking, manuscript drafting and revision, and preparation of figures and reproducibility materials. The authors retain responsibility for the experimental decisions, accuracy of the code and results, fidelity of citations, and all claims and submitted artifacts.

ETHICS STATEMENT

The experiments concern synthetically generated cyclic sequences and controlled interventions on small research models. The task data do not involve human participants or personal records. Claims are restricted to the analyzed models and do not imply reliability or safety guarantees for deployed systems.

REPRODUCIBILITY STATEMENT

Appendices A–K describe the experimental and analysis protocols and report run-level results. Code and model checkpoints are not included with this submission.

REFERENCES

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6243–6267, 2023. URL https://proceedings.mlr.press/v202/chughtai23a.html.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d3221cdb27e49d9c1cd35ad254feccfe-Abstract-Conference.html.

Hiroki Furuta, Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. Towards empirical interpretation of internal circuits and properties in grokked transformers on modular polynomials. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2402.16726.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL https://arxiv.org/abs/2404.15255. arXiv:2404.15255.

Michael Y. Hu, Angelica Chen, Naomi Saphra, and Kyunghyun Cho. Latent state models of training dynamics. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=NE2xXWo0LF.

Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James Simon, Michael R. Deweese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/06cbd2e81dfbd3bb4cb0abce95b32584-Abstract-Conference.html.

Andrew Kyle Lampinen, Stephanie C. Y. Chan, and Katherine Hermann. Learned feature representations are biased by complexity, learning order, position, and more. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2405.05847.

Belinda Z. Li, Zifan Carl Guo, and Jacob Andreas. (How) do language models track state? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34429–34452, 2025. URL https://proceedings.mlr.press/v267/li25r.html.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.13382.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.10749.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Ekdeep Singh Lubana, Eric J. Bigelow, Robert P. Dick, David Krueger, and Hidenori Tanaka. Mechanistic mode connectivity. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22965–23004, 2023. URL https://proceedings.mlr.press/v202/lubana23a.html.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/70b8505ac79e3e131756f793cd80eb8d-Abstract-Conference.html.

Gavin McCracken, Gabriela Moisescu-Pareja, Vincent Létourneau, Doina Precup, and Jonathan Love. Uncovering a universal abstract algorithm for modular addition in neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3edb234091dca2023308398dbf824850-Abstract-Conference.html.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/51157713f1f1f572a05647e1aef1b1ed-Abstract-Conference.html.

Depen Morwani, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. Feature emergence via margin maximization: Case studies in algebraic tasks. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.07568.

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.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177. arXiv:2201.02177.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding, 2021. URL https://arxiv.org/abs/2104.09864. arXiv:2104.09864.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/dbb98528c9870377f3f0d133aae6050b-Abstract-Conference.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

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

Zhiyu Zhang and Yupeng Li. Tracking states or tracking cosets? An algebraic account of learned state tracking, 2026. URL https://arxiv.org/abs/2609.29951. arXiv:2609.29951.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/56cbfbf49937a0873d451343ddc8c57d-Abstract-Conference.html.

A TASK, ARCHITECTURE, AND TRAINING PROTOCOL

Data. For each length-100 example, actions are sampled independently and uniformly from $\{0, \ldots, n-1\}$. Targets are the cumulative sums modulo n, including the current action. There is no next-token offset: output at position t predicts s_t from $a_{1:t}$. Chance state accuracy is $1/n$. The energy-regularized training implementation constructs a pool of 50,000 sequences with NumPy's `default_rng(seed)`, and samples minibatches with replacement from that pool. A common evaluation set contains 8,000 sequences generated with seed 12345. Evaluation accuracy over all tokens and accuracy at the last position are distinct logged quantities. The stopping criterion uses the latter.

Architecture. The implementation has a learned n-state embedding, eight causal self-attention/MLP blocks, and a final LayerNorm followed by a linear n-class head. Each block is pre-normalized:

$$u_\ell = h_{\ell-1} + \mathrm{Attention}_\ell(\mathrm{LN}_{\ell,1}(h_{\ell-1})), \tag{3}$$

$$h_\ell = u_\ell + \mathrm{MLP}_\ell(\mathrm{LN}_{\ell,2}(u_\ell)). \tag{4}$$

Width is 256, the number of heads is eight (head dimension 32), and the MLP expands to 1024 dimensions with GELU activation. Dropout is 0.1 during training and disabled during analysis. RoPE rotates queries and keys with base 10,000; maximum cached length is 132. The final representation used for endpoint analysis is h_8, before the final LayerNorm. This differs from both the final MLP input u_8 and its write $w_8 = h_8 - u_8$.

Optimization. The energy-regularized training implementation uses AdamW (Loshchilov & Hutter, 2019), batch size 128, initial learning rate 10^{-4}, weight decay 0.01, 300 warmup steps, and gradient-norm clipping at 1.0. It halves the base learning rate after 30 evaluations without an all-position accuracy improvement greater than 10^{-3}, with at most three reductions. The optimization budget is 200,000 updates; step labels are zero-based, so a budget-exhausted archive can report 199,999. The final saved model is the last training state. The separately saved best-accuracy model is not substituted for it in the endpoint analyses. The inspected training implementation and model configuration specify these settings; the run-level stopping records identify the actual endpoint step separately for each run.

Energy-balance objective. At each update, let $\mu_{t,s}$ be the minibatch mean of the final-block residual at position t conditional on state s. Subtract the equally weighted mean over states and take an orthonormal DFT along the state axis. For $M = (n-1)/2$, sum squared norms at frequencies m and $n-m$ to obtain $\widetilde{E}_{t,m}$. Let $p_{t,m} = \widetilde{E}_{t,m} / \max(\sum_j \widetilde{E}_{t,j}, 10^{-8})$. The loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + 0.5\, \frac{\sum_t v_t \sum_{m=1}^{M} (p_{t,m} - 1/M)^2}{\max(1, \sum_t v_t)}, \tag{5}$$

$$v_t = \mathbf{1}\{\text{each state occurs at least twice at } t\}.$$

The cross-entropy averages all supervised tokens. The regularizer acts at every update, on valid positions of that minibatch, and remains at coefficient 0.5 throughout the budget (the hold and end settings are both 200,001). Baseline omits this term. The loss balances fractions and does not directly maximize absolute harmonic energy, functional share, or independence.

Convergence and checkpoint cadence. At recorded evaluation step k, define $A_{\mathrm{last}}(k)$ as the fraction of evaluation sequences with a correct state at position 100. First-hit step is the first k with $A_{\mathrm{last}}(k) \geq 0.97$. Stable-hit step is the step of the *third* consecutive evaluation meeting this inequality; a failed evaluation resets the count. It is not the first member of the three-evaluation window and does not assert permanent accuracy above threshold at unobserved later steps. In the training records for all reported cohorts, including the 70 held-out runs, the final recorded step equals the stable-hit step for every converged run; the non-converged C_7 baseline seed 4 ends at step 199,999. These endpoints therefore correspond to different training durations rather than a common fixed step. The training implementation stops at the stable hit or the budget. Evaluation is every 200 updates and at the final budget step. Checkpoints are saved every 1k steps through 20k, every 2k through 40k, and every 5k thereafter, plus milestone and final checkpoints. These are separate

from the 1k sampling selected for the optional accuracy/energy trajectory summaries. The 200-step cadence defines convergence; the optional trajectories use 1k sampling.

B HARMONIC REPRESENTATIONS AND FUNCTIONAL METRICS

B.1 STATE CENTROIDS, ENERGIES, AND COORDINATE SYSTEMS

At a specified residual location and token position t, define $\mu_{t,s} = \mathbb{E}[h_t \mid s_t = s]$ and $\bar{\mu}_t = n^{-1} \sum_s \mu_{t,s}$. With the $1/n$ DFT convention,

$$f_{t,m} = \frac{1}{n} \sum_{s=0}^{n-1} (\mu_{t,s} - \bar{\mu}_t) e^{-2\pi i m s/n}, \qquad E_{t,m} = 2\|f_{t,m}\|_2^2. \tag{6}$$

The real pair at $m, n-m$ has mean squared contribution $E_{t,m}$. An orthonormal DFT gives the same quantity after dividing pair energy by n; energy fractions do not depend on this scaling. For the positionwise endpoint representation summaries in Figures 1–3, fractions are formed at each valid position and then averaged over positions. Thus $p_m = \mathrm{mean}_t \, E_{t,m} / \sum_j E_{t,j}$, with the division inside the mean. This differs in general from normalizing the mean energies.

The effective number and entropy balance are different summaries:

$$N_{\mathrm{eff}}^E = \frac{1}{\sum_m p_m^2}, \qquad B(p) = \frac{-\sum_m p_m \log p_m}{\log M}. \tag{7}$$

The two-mode threshold in Figures 1–2 is $N_{\mathrm{eff}}^E \geq 1.5$; it is distinct from entropy balance. Both figures also show continuous energy allocation or effective modes.

Two coordinate implementations are used and should not be conflated. The generic functional calibration first removes the empirical mean at each position, pools centered tokens across positions, and computes state centroids. Multiplying those centroids by the real Fourier basis gives a joint coordinate basis with two columns per mode. In the final-MLP assays, state centroids are fitted at position 100 separately at the input and output/write locations; each Fourier pair is orthonormalized by SVD. In both cases all modes are concatenated into B, and coordinates are obtained jointly with a pseudoinverse:

$$z = (h - \mu)(B^+)^\top, \qquad h_m = z_m B_m^\top. \tag{8}$$

Here h is a row vector and z_m denotes its two coordinates for mode m. This avoids assigning overlapping components by separately projecting onto nonorthogonal planes. The final-MLP implementations reject rank-deficient pairs and joint condition numbers above 10^6 and use pseudoinverse tolerance 10^{-10}. Generic functional calibration uses its documented 10^{-8} tolerance. All-mode removal retains $h - \sum_m h_m$, including its mean and non-harmonic components.

B.2 CAUSAL ALLOCATION FROM THE EXISTING READOUT

Let $\pi_s(h)$ be the probability of class s after the unchanged final LayerNorm and linear head. A semantic rotation of mode m toward state shift Δ edits its coordinates by angle $2\pi m\Delta/n$ in the generic harmonic basis, leaving the other modes intact. For target $s' = s + \Delta \bmod n$, the gain is

$$g_{m,\Delta} = \log \pi_{s'}(h^{(m,\Delta)}) - \log \pi_{s'}(h). \tag{9}$$

The functional pipeline uses three separately generated pools: plane calibration, residual-control matching, and final testing, each with 256 sequences. On the test pool, responses are restricted to tokens whose intact prediction equals the true state. Values are averaged within source-state/nonzero-shift cells and then macro-averaged over those cells. Twenty residual control draws are used. The control subspaces and transforms are fitted on the matching pool, including a mode-specific magnitude match. If $\bar{g}_m$ and $\bar{g}_m^{\mathrm{ctrl}}$ are the macro-averaged target log-probability gains, define

$$I_m^{\mathrm{rot}} = [\bar{g}_m - \bar{g}_m^{\mathrm{ctrl}}]_+, \quad q_m^{\mathrm{rot}} = \frac{I_m^{\mathrm{rot}}}{\sum_j I_j^{\mathrm{rot}}}, \quad B_F = B(q^{\mathrm{rot}}). \tag{10}$$

The positive part is applied *after* averaging and control subtraction. A zero total influence makes allocation undefined, rather than uniformly distributed. The source tables used for the reported cohorts have valid allocations.

Table 1: Cohort accounting for the reported results. All seed ranges are inclusive.

Task / condition	Seeds	Runs	Converged	Main role
C_5 baseline	1–30	30	30	Figures 1–2
C_5 regularized discovery	1–30	30	30	Figures 2–5
C_5 regularized held-out	31–100	70	70	Figure 5
C_7 baseline	1–30	30	29	Figure 6b–c
C_7 regularized	1–30	30	30	Figure 6b–f

Removal-derived allocation instead uses the positive macro-average decrease in the intact source state's log probability:

$$I_m^{\mathrm{rem}} = [\mathrm{macroavg}\big[\log \pi_s(h) - \log \pi_s(h - h_m)\big]]_+, \qquad q_m^{\mathrm{rem}} = I_m^{\mathrm{rem}} / \sum_j I_j^{\mathrm{rem}}. \tag{11}$$

No residual-control subtraction is included in this removal allocation. Figure 3 uses q^{rot}; Figure 6d also uses it, whereas (e) uses q^{rem} and (f) uses q^{rot}. These normalized responses quantify relative causal allocation under specified perturbations. Because the readout is nonlinear through LayerNorm and softmax, they do not sum to an additive decomposition of task performance.

B.3 Mode-only recovery and isolated rotation

The Figure 3 recovery implementation fits its pooled harmonic planes from a fixed 4,000-sequence probe (seed 888), averaging centered state centroids across positions before fitting SVD planes. It evaluates on that same probe, over all 100 positions. This shared fit/evaluation is explicitly a within-sample diagnostic. With $h_0 = h - \sum_j h_j$, predictions at h_0 give r_0, and predictions at $h_0 + h_m$ give r_m, each conditioned on the intact prediction being correct. The recovery score

$$R_m = (r_m - r_0)/(1 - r_0) \tag{12}$$

is one for full restoration and zero for no improvement over all-mode removal. It may be negative and is undefined if $r_0 = 1$.

For isolated rotation, a two-dimensional orthogonal Procrustes map is fitted between mode centroids for states s and $s + \Delta$. The map acts on z_m, and only this rotated component is added back to h_0. Target adoption is the fraction of originally correct tokens predicted as $s + \Delta$, averaged over shifts $\{-2, -1, 1, 2\}$. This orthogonal fit is distinct from the least-squares semantic map used for Figure 4. Ranking by q_m^{rot} precedes inspection of R_m and isolated adoption; the lower-ranked mode is not chosen separately for each assay.

C Cohorts, regime definitions, and endpoint selection

Seed matching refers to the 30 baseline/regularized training-seed pairs within a task, not to matching final representations. The C_5 fast/slow classification is $k_{\mathrm{stable}} < 30{,}000$ versus $\geq 30{,}000$ throughout the paper. It is not recomputed from each intervention metric. Discovery and held-out seeds are disjoint, but share architecture, task distribution, and analysis procedures; this is replication across training seeds rather than across task families.

The sole nonconverged C_7 baseline run is seed 4. Its endpoint allocation is retained in Figure 6b and its time is displayed at the censoring limit in (c); it is not counted as converged at that limit. Functional and directed-edge tables contain 29 eligible baseline runs and 30 regularized runs (177 mode rows and 354 directed edges). Panels (d–f) use only the regularized 30. No C_5 run in the displayed cohorts is excluded. For C_7, no fast/slow labels are assigned.

D Complete baseline and regularized convergence results

Table 2 records the seed-matched C_5 data from the Figure 2 source. The current Figure 1 and Figure 2 sources agree exactly on seed-level stable-hit values and on the 10 two-mode classifications.

Their baseline mode-1 fractions differ by at most 0.0021 across the two endpoint-analysis exports; we retain each figure's own numerical inputs rather than assert identical calibration values. The display of seeds 1–10 in Figure 1c is only a compact illustration; Figure 1d and all condition-level comparisons use 30 runs. The observed gap in convergence is visible in the continuous values, so the reported separation does not arise solely from choosing a binary label.

Table 2: All 30 C_5 seed pairs. Steps are in thousands; p_1 determines $p_2 = 1 - p_1$. B/R denote baseline/regularized.

Seed	k_B	k_R	$p_{1,B}$	$p_{1,R}$	N_B^E	N_R^E
1	44.6	14.8	0.969	0.491	1.064	1.999
2	44.4	11.4	0.040	0.501	1.083	2.000
3	17.2	15.0	0.461	0.491	1.988	1.999
4	54.0	19.0	0.044	0.513	1.091	1.999
5	46.4	14.4	0.964	0.489	1.074	1.999
6	43.8	14.8	0.960	0.513	1.084	1.999
7	19.4	17.6	0.510	0.493	1.999	2.000
8	47.4	13.0	0.038	0.505	1.079	2.000
9	19.2	15.4	0.653	0.515	1.829	1.998
10	41.2	48.8	0.956	0.442	1.092	1.974
11	40.2	42.2	0.032	0.557	1.066	1.974
12	15.8	15.6	0.535	0.529	1.990	1.993
13	43.2	46.6	0.964	0.454	1.075	1.983
14	47.4	16.6	0.950	0.488	1.105	1.999
15	46.2	12.4	0.968	0.508	1.065	1.999
16	48.8	15.4	0.964	0.527	1.075	1.994
17	14.6	19.2	0.384	0.496	1.899	2.000
18	19.0	17.2	0.595	0.487	1.930	1.999
19	48.2	12.4	0.035	0.472	1.073	1.994
20	42.4	18.8	0.035	0.483	1.073	1.998
21	50.6	45.0	0.032	0.499	1.067	2.000
22	43.6	11.0	0.962	0.488	1.079	1.999
23	46.4	16.4	0.958	0.481	1.087	1.997
24	49.8	15.4	0.039	0.505	1.080	2.000
25	16.4	14.4	0.450	0.524	1.980	1.995
26	19.0	16.6	0.576	0.493	1.955	2.000
27	19.6	13.6	0.336	0.494	1.806	2.000
28	48.4	36.4	0.044	0.565	1.091	1.967
29	46.4	18.4	0.970	0.500	1.062	2.000
30	19.6	13.4	0.602	0.468	1.919	1.992

Descriptive seed correspondence. The five slow regularized seeds (10, 11, 13, 21, and 28) are also single-mode-dominant and slow at baseline (Table 2). Their baseline dominant modes are respectively $1, 2, 1, 2, 2$, matching the regularized source modes in Table 4. The separate baseline and regularized notebooks share the model construction and main training settings. Both reset PyTorch, NumPy, and Python random seeds before constructing each new model, generate the training pool with `default_rng(seed)`, and draw minibatches with `default_rng(seed + 10000)`. Fresh starts under the same environment and configuration are therefore designed to match initial weights, training pools, and minibatch index sequences. Both implementations can instead load existing pools or resume checkpoints; complete equality across all archived runs has not been verified. The observed seed correspondence does not establish that initialization determines the solution type or that baseline training trajectories persist under regularization.

E FUNCTIONAL ALLOCATION UNDER ENERGY REGULARIZATION

Table 3 contains both ranks for every regularized discovery run. Energy shares use Figure 3's positionwise convention; causal shares use the independent-split functional pipeline. These calibration conventions remain distinct. Recovery and isolated adoption are within-sample assays (Appendix B). The lower-share mode in slow runs remains substantially usable: weak independent usability is relative, not zero contribution.

Table 3: Regularized C_5 functional assays. F/S: convergence label; L/H: lower/higher causal-share rank. A is isolated target adoption.

Seed	Group	m_L	p_L	q_L	R_L	R_H	A_L	A_H
1	F	2	0.509	0.467	0.923	0.933	0.937	0.935
2	F	1	0.501	0.479	0.937	0.923	0.946	0.935
3	F	2	0.509	0.477	0.949	0.961	0.959	0.964
4	F	2	0.487	0.492	0.979	0.980	0.979	0.982
5	F	2	0.511	0.498	0.908	0.890	0.923	0.908
6	F	2	0.487	0.493	0.984	0.977	0.986	0.979
7	F	1	0.493	0.488	0.960	0.958	0.964	0.962
8	F	1	0.505	0.444	0.917	0.977	0.928	0.975
9	F	2	0.485	0.468	0.834	0.957	0.863	0.963
10	S	2	0.558	0.223	0.743	0.995	0.758	0.987
11	S	1	0.557	0.263	0.720	0.995	0.742	0.986
12	F	2	0.471	0.500	0.910	0.942	0.920	0.950
13	S	2	0.546	0.189	0.546	0.993	0.670	0.965
14	F	1	0.488	0.491	0.932	0.968	0.942	0.973
15	F	1	0.508	0.490	0.927	0.955	0.938	0.962
16	F	1	0.527	0.476	0.935	0.946	0.944	0.951
17	F	2	0.504	0.488	0.957	0.910	0.966	0.928
18	F	2	0.513	0.492	0.933	0.932	0.943	0.942
19	F	1	0.472	0.478	0.924	0.966	0.936	0.967
20	F	2	0.517	0.480	0.972	0.919	0.965	0.918
21	S	1	0.499	0.288	0.618	0.996	0.734	0.942
22	F	1	0.488	0.499	0.970	0.968	0.973	0.972
23	F	2	0.519	0.463	0.928	0.915	0.940	0.920
24	F	1	0.505	0.468	0.934	0.980	0.945	0.978
25	F	2	0.476	0.476	0.919	0.942	0.933	0.946
26	F	1	0.493	0.495	0.948	0.951	0.957	0.958
27	F	1	0.494	0.498	0.975	0.826	0.978	0.854
28	S	1	0.565	0.251	0.796	0.998	0.753	0.989
29	F	2	0.500	0.474	0.926	0.919	0.940	0.933
30	F	2	0.532	0.495	0.976	0.941	0.977	0.951

F Complete cross-mode synthesis analysis

F.1 Layerwise decoding

The layerwise analysis includes all 30 regularized discovery runs. At each of the eight post-block residual locations, fit harmonic geometry using 2,000 sequences (seed 360001), and evaluate nearest state-centroid accuracy in each two-dimensional coordinate block using 1,000 independent sequences (seed 360003). The generator first produces length-150 sequences and then takes the length-100 prefix to preserve the archived random-number layout. The measurement is at position 100. Accuracy is the fraction of evaluation examples whose nearest centroid has the true state label; all evaluation examples enter this decoding score. Figure 4's high/low ordering is fixed by endpoint causal share. Its thin curves show individual runs across all eight layers. Chance is $1/5$.

F.2 MLP-mediated removal and rotation

The direct synthesis analysis separately fits geometry on 3,000 sequences (seed 48004801) and evaluates on 3,000 (seed 48004802), at position 100. It uses the final MLP input and final residual, with native LayerNorm recomputed after the edit. The calibrated input source direction is inherited from the archived orientation and checked against the current pre-MLP energy rank. Source is never selected by the largest observed intervention effect.

For edit δ to u, define

$$h_\delta^{\mathrm{rec}} = u + \delta + w(u+\delta), \qquad h_\delta^{\mathrm{byp}} = u + \delta + w(u). \tag{13}$$

For source removal, $\delta = -u_S$. Let $\mathcal{E}_T(H)$ mean the target harmonic's energy in evaluation state centroids after projecting H into the *fixed* post-MLP target plane. The raw mediated target loss is

$$L_{S\to T} = \frac{\mathcal{E}_T(h_\delta^{\mathrm{byp}}) - \mathcal{E}_T(h_\delta^{\mathrm{rec}})}{\max(\mathcal{E}_T(h), 10^{-12})}. \tag{14}$$

Four Gaussian residual-control edits are projected outside the full input harmonic span and individually rescaled to match each token's source-edit norm. Specific loss is the raw loss minus the mean analogous control loss. Energy uses all evaluation examples grouped by true state. It is not a decrease in classification accuracy, and values outside $[0, 1]$ are possible.

For rotation, the source-coordinate map $A_{S,\Delta}$ is fitted by least squares to calibration centroids $c_{s,S}A_{S,\Delta} \simeq c_{s+\Delta,S}$. The edit is $(z_S A_{S,\Delta} - z_S)B_S^\top$. With fixed output coordinates and desired target displacement $d_{s,T,\Delta} = c_{s+\Delta,T} - c_{s,T}$, define

$$P_T(H; h, s, \Delta) = \frac{(z_T(H) - z_T(h)) \cdot d_{s,T,\Delta}}{\|d_{s,T,\Delta}\|^2}. \tag{15}$$

Mediated progress is $P_T(h_\delta^{\mathrm{rec}}) - P_T(h_\delta^{\mathrm{byp}})$, followed by subtraction of the mean of four matched residual-control effects. Average over originally correct evaluation tokens and shifts $1, 2, 3, 4$. Zero denotes no specific movement toward the desired phase; one denotes a displacement equal to the calibration centroid difference along that direction, not necessarily perfect classification. The eligibility guard requires at least 750 clean-correct evaluation examples. Both intervention directions are evaluated for every run. Table 4 provides the complete run-level direction-by-intervention matrix.

Table 4: All 30 discovery runs: target energy fractions before/after the final MLP and both directions of specific removal loss (L) and rotation progress (P). Arrows indicate source-to-target or reverse.

Seed	m_S	p_T^{pre}	p_T^{post}	$L_\rightarrow$	$L_\leftarrow$	$P_\rightarrow$	$P_\leftarrow$
1	1	0.169	0.509	-0.085	-0.013	-0.023	-0.002
2	2	0.166	0.508	-0.074	0.005	-0.017	0.005
3	1	0.212	0.491	-0.081	0.016	-0.022	0.005
4	2	0.410	0.510	-0.004	-0.009	0.000	0.000
5	2	0.438	0.475	-0.096	0.077	-0.032	0.027
6	1	0.345	0.504	-0.061	0.034	-0.024	0.015
7	2	0.414	0.494	-0.003	0.044	-0.008	0.015
8	2	0.198	0.486	0.007	0.041	0.008	0.013
9	1	0.374	0.443	-0.208	0.093	-0.080	0.042
10	1	0.031	0.569	0.794	0.027	0.783	0.015
11	2	0.013	0.566	0.816	0.057	0.813	0.030
12	1	0.478	0.467	0.057	-0.011	0.029	-0.002
13	1	0.050	0.552	0.794	0.080	0.732	0.030
14	2	0.326	0.479	-0.002	0.074	-0.001	0.024
15	2	0.257	0.513	-0.055	0.021	-0.017	0.007
16	2	0.260	0.520	-0.116	0.081	-0.039	0.028
17	1	0.218	0.524	-0.043	-0.009	-0.009	-0.001
18	1	0.458	0.495	-0.011	-0.002	-0.003	0.002
19	2	0.193	0.463	-0.069	0.063	-0.026	0.021
20	1	0.254	0.526	-0.088	0.060	-0.022	0.015
21	2	0.027	0.486	0.776	0.030	0.755	0.017
22	2	0.450	0.476	0.020	0.028	0.013	0.011
23	1	0.027	0.519	-0.112	0.108	-0.030	0.031
24	2	0.223	0.493	-0.109	0.028	-0.037	0.010
25	1	0.312	0.456	-0.113	0.112	-0.038	0.038
26	2	0.321	0.517	-0.038	0.022	-0.011	0.012
27	2	0.358	0.505	-0.033	-0.004	-0.011	-0.001
28	2	0.005	0.583	0.767	0.148	0.830	0.057
29	1	0.100	0.503	-0.050	-0.008	-0.012	0.001
30	1	0.439	0.554	0.040	0.026	0.013	0.009

Table 5: Controls and data separation verified for the core results. Counts refer to training runs, not perturbations or tokens.

Analysis	Run count	Verified control or separation
Figure 3 causal share	30 discovery	Separate plane-fit, control-match, and final-test pools; 20 residual draws.
Figure 3 recovery / isolated rotation	30 discovery	Same fixed probe for geometry and response; explicit limitation.
Figure 4 layerwise decoding	30 discovery	Independent 2,000/1,000 fit/evaluation sequences.
Figure 4 final-MLP assays	30 discovery	Independent 3,000/3,000 fit/evaluation; both directions; four norm-matched residual draws; recompute/bypass contrast.
Figure 5 provenance	30 / 70	Separate 4,000-sequence fit, donor, and evaluation pools; two donor repeats and two residual draws; frozen direction and cutoff.
Figure 6 directed edges	59 eligible	All six directions; separate fit, donor, evaluation pools; matched residual controls for phase progress.

G Controls, scope, and calibration audit

Reverse-direction effects are a core result, not an optional control restricted to illustrative seeds. The MLP bypass contrast separates the response of the MLP from direct transport of the edited residual. Residual controls address nonspecific perturbation effects at matched norm; they do not guarantee that edited states remain on the training distribution. Native LayerNorm is part of the primary causal path and is recomputed for edited inputs.

The reported controls are those accompanying the numerical results in the table above. Optional frozen-LayerNorm and alternative-position branches in the implementation are not included as completed robustness results. Native-LayerNorm, bidirectional effects and norm-matched controls define the evidence used here.

H Provenance score and held-out protocol

H.1 A donor-based directional score

The provenance score measures the final MLP *write* $w(u)$, whereas the direct Figure 4 synthesis assay measures the final residual using a bypass contrast. Fit separate input and write geometries at position 100. For an evaluation example with state s, choose a donor with state $s + \Delta$ from the independent donor pool. Replace only source-mode component u_S by the donor component. For the edited write w', use

$$P^{w}_{S\to T} = \frac{(z_T(w') - z_T(w)) \cdot (c^{w}_{s+\Delta,T} - c^{w}_{s,T})}{\|c^{w}_{s+\Delta,T} - c^{w}_{s,T}\|^2}. \tag{16}$$

Two donor repeats are averaged for each nonzero shift. Two random residual edits outside the input harmonic span are norm-matched to the donor edit. Average donor-minus-control progress over shifts and originally correct examples to obtain $D_{S\to T}$. The source is the higher-energy mode at the input; ties are invalid under the frozen specification. Then $D_{\mathrm{prov}} = D_{S\to T} - D_{T\to S}$. There is no z-scoring, fitting of a removal weight, or combination with an ablation metric. The normalization is the squared calibration centroid displacement shown above. Positive values indicate stronger forward than reverse phase transfer; they are not probabilities.

The fit, donor, and evaluation pools each contain 4,000 sequences, with seeds 888, 3303301, and 3303302, respectively. Native LayerNorm is primary. The clean-correct eligibility guard is 1,000 examples. Source orientation is fixed before reading endpoint phenotype or convergence labels. The same assay is used in discovery and held-out scoring.

H.2 DISCOVERY-ONLY CUTOFF AND UNCHANGED HELD-OUT EVALUATION

Let F and S be the discovery runs with stable-hit step below and at least 30k, respectively. Compute

$$\tau = \frac{1}{2}\left(\max_{i\in F} D_{\mathrm{prov}\,i} + \min_{i\in S} D_{\mathrm{prov}\,i}\right). \tag{17}$$

The two extrema are 0.0586207655724138 and 0.7072251290082932, giving $\tau = 0.3829229472903535$. The score intervals are disjoint. No held-out score or label enters the calculation, and the cutoff is not readjusted on the held-out cohort. This is a discovery-derived analysis evaluated unchanged on additional runs, not a prospectively preregistered prediction of convergence speed.

Classify a score at least τ as slow. Among held-out runs, the 2×2 table contains 61 below-cutoff fast, zero above-cutoff fast, zero below-cutoff slow, and nine above-cutoff slow. Descriptive sensitivity, specificity, and balanced accuracy are each 1.0. These point values do not remove finite-sample uncertainty, especially for the nine slow runs. Table 6 includes all scores and times rather than only a classification summary.

Table 6: All 70 held-out regularized C_5 runs, shown in two blocks. Stable-hit step k is in thousands. Every score above the frozen cutoff belongs to the slow group.

Seed	k	D_{prov}	Group	Seed	k	D_{prov}	Group
31	49.4	0.6943	S	66	15.6	-0.0236	F
32	16.2	-0.0327	F	67	17.6	-0.0808	F
33	16.4	0.0199	F	68	12.6	0.0094	F
34	42.6	0.6821	S	69	15.8	-0.0745	F
35	40.0	0.6700	S	70	17.0	-0.0218	F
36	12.6	-0.0179	F	71	14.4	0.0313	F
37	16.2	-0.0725	F	72	15.2	-0.0401	F
38	21.0	-0.0589	F	73	13.4	0.0749	F
39	12.4	-0.0298	F	74	15.0	0.0087	F
40	15.0	0.0109	F	75	17.6	-0.0492	F
41	13.6	-0.0266	F	76	14.4	-0.0144	F
42	16.4	0.0115	F	77	41.6	0.6993	S
43	17.0	-0.0026	F	78	15.2	0.0883	F
44	47.0	0.6610	S	79	15.6	-0.0033	F
45	15.4	0.0350	F	80	13.2	-0.0917	F
46	21.0	0.0203	F	81	15.4	0.0212	F
47	14.4	-0.0214	F	82	14.4	0.0139	F
48	14.2	-0.0124	F	83	13.4	-0.0182	F
49	13.8	-0.0233	F	84	19.6	0.0905	F
50	48.4	0.6706	S	85	18.6	-0.0086	F
51	14.0	-0.0539	F	86	18.0	-0.0092	F
52	11.6	-0.0513	F	87	14.8	-0.0798	F
53	48.2	0.6869	S	88	11.8	-0.0763	F
54	15.4	-0.0528	F	89	13.2	-0.0100	F
55	16.4	-0.0549	F	90	16.0	-0.0359	F
56	15.0	-0.0196	F	91	11.6	0.0204	F
57	15.4	0.0293	F	92	13.8	-0.0432	F
58	17.6	-0.0372	F	93	13.4	-0.0671	F
59	18.8	-0.0360	F	94	11.6	0.0234	F
60	17.6	-0.0207	F	95	14.8	-0.0438	F
61	10.8	-0.0292	F	96	43.4	0.6783	S
62	13.0	-0.0282	F	97	16.0	-0.0617	F
63	15.8	0.0114	F	98	18.2	-0.0599	F
64	11.0	0.0103	F	99	42.4	0.7246	S
65	14.0	-0.2519	F	100	12.0	-0.0313	F

I THREE-MODE ANALYSIS IN C_7

The modes are $m = 1, 2, 3$, with ideal phases given by Equation 1. Figure 6 energy shares are the pooled endpoint calibration fractions used by the functional pipeline, renormalized per run. Its N_{eff}^{E} is computed from those same fractions. This is intentionally distinct from substituting the positionwise C_5 representation summary.

The directed graph uses independent 4,000-example fit, donor, and evaluation pools with the same seed convention as Appendix H, at position 100. For each of six source–target pairs, phase dependence is the donor-minus-random-residual progress in the target MLP-write coordinates. Removal dependence is the reduction in nearest-centroid accuracy of the target write code after removing the source component, conditioned on originally correct task predictions. This accuracy-drop quantity differs from the fixed-plane energy loss used in Figure 4.

For target m and edge assay a, the plotted incoming dependence is

$$d_m^a = \frac{1}{2} \sum_{j \neq m} \max(d_{j \to m}^a, 0), \qquad u_m^{\text{rem}} = p_m - q_m^{\text{rem}}, \quad u_m^{\text{rot}} = p_m - q_m^{\text{rot}}. \tag{18}$$

Both incoming edges contribute, with positive clipping *before* averaging. Within a run, $\sum_m u_m^a = 0$ by construction, adding dependence among the three observations. Figure 6e compares d^{rem} with u^{rem}; (f) compares phase dependence with u^{rot}. The reported Pearson/Spearman values are recomputed from the 90 regularized mode rows. The maximum incoming edge is a secondary aggregation already present in the source table; its correlations are listed in Table 7. This is sensitivity to aggregation of existing measurements, not another independent model cohort.

Table 7: Descriptive correlations for mean-positive incoming dependence (primary) and maximum-positive dependence (secondary), with allocation shortfall. All values use 90 modes nested in 30 regularized runs.

Assay	r_{mean}	ρ_{mean}	$r_{\max}$	$\rho_{\max}$
removal	0.704	0.532	0.672	0.505
phase	0.802	0.441	0.783	0.407

Table 8: All C_7 runs: stable-hit steps in thousands, effective modes, and regularized mode energy fractions. > 200 indicates right censoring at the nominal 200k budget (last recorded step 199,999), not an observed event.

Seed	k_B	k_R	N_B^E	N_R^E	$p_{1,R}$	$p_{2,R}$	$p_{3,R}$
1	76.6	25.2	1.192	2.999	0.336	0.337	0.327
2	65.8	23.4	1.218	2.997	0.345	0.319	0.336
3	25.4	18.8	2.095	2.997	0.345	0.319	0.336
4	> 200	15.4	1.044	2.987	0.337	0.305	0.358
5	155.8	21.0	1.079	2.986	0.324	0.311	0.365
6	98.0	19.8	1.122	2.994	0.323	0.355	0.322
7	87.0	43.6	1.156	2.968	0.293	0.378	0.329
8	20.4	14.4	2.330	2.994	0.350	0.334	0.315
9	75.4	18.0	1.205	2.948	0.392	0.322	0.285
10	92.0	17.8	1.178	2.964	0.301	0.384	0.315
11	16.0	22.0	2.703	2.967	0.316	0.382	0.301
12	25.6	16.8	2.022	2.999	0.335	0.327	0.338
13	84.4	22.8	1.211	2.978	0.320	0.373	0.307
14	24.8	17.0	2.360	2.999	0.326	0.338	0.336
15	18.6	21.2	2.377	2.992	0.357	0.316	0.326
16	116.4	18.6	1.130	2.969	0.320	0.300	0.380
17	73.2	16.4	1.154	2.998	0.325	0.331	0.344
18	21.6	15.6	2.088	2.997	0.327	0.326	0.347
19	68.2	26.0	1.092	2.980	0.307	0.371	0.322
20	20.0	14.8	2.050	2.997	0.338	0.318	0.344

Seed	k_B	k_R	N_B^E	N_R^E	$p_{1,R}$	$p_{2,R}$	$p_{3,R}$
21	22.0	21.8	2.295	2.968	0.380	0.321	0.299
22	20.6	17.6	2.049	2.973	0.377	0.303	0.320
23	25.2	14.0	2.069	2.996	0.317	0.339	0.344
24	21.8	20.8	2.057	2.972	0.333	0.294	0.373
25	21.6	22.0	2.025	2.996	0.325	0.351	0.324
26	22.4	16.4	2.035	2.970	0.313	0.306	0.381
27	24.4	18.8	2.030	2.986	0.364	0.310	0.326
28	20.8	17.2	2.014	2.998	0.345	0.330	0.325
29	43.2	13.2	1.756	2.993	0.327	0.317	0.356
30	25.0	19.6	1.996	2.965	0.384	0.314	0.302

J STATISTICAL REPORTING AND SOURCE-TABLE TRACEABILITY

The experimental unit is an independently trained seed/run. The main text reports medians, ranges, paired counts, and descriptive correlations. Figure 4's thick curves and connecting lines are means, whereas numerical summaries in the prose are explicitly medians. No test treats token count, checkpoint count, or mode count as the number of independent models. Figure 6 correlation bands come from 2,000 bootstrap samples of 30 complete runs with replacement, carrying all three mode rows together, fitting a line in each sample, and taking pointwise 2.5th/97.5th percentiles. Random seeds are 6060 and 6061 for the two panels. The band describes fitted-mean uncertainty, not a prediction interval for a new mode.

We do not report naive 90-observation correlation p-values. The held-out convergence correspondence is reported descriptively. There is no new multiple-comparison-adjusted significance claim across the six figures. Converged-only C_7 baseline medians are labeled as such and do not estimate an uncensored population mean.

Table 9: Numerical traceability for the figures; methods and run-level results appear in the preceding appendices.

Figure	Archived measurements
1	Baseline endpoint energy fractions and stable-hit steps
2	Seed-matched baseline/regularized convergence and energy fractions
3	Mode-wise energy, causal share, recovery, and isolated rotation
4	Layerwise decoding, pre/post-MLP energy, bidirectional effects, and source/rank alignment
5	Discovery and held-out provenance scores, convergence labels, and directed effects
6	Cohort convergence, mode allocation, all directed edges, and correlation summaries

The Appendix tables report the run-level values used for the central comparisons.

K REPRODUCIBILITY DETAILS

Document and figure reproduction. This PDF uses the official ICLR 2027 style and contains all six figures, references, numerical summaries, and appendices. The procedures and numerical definitions needed to interpret the figures are described above.

Scientific implementations and software. The experiments used separate baseline and energy-regularized training implementations, together with functional-assay, layerwise-decoding, final-MLP, and discovery/held-out provenance implementations. Repeating the reported interventions requires saved model checkpoints and relevant intermediate inputs. Dependencies include NumPy, pandas, PyTorch, Matplotlib, seaborn, scikit-learn, and SciPy; historical versions are unpinned. A saved regularized C_5 training output identifies an NVIDIA A100-SXM4-40GB GPU, without establishing hardware for other cohorts. GPU hours and peak memory are not inferred from update counts; cross-platform reproduction need not be bitwise identical.