A SMALL-WORLD PERSPECTIVE ON CAPABILITY EMERGENCE IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As model scale increases, large language models may exhibit abrupt performance gains on certain tasks—gains that are difficult to predict by extrapolating performance trends observed in smaller models. This phenomenon is commonly referred to as "emergence". To explain emergence, existing work has investigated it from perspectives such as tasks, evaluation metrics, and model scale; other work has further traced the internal changes within Transformers, for example by tracking the paths of information propagation under a specific input. However, one question remains unanswered: does the Transformer undergo a holistic qualitative change, rather than a single or local change, that gives rise to the dramatic leap in LLM performance? Inspired by the view that "Language Modeling Is Compression", we propose that model capability should be reflected in whether highly compressed information under a given input can be rapidly reached (finding content seen during pretraining) and rapidly composed (combining multiple related pieces of content). To demonstrate this hypothesis, we borrow the famous six degrees of reachability theorem from small-world theory—namely, that any two nodes in a graph are reachable within six steps, to characterize the internal structural changes of Transformers. Specifically, we first model the Transformer as a directed graph based on the information propagation mechanism of attention; we then extend small-world theory to the directed-graph setting, define metrics that characterize the small-world properties of the Transformer, and describe the quantitative relationship between reachability/composability and these metrics; finally, across different tasks and parameter scales, we demonstrate that changes in the small-world properties of the Transformer's internal organization are highly consistent with the phenomenon of emergence. This is a creative characterization of emergence and may offer inspiration for unlocking greater potential in LLMs. For reproducibility, the code used for the experiments and analyses is available in an anonymous repository: `https://anonymous.4open.science/r/iclr2027-A12C/`.

1 INTRODUCTION

Understanding capability emergence is important for anticipating what larger language models will be able to do. Scaling laws describe relatively smooth improvements in language-modeling loss with model size, training data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022). Emergent abilities concern a different observation: some downstream abilities appear absent at smaller scales but become observable at larger scales, with sharp gains under the chosen evaluation (Wei et al., 2022). These transitions motivate examining how improvements in language prediction relate to the development of usable task capabilities.

Research has examined emergence through measurement, generalization, and computational mechanisms. Schaeffer et al. (2023) show that nonlinear or discontinuous evaluation metrics can make gradual underlying improvements appear abrupt. Gong et al. (2025) explain in-context learning through generalization across sequences and topics in autoregressive pretraining. At the component level, Aoyama & Wilcox (2025) associate the formation of specialized attention heads with a training phase transition and changes in alignment with human reading behavior. More recently, Lavie et al. (2026) derive a Bayesian account of abrupt copy-head emergence in a controlled attention

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.

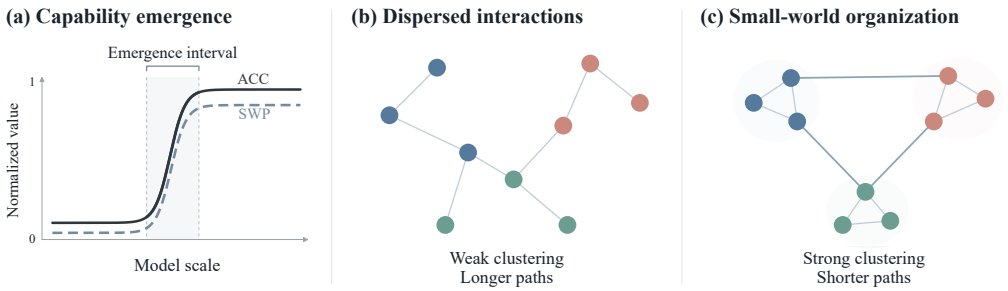


Figure 1: **A structural perspective on capability emergence. (a)** Schematic accuracy (ACC) and Small-World Propensity (SWP) curves illustrate their hypothesized association across model scales, with shading marking the emergence interval. **(b, c)** Conceptual graphs contrast dispersed interactions with weak local clustering and longer paths against small-world organization with stronger local clustering and shorter paths. All panels are schematic and do not imply causality or a universal ACC–SWP relationship.

model. These studies offer complementary explanations, but leave open how capability gains relate to functional interactions among components across model scales.

The established connection between language prediction and data compression offers a starting point (Delétang et al., 2024). Knowledge-capacity studies further demonstrate that Transformers can encode substantial factual information within their parameters under controlled training conditions (Allen-Zhu & Li, 2025). Scaling studies further show that model quality can continue to improve at training-token-to-parameter ratios far beyond conventional compute-optimal regimes, reaching up to 10,000 tokens per parameter in controlled experiments (Sardana et al., 2024). These findings motivate an important distinction between information that is encoded during training and information that can be effectively accessed and composed at inference time. This distinction is reflected in several observed limitations: models may memorize factual text without reliably extracting its content through question answering (Allen-Zhu & Li, 2024), or learn a relation without generalizing to its reverse (Berglund et al., 2024). Recent work also distinguishes representations available for flexible reasoning and reporting from other internal representations (Gurnee et al., 2026).

We therefore hypothesize that capability gains may depend not only on what information is encoded in model parameters, but also on how computational components interact to make that information accessible and composable in response to a prompt. Building on circuit analyses of Transformer computation (Conmy et al., 2023), we investigate the functional organization of these interactions.

"Six degrees of separation" illustrates how short chains can connect individuals in a social network (Travers & Milgram, 1969). Small-world theory combines short global paths with strong local clustering (Watts & Strogatz, 1998), providing a structural perspective on local aggregation and connectivity. These properties offer candidate structural correlates of organization supporting access to and composition of information, rather than direct measures of knowledge retrieval. Figure 1 illustrates this connection.

Applying this perspective to Transformer computation requires a functional representation. Although weighted small-world measures already exist (Muldoon et al., 2016), their application must account for layered, directional computation and distinguish connectivity in an undirected projection from actual information-flow routes.

Our central empirical question is whether intervals of abrupt performance improvement coincide with pronounced increases in small-world organization within the model's functional topology. We introduce directed, weighted functional graphs of attention-head interactions grounded in the Transformer's layered architecture. Their undirected, weighted projections enable a reference-based assessment using Small-World Propensity (SWP), relative to matched lattice and random networks (Muldoon et al., 2016).

Our contributions are threefold.

- **An information-flow perspective on capability emergence.** We argue that capability emergence depends not only on knowledge storage, but also on efficient access to and composition of information, and propose small-world organization as a candidate structural signature of this computational accessibility.

- **An extension of small-world analysis to Transformer computation.** We construct directed weighted graphs and use their topology to characterize changes in internal computational organization, enabling network-level structural analysis.

- **Behavioral–structural correspondence and functional relevance.** Across model families and tasks, we observe that several intervals of rapid capability improvement coincide with stronger small-world organization; intervention experiments further support the functional relevance of the identified components.

2 Related Work

Scaling and capability emergence. Empirical scaling laws describe how language-modeling loss varies with model size, training data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022). At the task level, studies of emergent abilities report rapid performance gains across model scales (Wei et al., 2022), while broken scaling laws model changes in scaling behavior through transitions between power-law regimes (Caballero et al., 2023). The interpretation of abrupt gains remains contested: Schaeffer et al. (2023) show that discrete evaluation metrics can produce apparently emergent behavior from gradual underlying improvements. These studies motivate distinguishing observed performance transitions from their explanations. Our work adds a structural perspective by examining whether behavioral changes are associated with changes in the organization of interactions within the model.

Mechanistic interpretability and circuit development. Mechanistic studies connect model behavior to specific internal computations. Olsson et al. (2022) associate the formation of induction heads with a sharp improvement in in-context learning during training, providing causal evidence in small attention-only models and correlational evidence in larger models. Task-vector and function-vector studies further show that internal representations can encode task information and causally influence task execution (Hendel et al., 2023; Todd et al., 2024). Across training and model scale, Tigges et al. (2024) examine component emergence, algorithmic consistency, circuit size, and node-set stability. Their results establish that stable algorithmic roles can coexist with changing circuit constituents. Our work complements these component-level analyses by examining how the joint topological organization of functional interactions changes across model scales.

Graph representations of Transformers. Graph-based analyses differ in both their units of representation and the meaning of their connections. Attention rollout and attention flow aggregate cross-layer attention to estimate input-token relevance (Abnar & Zuidema, 2020). The Transformer circuits framework decomposes attention into interpretable operations and studies how heads compose through the residual stream (Elhage et al., 2021). Task-circuit studies then identify and validate computational subgraphs, including the indirect object identification circuit (Wang et al., 2023). Automated circuit discovery uses interventions to identify connections that preserve task behavior (Conmy et al., 2023). These approaches provide foundations for representing functional dependencies among model components. Our analysis focuses on head interactions and compares their organization across scales, linking component-level analysis to network-level capability changes.

Small-world organization in weighted networks. Watts & Strogatz (1998) demonstrate that introducing shortcuts into a locally connected network can sharply reduce characteristic path length while retaining high clustering. Muldoon et al. (2016) introduce Small-World Propensity (SWP), which jointly evaluates clustering and path length relative to lattice and random references, including for weighted networks. This joint characterization is relevant to our hypothesis because clustering alone does not capture connectivity between distant component groups, while short paths alone do not characterize local organization. We adopt SWP as an established measure to examine whether changes in small-world organization, relative to these references, coincide with behavioral improvements across model scales.

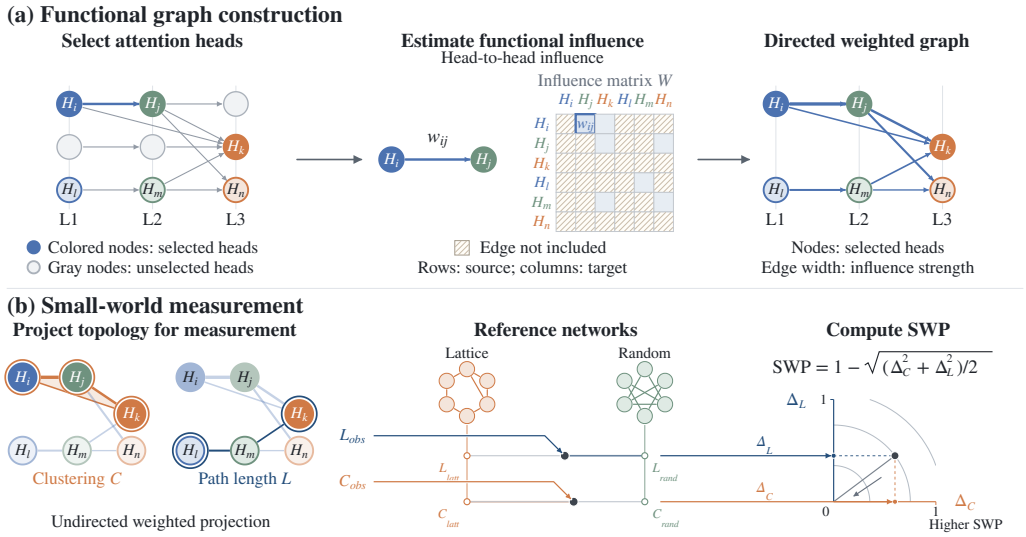


Figure 2: **From functional graphs to small-world measurement. (a)** Selected attention heads and their task-conditioned influences form a directed weighted functional graph; edge widths indicate influence strength. **(b)** Clustering C_{obs} and characteristic path length L_{obs} of the corresponding undirected weighted graph are normalized against lattice and random references to obtain Δ_C and Δ_L, which jointly determine Small-World Propensity (SWP). Higher SWP indicates smaller combined deviations. All graphs are schematic. SWP characterizes the undirected weighted projection, not directed information propagation.

3 Functional Graph Construction and Small-World Analysis

Our analysis proceeds in two stages, as illustrated in Figure 2. We first identify functionally salient attention heads and directed interactions among them, and estimate the strength of these interactions to obtain a directed, weighted functional graph (Figure 2a). We then quantify its small-world organization using Small-World Propensity (SWP; Muldoon et al., 2016) on an undirected weighted projection (Figure 2b).

3.1 Constructing Directed Weighted Graphs

To characterize how distributed internal computation is functionally organized, we extract stable functional interactions among Transformer attention heads and organize them into a directed weighted topology:

$$G^{\rightarrow} = (V, E, W). \tag{1}$$

The construction extracts stable functional structure in three stages: identifying functionally salient heads, selecting stable directed interactions among them, and quantifying their influence strengths on held-out examples.

Here, V consists of the selected attention heads, and E contains directed edges representing functional influences between heads. An edge $i \to j$ represents the functional influence of source head i on target head j. The non-negative weighted adjacency matrix W quantifies the strength of these influences, with $W_{ij} = 0$ for $(i, j) \notin E$.

To obtain a stable topology while separating structure discovery from functional evaluation, we partition the samples into two disjoint subsets: a discovery set and a held-out set. The discovery set is used to recover a stable interaction topology by identifying functionally salient nodes and stable directed interactions, whereas the held-out set is used to quantify their functional influence on unseen examples and derive the corresponding edge weights. The node and edge sets are fixed before held-out evaluation.

3.1.1 NODE SELECTION

We identify attention heads with stable attribution-based importance across discovery subsamples. In run r, we sample $P\%$ of the discovery examples, forming S_r^h, and score each head by its mean absolute attribution with respect to the answer margin, which measures the model's preference for the correct answer over the strongest incorrect alternative:

$$a_h^{(r)} = \frac{1}{|S_r^h|} \sum_{x \in S_r^h} \left| \mathrm{Attr}_h(x) \right|. \tag{2}$$

We retain the highest-scoring fraction ρ_h of heads to form V_r. Repeating this procedure N times, we define

$$c_h = \sum_{r=1}^{N} \mathbf{1}\{h \in V_r\}, \qquad V = \{h \in \mathcal{H}_\theta : c_h > K\}. \tag{3}$$

Thus, V retains heads whose attribution-based importance is stable across discovery subsamples. We then evaluate the selected heads on held-out examples without changing V. Repeated subsampling, attribution scoring, and selection settings are detailed in Appendix A.2.

3.1.2 EDGE SELECTION AND EDGE-WEIGHT ESTIMATION

After identifying stable nodes, we next recover the directed interaction structure among them. Candidate edges are constrained by the Transformer layer order:

$$E_{\mathrm{cand}} = \{(i,j) \in V \times V : \ell_i < \ell_j\}, \tag{4}$$

where ℓ_i denotes the layer containing head i. Thus, candidate edges connect heads in earlier layers to heads in later layers.

On each discovery subsample, candidate edges are ranked using Pair-EAP (Syed et al., 2024), and the highest-scoring fraction forms E_r. After N runs, we retain edges selected more than K times:

$$E = \left\{ (i,j) \in E_{\mathrm{cand}} : \sum_{r=1}^{N} \mathbf{1}\{(i,j) \in E_r\} > K \right\}. \tag{5}$$

Pair-EAP is used only to identify stable interaction structure; the final edge weights are estimated separately from held-out intervention effects. We then fix E and estimate edge weights on held-out examples. Let $m_{ij}^{\mathrm{int}}(x)$ and $m_{ij}^{\mathrm{base}}(x)$ denote the answer margins under the connection intervention and its corresponding baseline, respectively. We define

$$W_{ij} = \frac{1}{|\mathcal{D}_{\mathrm{held}}|} \sum_{x \in \mathcal{D}_{\mathrm{held}}} \left| m_{ij}^{\mathrm{int}}(x) - m_{ij}^{\mathrm{base}}(x) \right|, \qquad (i,j) \in E, \tag{6}$$

with $W_{ij} = 0$ otherwise. Thus, W_{ij} quantifies the magnitude of functional influence associated with an edge, rather than its sign. The resulting graph provides a weighted topology of stable functional interactions, forming the basis for the network-level analysis that follows.

Pair-EAP implementation and intervention details are provided in Appendix A.3, and the complete graph-construction procedure is provided in Appendix A.4 (Algorithm 1).

3.2 QUANTIFYING SMALL-WORLD ORGANIZATION

Having derived the directed weighted graph, we next characterize its network-level organization through local clustering and global path structure. Small-world organization jointly captures these properties by combining strong local clustering with short global paths. Local clustering describes how closely a node's neighbors are interconnected, whereas characteristic path length measures the average shortest-path distance between node pairs. We jointly quantify these properties using Small-World Propensity (SWP) (Muldoon et al., 2016), computed on the undirected weighted graph corresponding to each directed weighted graph. We assess clustering relative to a lattice reference and path length relative to a random reference, normalize the deviations using statistics from both references, and combine them into a single score. Higher SWP indicates stronger small-world organization relative to these references.

3.2.1 Reference Network Construction

For each observed graph, we construct ensembles of lattice and random reference networks matched to its number of nodes, number of edges, and empirical edge-weight distribution. The lattice references provide a baseline for strong local clustering, while the random references provide a baseline for short global paths. We average the graph statistics across realizations within each ensemble. Let C_{latt} and L_{latt} denote the mean weighted clustering coefficient and weighted characteristic path length of the lattice ensemble, and let C_{rand} and L_{rand} denote the corresponding means for the random ensemble. Reference-generation procedures and parameter settings are provided in Appendix B.

3.2.2 Normalized Deviations and SWP Computation

Let C_{obs} and L_{obs} denote the weighted clustering coefficient and weighted characteristic path length of the observed undirected weighted graph. Path lengths are computed by converting positive edge weights into costs, with stronger weights corresponding to lower costs. For disconnected graphs, path length is evaluated within the largest connected component. The conversion to an undirected graph, the weighted clustering definition, and the path-cost calculation are detailed in Appendix B.

Following the SWP formulation, we define the normalized clustering and path-length deviations as

$$\begin{aligned} \Delta_C &= \mathrm{clip}_{[0,1]}\left(\frac{C_{\mathrm{latt}} - C_{\mathrm{obs}}}{C_{\mathrm{latt}} - C_{\mathrm{rand}}}\right), \\ \Delta_L &= \mathrm{clip}_{[0,1]}\left(\frac{L_{\mathrm{obs}} - L_{\mathrm{rand}}}{L_{\mathrm{latt}} - L_{\mathrm{rand}}}\right), \end{aligned} \tag{7}$$

where $\mathrm{clip}_{[0,1]}$ restricts each value to $[0, 1]$. These deviations are combined as

$$\mathrm{SWP} = 1 - \sqrt{\frac{\Delta_C^2 + \Delta_L^2}{2}}, \qquad \mathrm{SWP} \in [0, 1]. \tag{8}$$

Higher SWP indicates smaller combined deviations from the reference clustering and path-length levels and characterizes the organization of the undirected weighted graph.

4 Experiments: Behavioral Transitions and Small-World Organization

4.1 Experimental Setup

Models and tasks. We evaluate 21 checkpoints from four model families: Pythia (Biderman et al., 2023), Qwen2.5 (Yang et al., 2024), Qwen3 (Yang et al., 2025), and Gemma-3 (Gemma Team et al., 2025). Pythia is evaluated on Control and Overall arithmetic; Qwen3 on Plus-one arithmetic, Logical Fallacies, and Machine Learning; Qwen2.5 on Plus-one arithmetic and logical deduction; and Gemma-3 on High-school Geography, World Religions, Logical Deduction, and Temporal Sequences. Detailed experimental settings are provided in Appendix C.

Behavior and topology. We compare task accuracy with the SWP of functional graphs using the methods in Sections 3.1 and 3.2. Continuous answer margins provide a complementary behavioral measure and are compared only within matched model, task, and prompting settings. Experimental details are provided in Appendix C, with additional accuracy–margin results in Appendix E.

Evaluation design. We assess the proposed behavioral–structural correspondence across model families, tasks, prompting conditions, and scales through complementary analyses. Within matched settings, accuracy–SWP trajectories show whether capability gains coincide with stronger small-world organization, as measured by SWP. Continuous answer margins test whether the behavioral changes are also visible beyond a discrete accuracy measure. A layer-order-preserving random-edge null model tests whether layer order alone can explain the observed SWP, while interventions assess the functional relevance of selected heads and component groups.

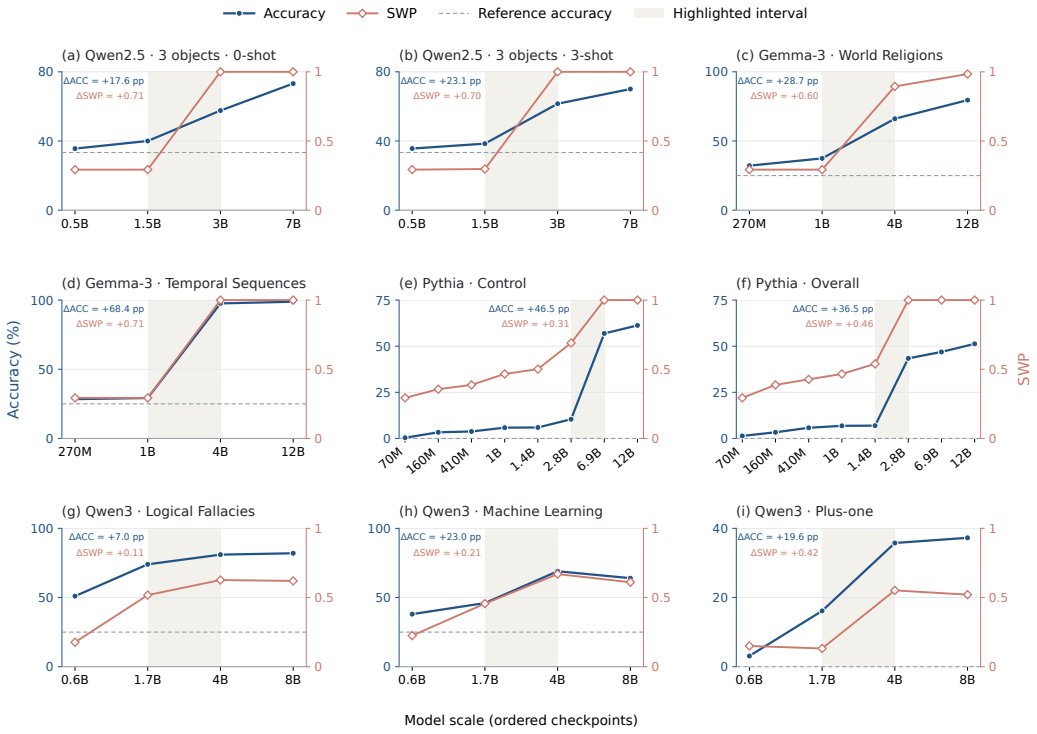


Figure 3: **Behavioral and structural changes across model families and tasks.** Blue curves show accuracy (left axes), and red curves show Small-World Propensity (SWP; right axes). Shading marks the highlighted adjacent-scale intervals, with the corresponding changes in accuracy and SWP reported in each panel. Panels show representative settings from Qwen2.5, Gemma-3, Pythia, and Qwen3. Dashed lines denote chance accuracy for multiple-choice tasks and a zero-accuracy reference for arithmetic. Model scales are shown as ordered checkpoints, and accuracy-axis ranges vary across panels. Complete ACC–SWP trajectories are reported in Appendix D.

4.2 Behavioral and Topological Transitions Across Scale

We examine whether rapid behavioral improvements coincide with changes in small-world organization. Figure 3 compares accuracy–SWP trajectories across four model families. Comparisons are made within each family, task, and prompting condition; additional results appear in Appendix D.

Alignment across model scales. Several settings show pronounced concurrent increases in accuracy and SWP over the highlighted intervals. In Qwen2.5 logical deduction, substantial SWP increases accompany clear accuracy gains under both zero-shot and three-shot prompting. Strong joint changes are also observed for Gemma-3 World Religions and Temporal Sequences, as well as Pythia Control and Overall arithmetic. Qwen3 Plus-one provides an additional example of aligned behavioral and structural improvement. Together, these results provide cross-family evidence that capability improvements can coincide with stronger small-world organization.

Similar structural changes, different behavioral gains. Panels (a) and (b) of Figure 3 show representative Qwen2.5 logical-deduction settings. The full set of six object-count–prompting combinations is reported in Appendix D.2 (Figures 7 and 8). Between 1.5B and 3B, SWP increases by approximately 0.70 in each combination, while accuracy gains range from 10.4 to 23.1 percentage points. Thus, similar increases in small-world organization, as measured by SWP, accompany behavioral gains of different magnitudes.

Continuous behavioral evidence. Schaeffer et al. (2023) argue that apparent emergence can arise from nonlinear or discontinuous evaluation metrics and may become smooth under continuous alternatives. However, subsequent work shows that emergent behavior can persist under continuous

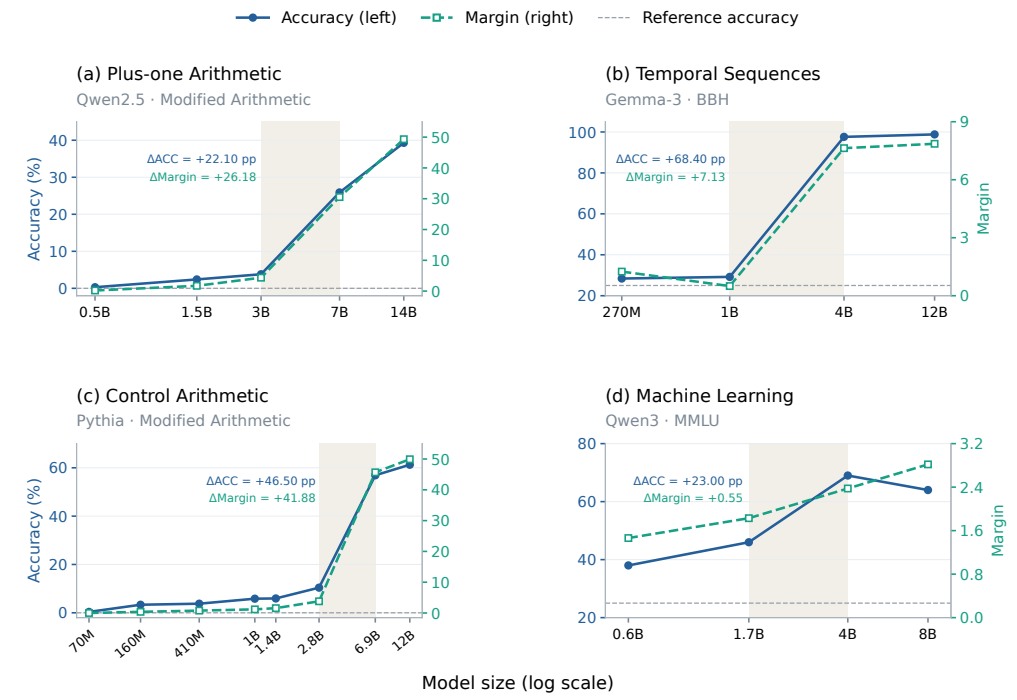


Figure 4: **Representative accuracy–margin trajectories across model families.** Shading marks the highlighted scale intervals, with the corresponding ΔACC (pp) and ΔMargin annotated in each panel. Dashed lines indicate reference accuracy. Additional results appear in Appendix E.

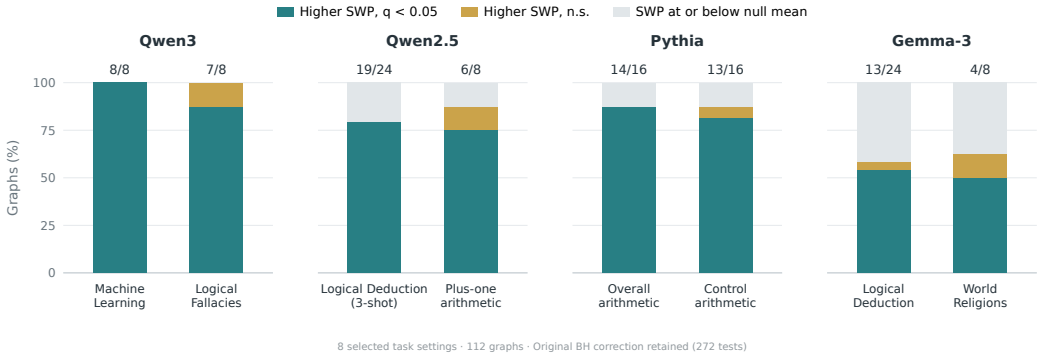


Figure 5: **Functional topology relative to a layer-order-preserving random-edge null model.** Each observed graph is compared with 5,000 random-edge nulls constrained by the same Transformer layer order. Bars summarize eight representative task settings (112 graphs); 84/112 (75.00%) have SWP above the null mean with BH-adjusted $q < 0.05$. The original BH correction over 272 tests is retained, with complete results reported in Appendix G.

probability-based measures (Snell et al., 2024). Consistent with this observation, our representative settings show that the highlighted accuracy gains are accompanied by increases in continuous answer margin (Figure 4), indicating that the behavioral changes are also reflected beyond the discrete accuracy metric. Additional results are reported in Appendix E.

Comparison with a layer-ordered null model. Against a layer-order-preserving random-edge null model, 112 of 160 graphs (70.0%) across all 12 task settings have SWP above the null mean with BH-adjusted $q < 0.05$ (Appendix G; representative results in Figure 5). This suggests that layer order alone does not explain the observed small-world organization.

Table 1: **Intervention evidence across tasks.** Entries report supported/evaluated units. The Heads column counts model–task–prompt combinations satisfying the Selected–Random criterion in both intervention directions and both splits. The Hub and Shortcut columns count individual model–split–direction comparisons under their respective protocols. Deduction rows aggregate the 3-, 5-, and 7-object settings. Full results appear in Appendix F.

Family	Task	Heads	Hub	Shortcut
Pythia	Overall arithmetic	6/8	23/28	28/32
	Control arithmetic	6/8	21/28	26/32
Qwen3	Plus-one arithmetic	4/4	11/16	14/16
	Logical Fallacies	4/4	12/16	14/16
	Machine Learning	4/4	13/16	12/16
Qwen2.5	Plus-one arithmetic	4/4	14/16	15/16
	Deduction: 3, 5, 7 objects (0-shot)	5/12	40/48	41/48
	Deduction: 3, 5, 7 objects (3-shot)	7/12	35/48	38/48
Gemma-3	HS Geography	3/4	11/16	14/16
	World Religions	3/4	13/16	11/16
	Logical Deduction	4/4	15/16	14/16
	Temporal Sequences	4/4	14/16	15/16
Total		**54/72 (75.00%)**	**222/280 (79.29%)**	**242/288 (84.03%)**

Convergent evidence across analyses. In several highlighted scale intervals, gains in accuracy and continuous answer margins coincide with stronger small-world organization, supporting a behavioral–structural correspondence. Complementary interventions (Section 4.3) further support the functional relevance of the identified components.

4.3 Functional Evidence from Interventions

Interventions separately assess whether components identified in the functional-graph analysis are relevant to task behavior. We test discovery-selected heads through held-out disruption and restoration, and hub- and shortcut-associated groups under separate protocols. Table 1 summarizes support; its columns use different counting units.

Selected attention heads. Head sets are fixed using discovery data. Of 72 reported model–task–prompt combinations, 54 meet the Selected–Random criterion with positive contrasts and global BH-adjusted $q < 0.05$ in both intervention directions and splits. The correction covers all 928 comparisons; detailed effects appear in Appendix F.3.

Hub- and shortcut-associated components. Across 16 task–prompt settings, support is found in 222/280 hub and 242/288 shortcut comparisons. These are individual model–split–direction comparisons and support the functional relevance of the identified groups. The interventions do not isolate the causal contribution of small-world topology. Protocols and task-level results appear in Appendix F.4.

5 Conclusion

This work investigates changes in the internal organization of computation during capability emergence in language models from a network-topological perspective. We represent functional interactions within Transformers as directed weighted graphs and characterize their topology with the aid of Small-World Propensity (SWP). Our results show that rapid capability improvements are often accompanied by stronger small-world organization, suggesting that the reorganization of internal interactions may help make information more accessible and composable. Intervention experiments further support the functional relevance of the identified components and structural features. Overall, this work extends the study of capability emergence from individual components to the global organization of interactions among components, providing a new structural perspective on how models aggregate, access, and compose internal information.

AI USE STATEMENT

We used generative AI tools only for language polishing. All AI-assisted edits were manually checked and reviewed by the authors, who take full responsibility for the final content of the manuscript.

ETHICS STATEMENT

The authors have reviewed the ICLR Code of Ethics and believe that this work does not raise any ethical concerns. This study does not involve human subjects, personal data, or privacy-sensitive information. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the code used for the experiments and analyses reported in this paper in an anonymous repository at `https://anonymous.4open.science/r/iclr2027-A12C/`.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1067–1077. PMLR, 2024. URL https://proceedings.mlr.press/v235/allen-zhu24a.html.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/26d3c9a66836ded8f34a944f2bfe868e-Abstract-Conference.html.

Tatsuya Aoyama and Ethan Wilcox. Language models grow less humanlike beyond phase transition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24938–24958. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1214.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Stickland, Tomek Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5178b2f2d7c44aa390c0777dc77b3f0c-Abstract-Conference.html.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sckjveqlCZ.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information*

Processing Systems, volume 36, pp. 16318–16352. Curran Associates, Inc., 2023. doi: 10.52202/075280-0719. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Paper-Conference.pdf.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jznbgiynus.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for Transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Gemma Team et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. doi: 10.48550/arXiv.2503.19786.

Zixuan Gong, Xiaolin Hu, Huayi Tang, and Yong Liu. Towards auto-regressive next-token prediction: In-context learning emerges from generalization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gK1rl98VRp.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. doi: 10.52202/068431-2176.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Itay Lavie, Kirsten Fischer, Andrey Lekov, Frederic Van Maele, Zohar Ringel, and Moritz Helias. Phase transitions in attention: A Bayesian theory of copy head emergence. *arXiv preprint arXiv:2606.12058*, 2026. URL https://arxiv.org/abs/2606.12058.

Sarah Feldt Muldoon, Eric W. Bridgeford, and Danielle S. Bassett. Small-world propensity and weighted brain networks. *Scientific Reports*, 6(1):22057, 2016. doi: 10.1038/srep22057.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond Chinchilla-optimal: Accounting for inference in language model scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43445–43460. PMLR, 2024. URL https://proceedings.mlr.press/v235/sardana24a.html.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pp. 55565–55581, 2023. doi: 10.52202/075280-2425. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/adc98a266f45005c403b8311ca7e8bd7-Abstract-Conference.html.

Charlie Victor Snell, Eric Wallace, Dan Klein, and Sergey Levine. Predicting emergent capabilities by finetuning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=vL8BIGuFTF.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.824.

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.

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, pp. 40699–40731, 2024. doi: 10.52202/079017-1287. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/47c7edadfee365b394b2a3bd416048da-Abstract-Conference.html.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969. doi: 10.2307/2786545. URL https://www.jstor.org/stable/2786545.

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

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393(6684):440–442, 1998. doi: 10.1038/30918. URL https://doi.org/10.1038/30918.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

A Graph Construction Details

A.1 Answer Scoring and Answer Margin

We evaluate attribution scores and intervention effects with respect to the answer margin. For an example x, let $\mathcal{A}(x)$ denote the set of candidate answers and let $a^* \in \mathcal{A}(x)$ denote the correct answer. For a candidate answer consisting of tokens $a = (a_1, \ldots, a_{T_a})$, its score is the length-normalized conditional log-probability:

$$s_\theta(a \mid x) = \frac{1}{T_a} \sum_{t=1}^{T_a} \log p_\theta(a_t \mid x, a_{<t}). \tag{9}$$

Here, T_a is the number of answer tokens, $a_{<t} = (a_1, \ldots, a_{t-1})$ is the preceding answer prefix, and $p_\theta(a_t \mid x, a_{<t})$ is the model's conditional probability of token a_t. Scores are computed by teacher forcing over each candidate answer, excluding prompt tokens from the sum.

The answer margin is the difference between the correct answer's score and the highest score among incorrect candidates:

$$m_\theta(x) = s_\theta(a^* \mid x) - \max_{a \in \mathcal{A}(x) \setminus \{a^*\}} s_\theta(a \mid x). \tag{10}$$

A positive margin indicates that the correct answer scores higher than every incorrect candidate. The baseline and intervened evaluations use the same candidate answers and scoring rule.

A.2 Repeated Subsampling and Node Selection

Let $\mathcal{D}_{\text{disc}}$ and $\mathcal{D}_{\text{held}}$ denote the discovery and held-out sets, respectively, with

$$\mathcal{D}_{\text{disc}} \cap \mathcal{D}_{\text{held}} = \varnothing. \tag{11}$$

Let $\mathcal{H}_\theta$ denote the set of attention heads in the model. In each run r, we draw a subset $S_r^h \subseteq \mathcal{D}_{\text{disc}}$ containing $P\%$ of the discovery examples. Sampling is without replacement within each run; subsamples may overlap across runs.

For head h, let $\text{Attr}_h(x)$ denote its attribution score with respect to the answer margin. Its score in run r is the mean absolute attribution:

$$a_h^{(r)} = \frac{1}{|S_r^h|} \sum_{x \in S_r^h} \left| \text{Attr}_h(x) \right|. \tag{12}$$

We retain the highest-scoring fraction ρ_h of the heads to obtain V_r. After N runs, the final node set is

$$c_h = \sum_{r=1}^{N} \mathbf{1}\{h \in V_r\}, \qquad V = \{h \in \mathcal{H}_\theta : c_h > K\}. \tag{13}$$

Here, $\mathbf{1}\{\cdot\}$ is the indicator function. The threshold is strict: a head must be selected more than K times, where $0 \le K < N$.

Selection settings. The example-subsampling percentage P is distinct from the head- and edge-selection fractions ρ_h and ρ_e. The node and edge procedures use the same P, N, and K, while ρ_h and ρ_e control their respective per-run selection sizes.

A.3 Edge Selection and Held-Out Weight Estimation

Candidate edges. After fixing V, we form the candidate edge set

$$E_{\text{cand}} = \{(i, j) \in V \times V : \ell_i < \ell_j\}, \tag{14}$$

where ℓ_i and ℓ_j are the Transformer layers containing heads i and j, respectively. This permits connections between both adjacent and non-adjacent layers and excludes same-layer and backward connections.

Pair-EAP scoring and stable edge selection. We use a receiver-specific implementation of EAP (Syed et al., 2024), termed Pair-EAP here, to score the source head's contribution to a specified target head. Let $\mathrm{PairEAP}_{i\to j}(x)$ denote the resulting edge attribution with respect to the answer margin.

In run r, we draw a discovery subsample S_r^e using the same subsampling rule as for node selection. The score of candidate edge (i, j) is

$$b_{ij}^{(r)} = \frac{1}{|S_r^e|} \sum_{x \in S_r^e} \left| \mathrm{PairEAP}_{i\to j}(x) \right|. \tag{15}$$

We retain the highest-scoring fraction ρ_e of candidate edges to form E_r. The final edge set is

$$c_{ij} = \sum_{r=1}^{N} \mathbf{1}\{(i, j) \in E_r\}, \qquad E = \{(i, j) \in E_{\mathrm{cand}} : c_{ij} > K\}. \tag{16}$$

Held-out evaluation. We fix both V and E before evaluation on $\mathcal{D}_{\mathrm{held}}$. Head interventions assess the functional effects of selected heads on examples not used during selection. Connection interventions assess the effects of selected edges and provide the measurements used for weighting. Neither evaluation changes membership in V or E.

Edge-weight calculation. For an edge $i \to j$ and held-out example x, let $m_{ij}^{\mathrm{int}}(x)$ denote the answer margin under the connection intervention, and let $m_{ij}^{\mathrm{base}}(x)$ denote its corresponding baseline margin. Both are computed using the scoring rule in Appendix A.1. The intervention effect is

$$\Delta m_{ij}(x) = m_{ij}^{\mathrm{int}}(x) - m_{ij}^{\mathrm{base}}(x). \tag{17}$$

We define the edge weight as the mean absolute intervention effect:

$$w_{ij} = \frac{1}{|\mathcal{D}_{\mathrm{held}}|} \sum_{x \in \mathcal{D}_{\mathrm{held}}} \left| \Delta m_{ij}(x) \right|. \tag{18}$$

Taking absolute values before averaging prevents positive and negative effects across examples from canceling. Thus, weights quantify influence magnitude without encoding its sign. Within the same graph and under the same intervention protocol, larger weights indicate stronger average functional effects. Finally, $W_{ij} = w_{ij}$ for $(i, j) \in E$ and $W_{ij} = 0$ otherwise.

A.4 Complete Construction Algorithm

Algorithm 1 provides the full construction procedure. $\mathrm{Subsample}(D, q)$ samples a fraction q of the examples without replacement within each run; subsamples may overlap across runs. $\mathrm{TopFraction}(U, s, \rho)$ selects the highest-scoring fraction ρ of elements in U according to scores s. The attribution methods, intervention protocols, and selection settings are specified in the preceding subsections.

Algorithm 1 Detailed Procedure for Directed Weighted Graph Construction

Require: Transformer f_θ with attention-head set $\mathcal{H}_\theta$;
disjoint discovery and held-out sets $\mathcal{D}_{\text{disc}}$ and $\mathcal{D}_{\text{held}}$;
subsampling percentage P; selection fractions $\rho_h, \rho_e \in (0, 1]$;
number of repetitions N; integer threshold K, where $0 \le K < N$.
Ensure: Directed weighted graph $G^{\rightarrow} = (V, E, W)$.

Stage 1: Discovery-based node selection
Initialize $c_h \leftarrow 0$ for every $h \in \mathcal{H}_\theta$.
for $r = 1, \ldots, N$ **do**
  $S_r^h \leftarrow \text{Subsample}(\mathcal{D}_{\text{disc}}, P/100)$
  Compute $a_h^{(r)} \leftarrow \frac{1}{|S_r^h|} \sum_{x \in S_r^h} |\text{Attr}_h(x)|$ for every $h \in \mathcal{H}_\theta$.
  $V_r \leftarrow \text{TopFraction}(\mathcal{H}_\theta, a^{(r)}, \rho_h)$
  Increment c_h for every $h \in V_r$.
end for
$V \leftarrow \{h \in \mathcal{H}_\theta : c_h > K\}$; fix V.

Stage 2: Discovery-based edge selection
$E_{\text{cand}} \leftarrow \{(i, j) \in V \times V : \ell_i < \ell_j\}$
Initialize $c_{ij} \leftarrow 0$ for every $(i, j) \in E_{\text{cand}}$.
for $r = 1, \ldots, N$ **do**
  $S_r^e \leftarrow \text{Subsample}(\mathcal{D}_{\text{disc}}, P/100)$
  Compute $b_{ij}^{(r)} \leftarrow \frac{1}{|S_r^e|} \sum_{x \in S_r^e} |\text{PairEAP}_{i \to j}(x)|$ for every $(i, j) \in E_{\text{cand}}$.
  $E_r \leftarrow \text{TopFraction}(E_{\text{cand}}, b^{(r)}, \rho_e)$
  Increment c_{ij} for every $(i, j) \in E_r$.
end for
$E \leftarrow \{(i, j) \in E_{\text{cand}} : c_{ij} > K\}$; fix E.

Stage 3: Held-out evaluation and edge weighting
Evaluate selected heads through interventions on $\mathcal{D}_{\text{held}}$, without changing V.
Initialize $W \leftarrow \mathbf{0}_{|V| \times |V|}$.
for all $(i, j) \in E$ **do**
  Intervene on the connection $i \to j$ for each $x \in \mathcal{D}_{\text{held}}$.
  Compute $\Delta m_{ij}(x) \leftarrow m_{ij}^{\text{int}}(x) - m_{ij}^{\text{base}}(x)$ for each held-out example.
  $W_{ij} \leftarrow \frac{1}{|\mathcal{D}_{\text{held}}|} \sum_{x \in \mathcal{D}_{\text{held}}} |\Delta m_{ij}(x)|$
end for
return $G^{\rightarrow} = (V, E, W)$.

Here, $\text{Attr}_h(x)$ denotes the attribution of head h with respect to the answer margin. $\text{PairEAP}_{i \to j}(x)$ denotes the receiver-specific edge attribution based on EAP (Syed et al., 2024). The quantities $m_{ij}^{\text{int}}(x)$ and $m_{ij}^{\text{base}}(x)$ are the answer margins under the connection intervention and its corresponding baseline, respectively. Held-out evaluation does not alter the selected node or edge sets.

B SWP Implementation Details

B.1 Reference Network Generation

For each observed undirected weighted graph, we generate lattice and random reference networks following the weighted reference-network construction procedure of Muldoon et al. (2016). Both reference ensembles are matched to the observed graph in the number of nodes, the number of edges, and the empirical edge-weight distribution.

The lattice references favor local connections and assign stronger weights to shorter distances in the lattice arrangement, providing a baseline for strong local clustering. The random references random-

ize connections and the placement of edge weights, providing a baseline for short global paths. We use the original SWP reference-generation procedure rather than introduce a new reference model.

For each observed graph, we generate $R = 5{,}000$ realizations of each reference type. Let $C_t^{(r)}$ and $L_t^{(r)}$ denote the weighted clustering coefficient and weighted characteristic path length of realization r, respectively, where $t \in \{\mathrm{latt}, \mathrm{rand}\}$. The reference statistics are obtained by averaging over realizations:

$$C_t = \frac{1}{R}\sum_{r=1}^{R} C_t^{(r)}, \qquad L_t = \frac{1}{R}\sum_{r=1}^{R} L_t^{(r)}, \qquad R = 5{,}000. \tag{19}$$

The resulting C_{latt}, L_{latt}, C_{rand}, and L_{rand} are used to compute the normalized deviations and SWP as defined in the main text. Thus, we average the reference statistics before computing SWP, rather than average SWP scores computed separately for individual realizations.

C Detailed Experimental Settings

C.1 Model Families and Scales

Table 2 lists the model scales used in the accuracy–SWP experiments. Pythia, Qwen2.5, and Qwen3 use base checkpoints, whereas Gemma-3 uses instruction-tuned checkpoints. Comparisons across families therefore assess whether similar patterns occur under different model configurations; they do not isolate the effect of parameter count.

C.2 Tasks, Prompts, and Behavioral Samples

Behavioral evaluation sample counts are summarized in Table 3.

Modified arithmetic. Pythia is evaluated on Control and Overall arithmetic. Qwen3 and Qwen2.5 are evaluated on Plus-one arithmetic.

MMLU. Qwen3 is evaluated on Logical Fallacies and Machine Learning from MMLU (Hendrycks et al., 2021). Gemma-3 is evaluated on High-school Geography and World Religions.

BIG-Bench Hard. Qwen2.5 is evaluated on Logical Deduction from BIG-Bench Hard (Suzgun et al., 2023) with 3, 5, and 7 objects under zero-shot and three-shot prompting. Gemma-3 is evaluated on Logical Deduction and Temporal Sequences.

C.3 Samples and Splits for SWP Analysis

MMLU. Qwen3 Logical Fallacies and Machine Learning, and Gemma-3 High-school Geography and World Religions use samples from the corresponding subjects with stratified splits by answer category.

BIG-Bench Hard. Qwen2.5 Logical Deduction uses answer-label-stratified splits for each object count and prompting condition. Gemma-3 Logical Deduction and Temporal Sequences use the corresponding BBH samples and split configurations.

C.4 Accuracy References

For MMLU and BBH, uniform guessing among four options gives 25% accuracy. For logical deduction with K objects, uniform guessing among the K answer options gives 1/K: approximately 33.33%, 20.00%, and 14.29% for 3, 5, and 7 objects, respectively. For open-ended arithmetic, chance accuracy depends on the assumed distribution over integer answers. The 0% line in the arithmetic plots is therefore used as a zero-accuracy reference, not as a derived chance level.

Table 2: Model families and parameter scales used in the accuracy–SWP experiments. M and B denote millions and billions of parameters.

Family	Parameter scales	Count
Pythia	70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, 12B	8
Qwen2.5	0.5B, 1.5B, 3B, 7B, 14B	5
Qwen3	0.6B, 1.7B, 4B, 8B	4
Gemma-3	270M, 1B, 4B, 12B	4
Total		21

Table 3: Behavioral evaluation sample counts. For Qwen2.5 logical deduction, counts are per object-count setting.

Family	Task or setting	Examples
Pythia	Control arithmetic	3,000
Pythia	Overall arithmetic	6,000
Qwen3	Plus-one arithmetic	3,000
Qwen3	MMLU: Logical Fallacies	163
Qwen3	MMLU: Machine Learning	112
Qwen2.5	Plus-one arithmetic	3,000
Qwen2.5	BBH: Logical Deduction, 0-shot, each of 3/5/7 objects	250
Qwen2.5	BBH: Logical Deduction, 3-shot, each of 3/5/7 objects	250
Gemma-3	MMLU: High-school Geography	198
Gemma-3	MMLU: World Religions	171
Gemma-3	BBH: Logical Deduction	750
Gemma-3	BBH: Temporal Sequences	250

D Additional Accuracy–SWP Results

This appendix reports the complete accuracy–SWP trajectories complementing Figure 3. Results are grouped by model family to facilitate comparison across tasks and prompting conditions.

D.1 Gemma-3 Tasks

Across the evaluated Gemma-3 tasks, the highlighted intervals show concurrent increases in accuracy and SWP, with the magnitude of both changes varying across tasks.

D.2 Qwen2.5 Tasks

Figure 7 compares Plus-one arithmetic with zero-shot logical deduction across three difficulty levels. In logical deduction, SWP increases sharply over the highlighted interval, while the corresponding accuracy gains vary with task difficulty.

Figure 8 shows the corresponding three-shot logical-deduction results. The structural transition occurs over the same scale interval across the three difficulty levels, whereas the behavioral improvements differ in magnitude.

D.3 Qwen3 Tasks

The Qwen3 trajectories show different degrees of behavioral–structural correspondence across tasks, with the highlighted intervals exhibiting joint changes in accuracy and SWP of varying magnitude.

D.4 Pythia Tasks

Both Pythia arithmetic settings exhibit joint behavioral and structural changes across scale, while the magnitudes and trajectories differ between Control and Overall arithmetic.

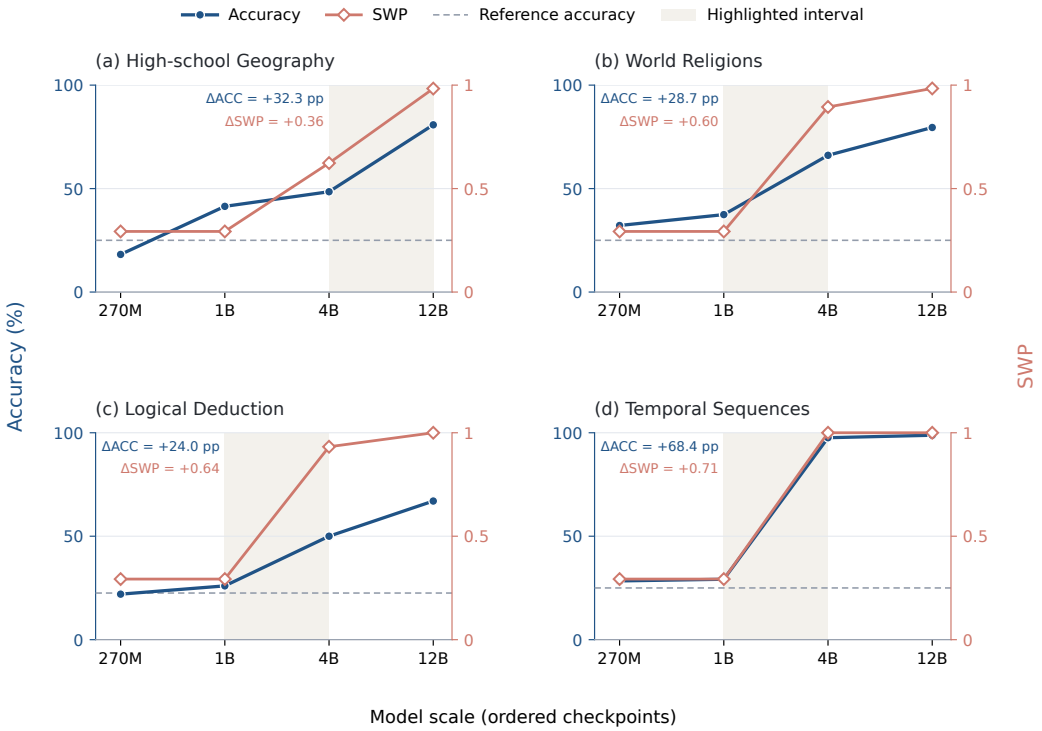


Figure 6: **Accuracy–SWP trajectories across Gemma-3 tasks.** Shading marks highlighted adjacent-scale intervals; dashed lines indicate reference accuracy.

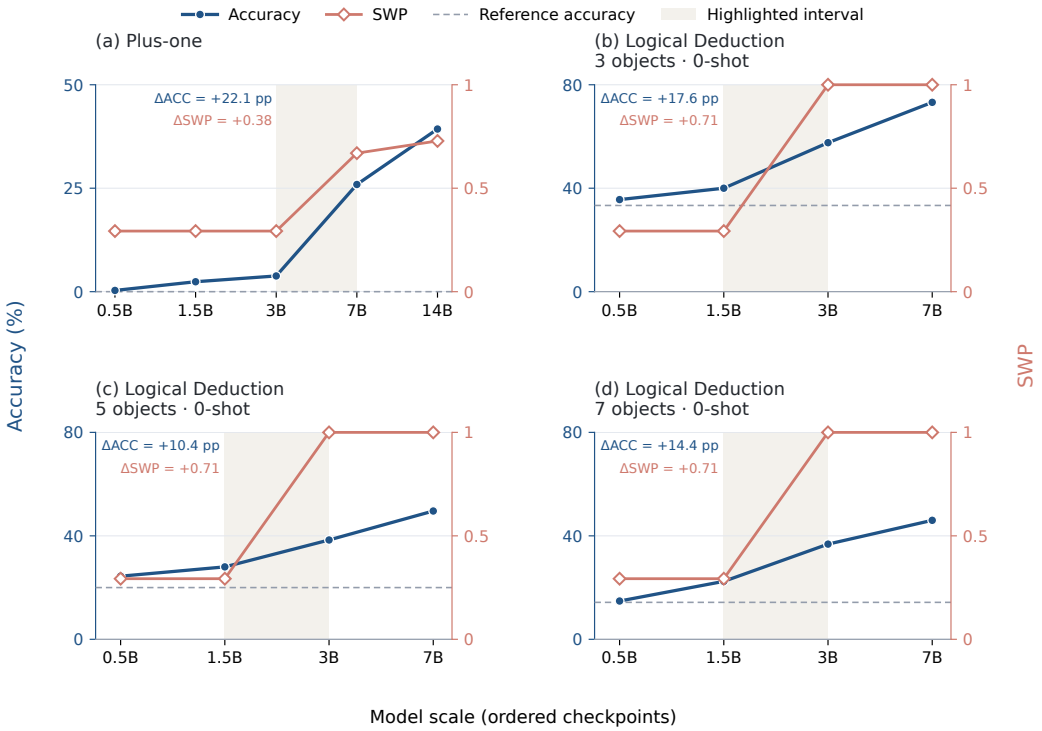


Figure 7: **Accuracy–SWP trajectories for Qwen2.5 Plus-one and zero-shot logical deduction.** Shading marks highlighted adjacent-scale intervals; dashed lines indicate reference accuracy.

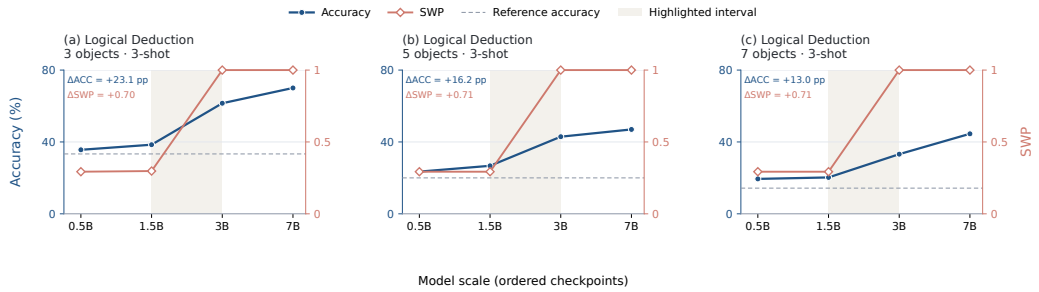


Figure 8: **Accuracy–SWP trajectories for Qwen2.5 three-shot logical deduction.** Shading marks highlighted adjacent-scale intervals; dashed lines indicate reference accuracy.

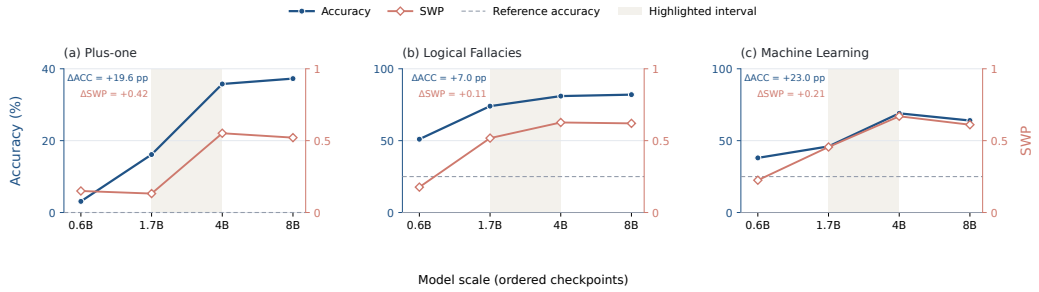


Figure 9: **Accuracy–SWP trajectories across Qwen3 tasks.** Shading marks highlighted adjacent-scale intervals; dashed lines indicate reference accuracy.

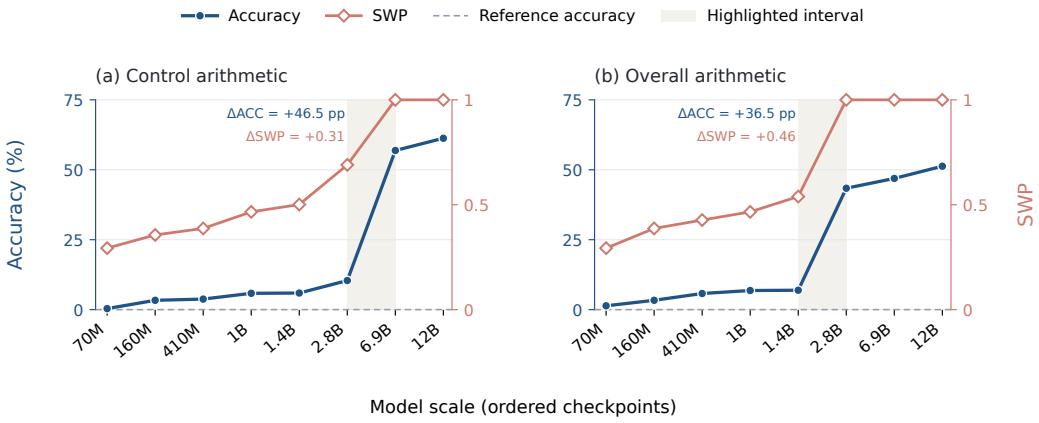


Figure 10: **Accuracy–SWP trajectories across Pythia arithmetic settings.** Shading marks highlighted adjacent-scale intervals; dashed lines indicate the zero-accuracy reference.

E CONTINUOUS BEHAVIORAL EVIDENCE FROM ANSWER MARGINS

We complement the discrete accuracy measurements with continuous answer-margin trajectories across model scales. Because margin scales differ across model families and task settings, comparisons are made within matched model–task settings rather than across absolute margin values. Shading marks the same highlighted scale intervals used for the corresponding behavioral comparisons.

E.1 GEMMA-3

Figure 11 reports the Gemma-3 trajectories across MMLU and BBH tasks. The highlighted intervals show concurrent increases in accuracy and answer margin across the evaluated settings.

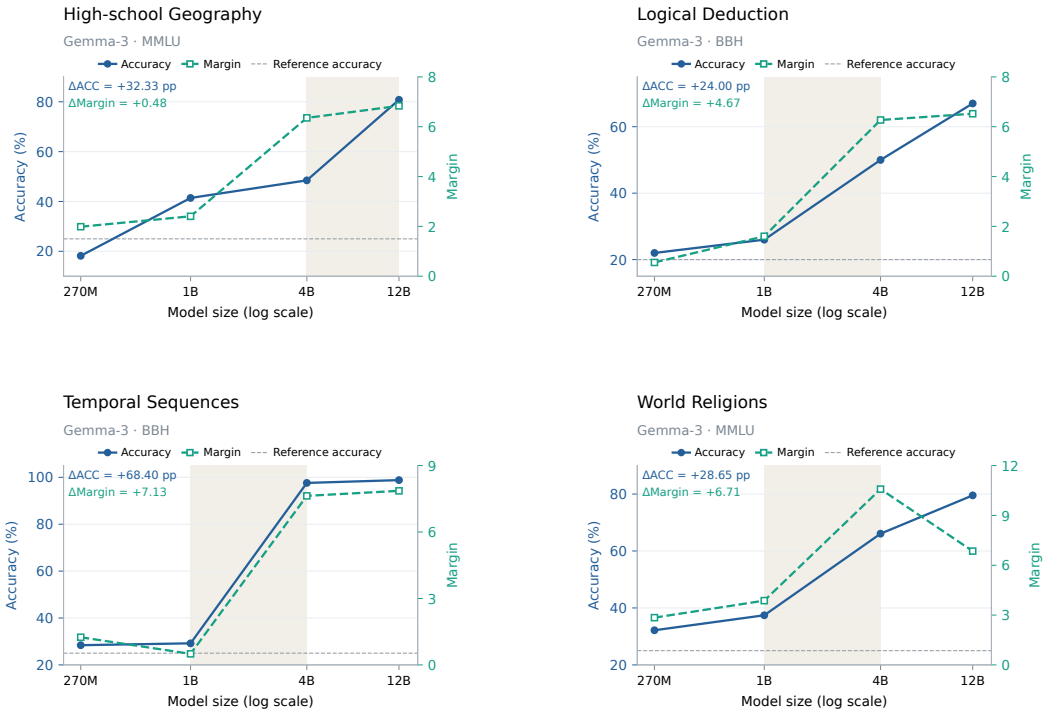


Figure 11: **Accuracy–margin trajectories across Gemma-3 tasks.** Shading marks the highlighted scale intervals; annotations report the corresponding ΔACC (pp) and ΔMargin. Dashed lines indicate reference accuracy.

E.2 PYTHIA

Figure 12 reports Control and Overall arithmetic. Both settings exhibit pronounced increases in accuracy and answer margin over the highlighted scale interval.

E.3 QWEN3

Figure 13 reports Plus-one arithmetic and the two MMLU tasks. Accuracy and answer margin increase together over the highlighted intervals across the evaluated settings.

E.4 QWEN2.5

The Qwen2.5 Plus-one trajectory is included in Figure 4. Figures 14 and 15 report logical deduction across the 3-, 5-, and 7-object settings under zero-shot and three-shot prompting. Across the

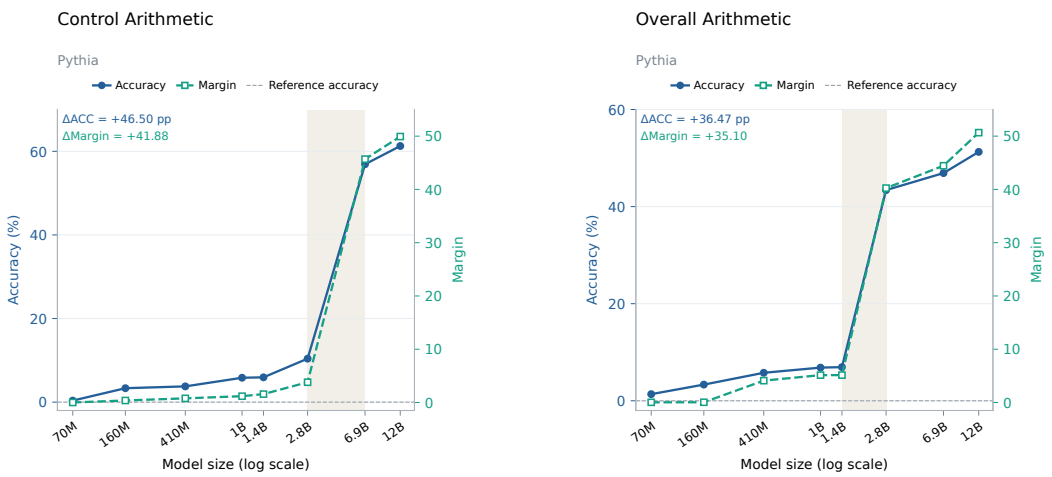


Figure 12: **Accuracy–margin trajectories across Pythia arithmetic settings.** Shading marks the highlighted scale intervals; annotations report the corresponding ΔACC (pp) and ΔMargin. Dashed lines indicate the zero-accuracy reference.

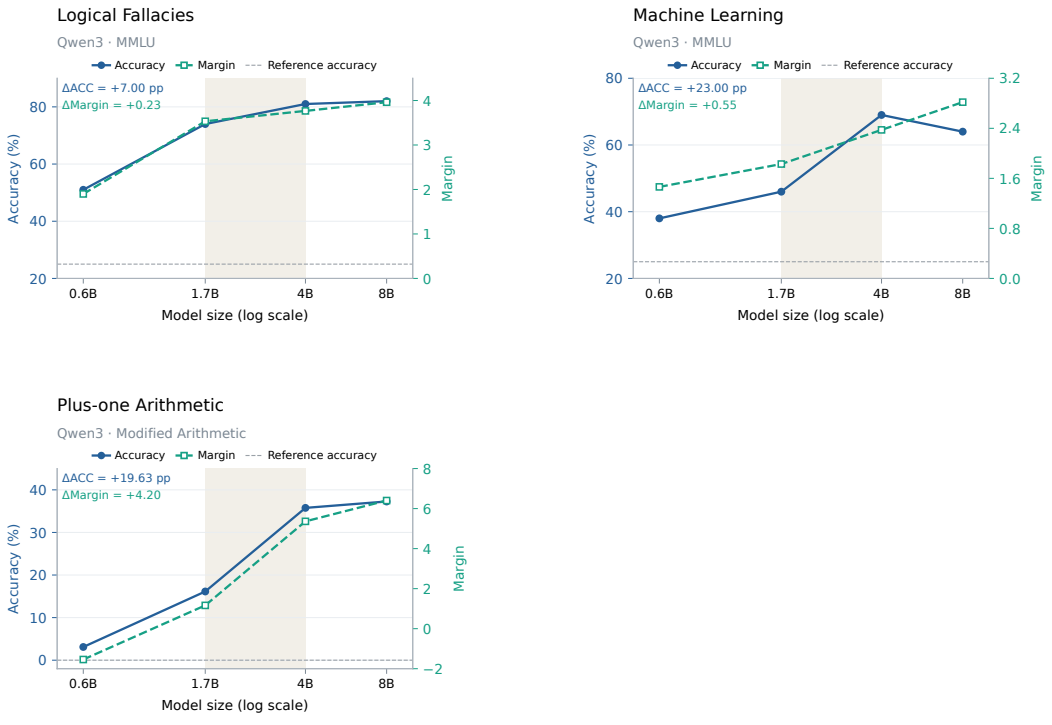


Figure 13: **Accuracy–margin trajectories across Qwen3 tasks.** Shading marks the highlighted scale intervals; annotations report the corresponding ΔACC (pp) and ΔMargin. Dashed lines indicate reference accuracy.

highlighted 1.5B–3B interval, accuracy and answer margin increase across all evaluated difficulty levels.

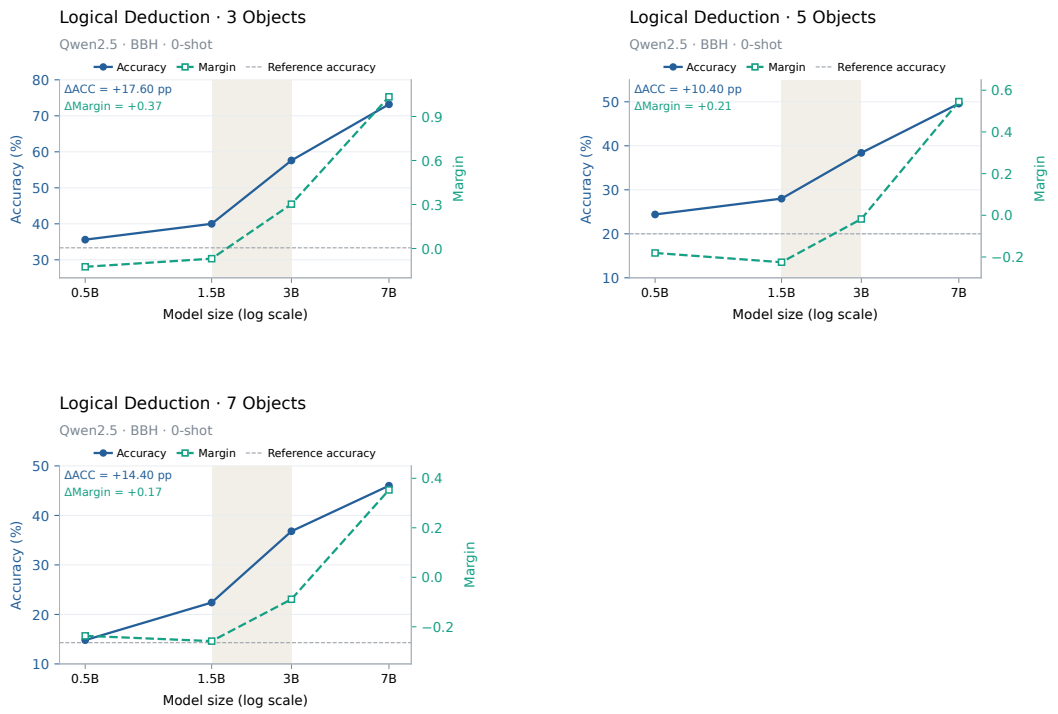


Figure 14: **Accuracy–margin trajectories for Qwen2.5 zero-shot logical deduction.** Panels show the 3-, 5-, and 7-object settings. Shading marks the highlighted scale interval; annotations report ΔACC (pp) and ΔMargin. Dashed lines indicate reference accuracy.

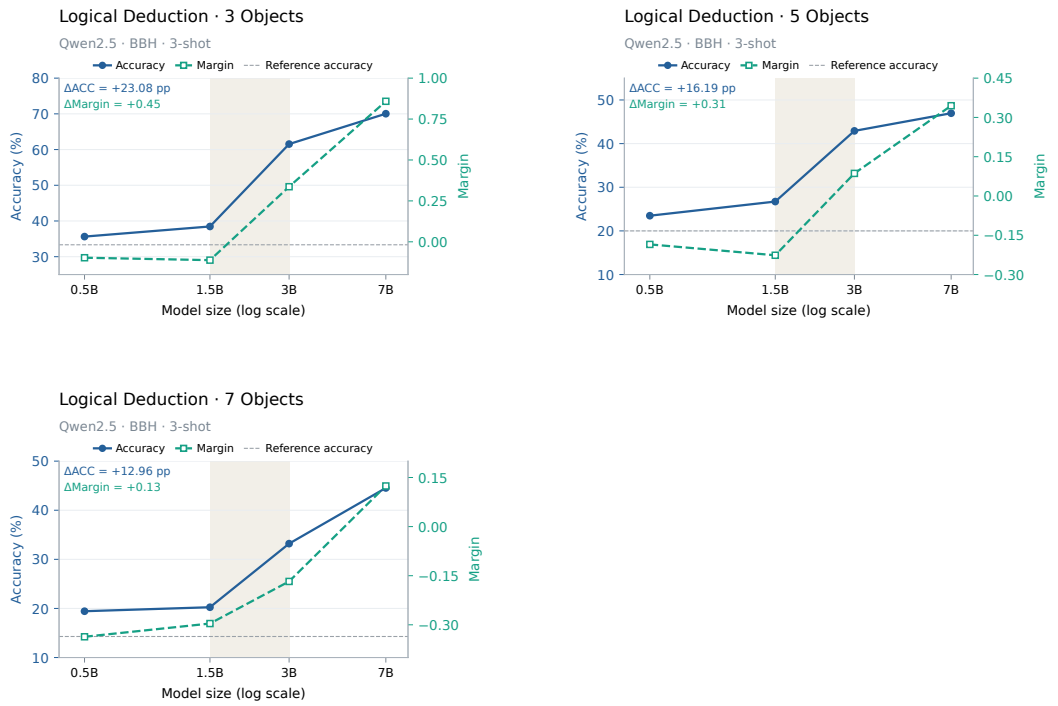


Figure 15: **Accuracy–margin trajectories for Qwen2.5 three-shot logical deduction.** Panels show the 3-, 5-, and 7-object settings. Shading marks the highlighted scale interval; annotations report ΔACC (pp) and ΔMargin. Dashed lines indicate reference accuracy.

F Intervention Validation of Selected Attention Heads

F.1 Evaluation Protocol

Reported scope and selection. We report 20 checkpoints from four model families, covering 16 family-specific task–prompt settings and 72 model–task–prompt combinations. The reported scope includes Pythia Control and Overall arithmetic; Qwen3 Plus-one arithmetic, Logical Fallacies, and Machine Learning; Qwen2.5 Plus-one arithmetic and Logical Deduction with 3, 5, and 7 objects under zero-shot and three-shot prompting; and Gemma-3 High-school Geography, World Religions, Logical Deduction, and Temporal Sequences. Head sets are determined within each discovery partition using five stratified 80% resamples: heads appearing in the top 15% by attribution magnitude in at least three resamples are retained. Frozen sets contain 13.89%–18.75% of all heads. Held-out effects are not used to revise the sets. All reported q-values retain the pre-specified global BH adjustment over the full consolidated analysis; the correction is not recomputed after restricting the reported task set.

Control sets. Each of ten random sets has the same size as the selected set and excludes both selected and bottom-ranked heads. Sampling is without replacement within a set; different random sets may overlap. Random controls are matched exactly by layer when possible; otherwise, head counts are matched within four relative-depth bins. All units successfully construct ten random controls. Bottom-ranked controls contain an equal number of heads selected by ascending discovery mean absolute attribution after excluding selected heads; ties are resolved by layer and head index. Bottom-ranked sets are not matched by layer.

Splits and interventions. Split seeds define two discovery/held-out partitions. Within each split, selection and evaluation use separate examples; held-out sets overlap across splits, making the second split a partition-consistency check rather than an independent replication. Head-output slices before the output projection are patched at all token-aligned sequence positions, with all heads in a set patched simultaneously. Let m denote the correct-minus-designated-incorrect answer margin.

Disruption and restoration effects are

$$D = m_{\text{clean}} - m_{\text{destroy}}, \qquad R = m_{\text{restore}} - m_{\text{corrupted}}.$$

Disruption transfers corrupted-source activations into a clean run; restoration transfers clean-source activations into a corrupted run. For Qwen and Gemma, m uses sequence-summed answer log-probabilities. Pythia averages token log-probabilities within each answer before taking their difference. Raw effects are not directly comparable between these two implementations.

Statistical analysis. For each held-out example, effects from the ten random sets are averaged before forming the Selected–Random contrast. Selected–Bottom contrasts are likewise paired within examples. Positive contrasts indicate larger intervention effects for selected heads, but do not by themselves establish that the selected-set effect is positive. Intervals use 2,000 percentile bootstrap resamples.

For Pythia Overall arithmetic, Random and Selected–Random intervals resample both examples and random sets; its other intervals resample examples only. In all remaining reported settings, intervals resample examples after averaging random-set effects within each example.

One-sided paired sign-flip tests use 2,000 Monte Carlo draws under the null sign-exchangeability assumption for paired differences. The reported subset contains 576 contrasts (72×2 directions $\times 2$ splits $\times 2$ controls). All reported q-values retain global BH adjustment across the full 928 contrasts in the consolidated analysis; correction is not rerun after restricting the reported settings.

A model-by-direction comparison is counted as consistent only when its mean contrast is positive and global $q < 0.05$ in both splits. The Both column additionally requires this criterion in both intervention directions. Confidence intervals and adjusted one-sided tests are distinct summaries; significance is not inferred from whether an interval crosses zero.

Verification scope. The reported counts, contrast differences, and global BH-adjusted values were checked against the consolidated summary records. This summary-level verification does not independently reproduce sample-level interventions or sign-flip tests.

F.2 Summary Across Tasks

Table 4 summarizes all 16 reported family-specific task–prompt settings. Consistent Selected–Random advantages occur in 55 of 72 model–task–prompt combinations for each intervention direction; 54 combinations satisfy the criterion in both directions.

F.3 Intervention Effects Across Models and Tasks

The following figures report intervention effects across the complete set of evaluated settings. Each figure compares Selected–Random and Selected–Bottom contrasts under disruption and restoration across the two data splits. Points show paired mean differences with 95% bootstrap confidence intervals; asterisks indicate positive contrasts with global BH-adjusted $q < 0.05$. Units are reported within each figure.

Pythia Figures 16 and 17 report intervention effects for Control and Overall arithmetic, respectively. Both settings show consistent Selected–Random advantages across model scales, with six of eight checkpoints satisfying the cross-split, cross-direction consistency criterion in each setting.

Qwen3 Figure 18 reports intervention effects for Plus-one arithmetic, Logical Fallacies, and Machine Learning. Across all three tasks, all four evaluated checkpoints satisfy the Selected–Random criterion in both intervention directions and both splits.

Qwen2.5 Figures 19 and 20 report intervention effects for Plus-one arithmetic and logical deduction across object counts and prompting conditions. Plus-one arithmetic satisfies the consistency criterion at all evaluated checkpoints, while logical deduction shows repeated support under both zero-shot and three-shot prompting, with stronger aggregate consistency under three-shot prompting.

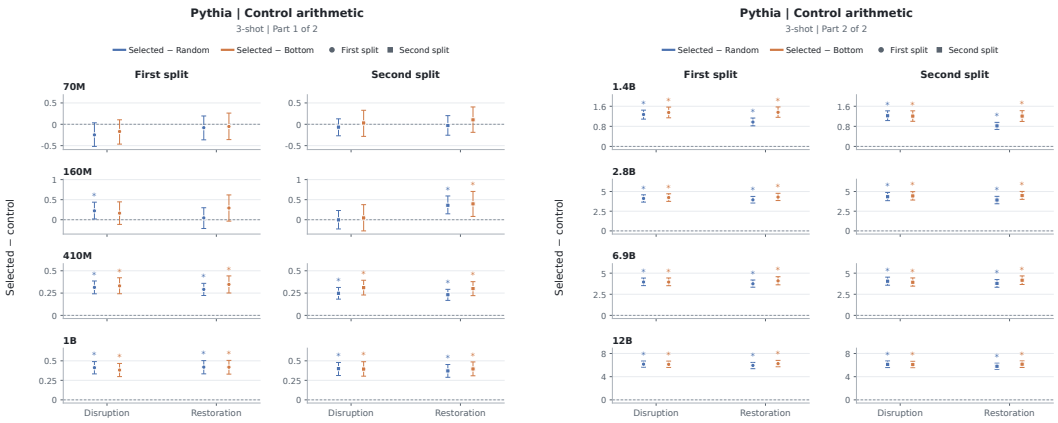


Figure 16: **Selected-head intervention effects for Pythia Control arithmetic.** Left and right panels cover the smaller and larger checkpoints, respectively.

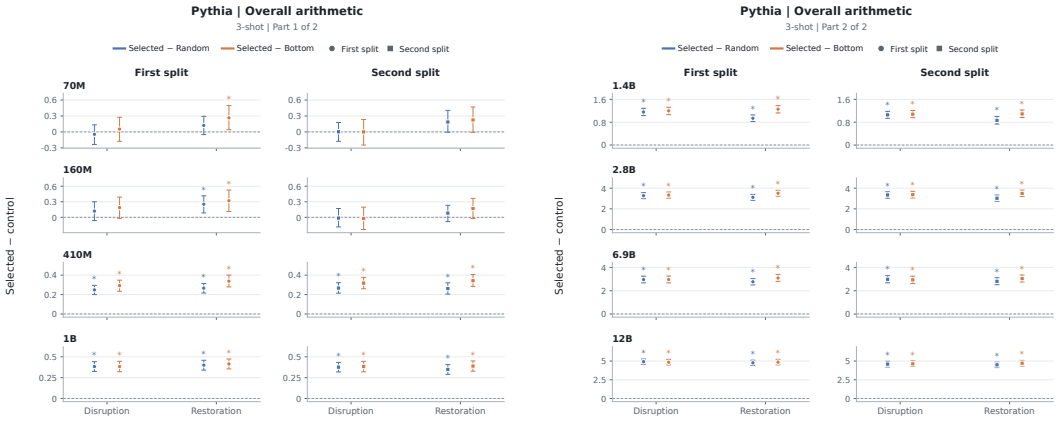


Figure 17: **Selected-head intervention effects for Pythia Overall arithmetic.** Left and right panels cover the smaller and larger checkpoints, respectively.

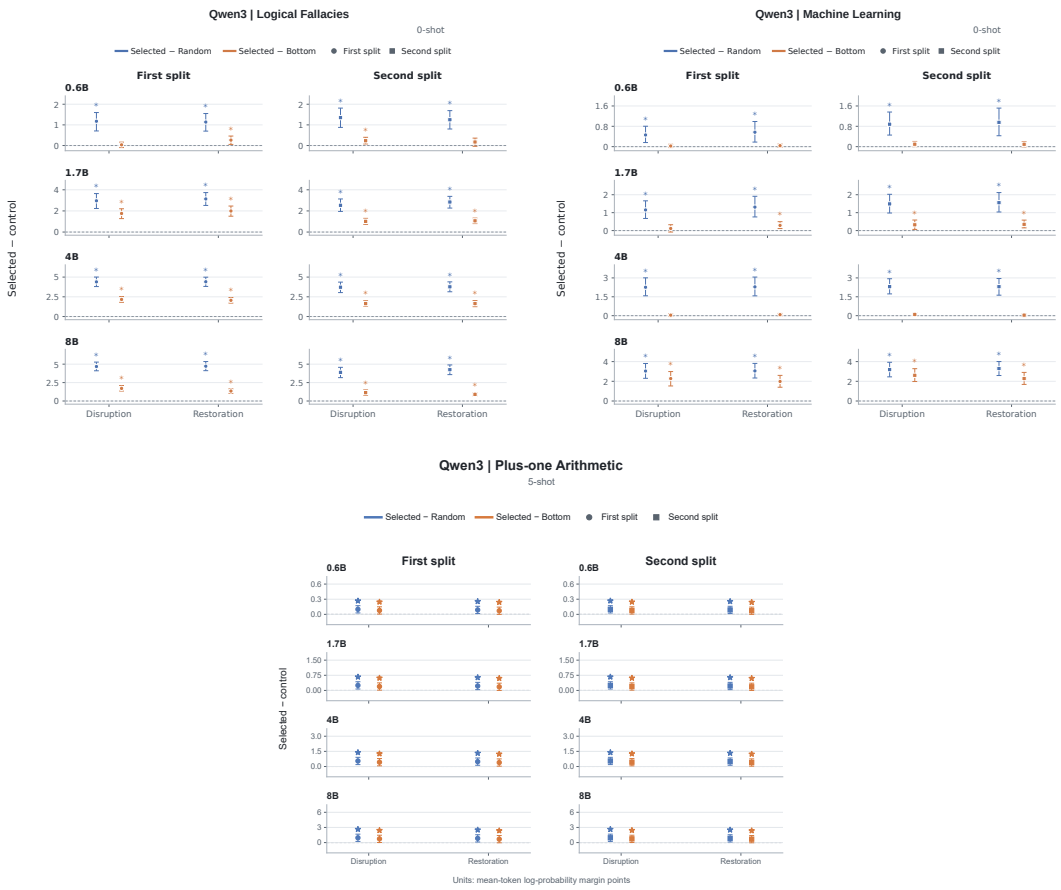


Figure 18: **Selected-head intervention effects across Qwen3 tasks.** Top: Logical Fallacies and Machine Learning. Bottom: Plus-one arithmetic.



Figure 19: **Selected-head intervention effects for Qwen2.5 Plus-one and 3-object logical deduction.** The logical-deduction panels compare zero-shot and three-shot prompting.

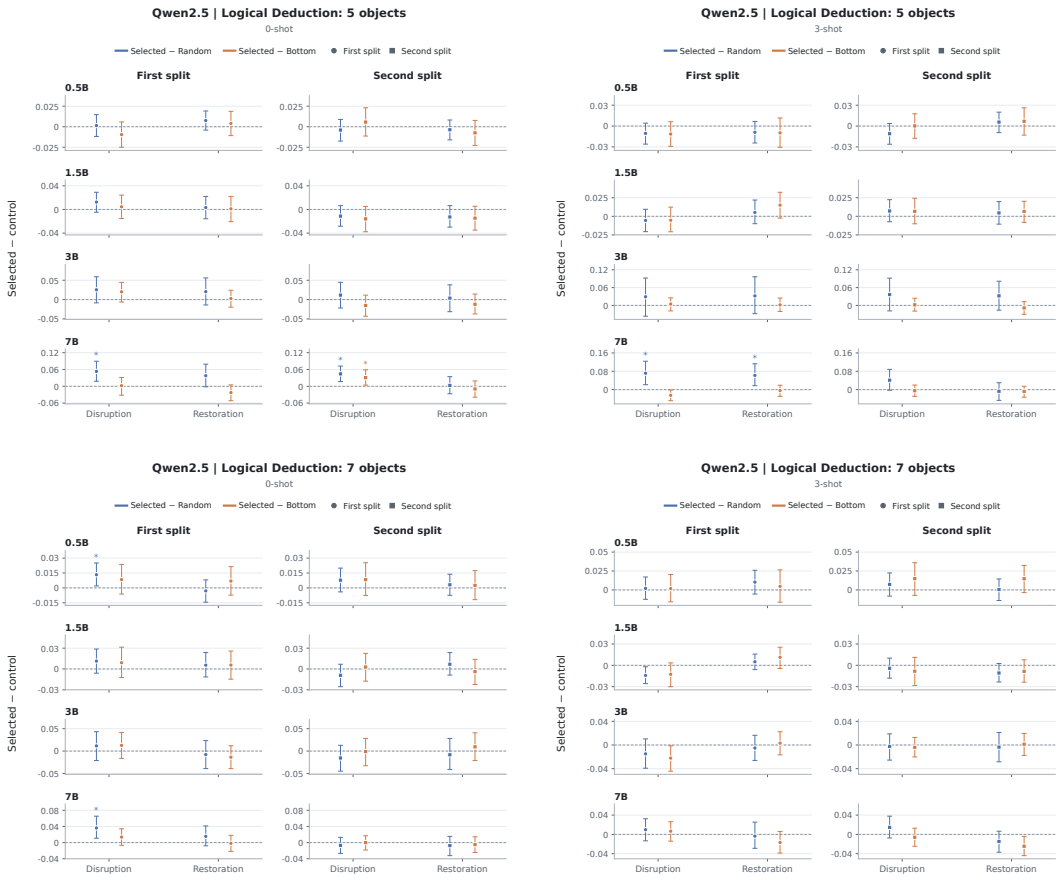


Figure 20: **Selected-head intervention effects for Qwen2.5 logical deduction.** Rows show the 5- and 7-object settings; columns compare zero-shot and three-shot prompting.

Table 4: **Consistency of Selected–Random intervention effects.** Entries report supported/evaluated model–task–prompt combinations. Each direction requires a positive contrast and global $q < 0.05$ in both splits; Both requires this criterion in both intervention directions. Qwen2.5 Logical Deduction rows aggregate the 3-, 5-, and 7-object settings. The reported settings retain the global BH adjustment over all 928 comparisons.

Family	Task	Shots	Disruption	Restoration	Both
Pythia	Overall arithmetic	3	6/8	6/8	6/8
Pythia	Control arithmetic	3	6/8	6/8	6/8
Qwen3	Plus-one arithmetic	5	4/4	4/4	4/4
Qwen3	MMLU: Logical Fallacies	0	4/4	4/4	4/4
Qwen3	MMLU: Machine Learning	0	4/4	4/4	4/4
Qwen2.5	Plus-one arithmetic	0	4/4	4/4	4/4
Qwen2.5	Logical Deduction: 3/5/7 objects	0	5/12	6/12	5/12
Qwen2.5	Logical Deduction: 3/5/7 objects	3	8/12	7/12	7/12
Gemma-3	MMLU: High-school Geography	0	3/4	3/4	3/4
Gemma-3	MMLU: World Religions	0	3/4	3/4	3/4
Gemma-3	BBH: Logical Deduction	0	4/4	4/4	4/4
Gemma-3	BBH: Temporal Sequences	0	4/4	4/4	4/4
Total	16 task–prompt settings		**55/72** (76.39%)	**55/72** (76.39%)	**54/72** (75.00%)

Totals count model–task–prompt combinations rather than unique checkpoints. Disruption and restoration together yield 110/144 supported model-by-direction comparisons.

Gemma-3 Figure 21 reports intervention effects across the Gemma-3 MMLU and BBH tasks. Logical Deduction and Temporal Sequences satisfy the Selected–Random criterion at all four evaluated checkpoints, while High-school Geography and World Religions satisfy it at three of four checkpoints.

F.4 Hub- and Shortcut-Associated Components

Coverage and summary. Table 5 summarizes 16 family-specific task–prompt settings across 72 model–task–prompt combinations. Each available comparison corresponds to one model, one split, one intervention direction, and one structural group type. The reported support counts are 222/280 (79.29%) for hub-associated groups and 242/288 (84.03%) for shortcut-associated groups.

For Pythia 70M, matched hub controls cannot be constructed for either arithmetic setting in either split and intervention direction, leaving eight model–split–direction hub comparisons unavailable. These comparisons are excluded from the denominator rather than counted as unsupported outcomes.

Counting convention. The counts in Table 5 refer to individual comparisons. They do not require agreement across both splits or both intervention directions. Consequently, their proportions are not directly comparable to the consistency counts reported for selected head sets.

Interpretation. The reported support proportions vary across tasks and structural group types. These comparisons concern component groups identified through graph structure. They do not by themselves establish that hub or shortcut connectivity has a causal contribution independent of the components involved. Nor does the higher aggregate shortcut support proportion establish a larger shortcut effect.

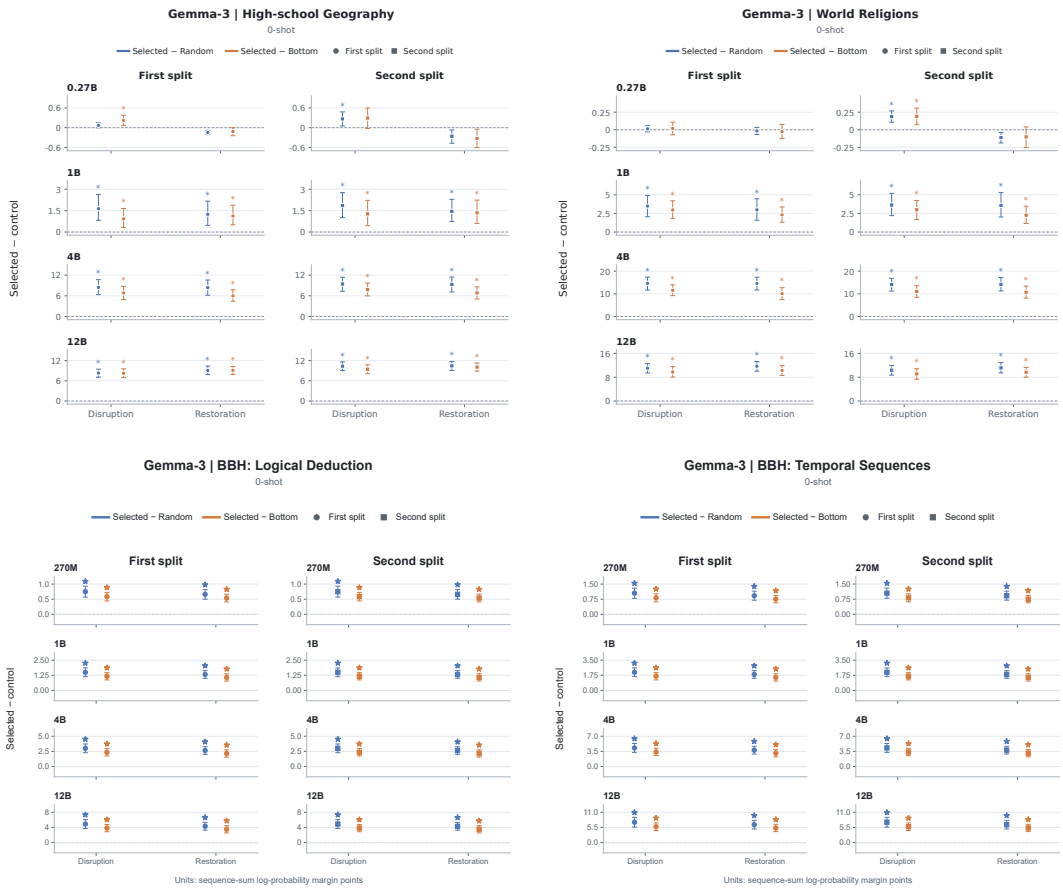


Figure 21: **Selected-head intervention effects across Gemma-3 tasks.** Top: High-school Geography and World Religions. Bottom: Logical Deduction and Temporal Sequences.

Table 5: **Reported support for hub- and shortcut-associated groups.** Entries give supported comparisons over available comparisons. Each comparison corresponds to one model, split, and intervention direction. Qwen2.5 Logical Deduction rows aggregate the 3-, 5-, and 7-object settings. The two columns describe separate sets of comparisons and should not be interpreted as a direct comparison of hub and shortcut effect sizes.

Family	Task	Shots	Hub	Shortcut
Pythia	Overall arithmetic	3	23/28	28/32
Pythia	Control arithmetic	3	21/28	26/32
Qwen3	Plus-one arithmetic	5	11/16	14/16
Qwen3	MMLU: Logical Fallacies	0	12/16	14/16
Qwen3	MMLU: Machine Learning	0	13/16	12/16
Qwen2.5	Plus-one arithmetic	0	14/16	15/16
Qwen2.5	Logical Deduction: 3/5/7 objects	0	40/48	41/48
Qwen2.5	Logical Deduction: 3/5/7 objects	3	35/48	38/48
Gemma-3	MMLU: High-school Geography	0	11/16	14/16
Gemma-3	MMLU: World Religions	0	13/16	11/16
Gemma-3	BBH: Logical Deduction	0	15/16	14/16
Gemma-3	BBH: Temporal Sequences	0	14/16	15/16
Total	16 task–prompt settings		**222/280** (79.29%)	**242/288** (84.03%)

G Complete Layer-Ordered Null-Model Results

We report the complete layer-ordered null-model comparison across all 12 task settings. Across 160 functional graphs, 112 (70.00%) have SWP above the null mean with BH-adjusted $q < 0.05$. The q-values retain the original global BH correction over 272 SWP tests, consistent with the analysis reported in the main text.

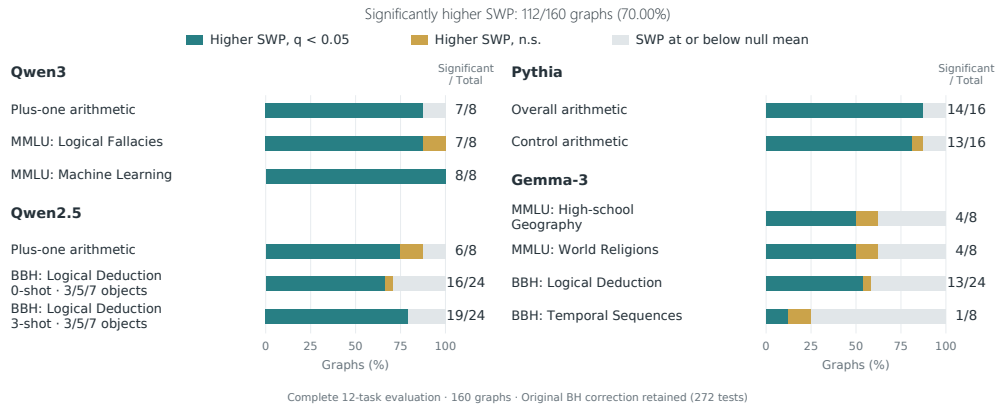


Figure 22: **Complete SWP comparison with the layer-ordered random-edge null model.** Bars summarize all 12 task settings (160 graphs), with 5,000 randomizations per graph. Overall, 112/160 (70.00%) have SWP above the null mean with BH-adjusted $q < 0.05$. The original BH correction over 272 tests is retained.