

Belief or Circuitry? Causal Evidence for In-Context Graph Learning

Katharine Kowalyshyn* and Timothy Duggan*

Daniel Little and Michael C. Hughes

Department of Computer Science, Tufts University, Medford, United States

katharine.kowalyshyn@tufts.edu

Abstract

How do LLMs learn in-context? Is it by pattern-matching recent tokens, or by inferring latent structure? We probe this question using a toy graph random-walk across two competing graph structures. We present behavioral, representational, and causal evidence that neither account alone is sufficient. First, a complexity-weighted sigmoid model captures topology-dependent evidence accumulation, and PCA shows that both graph topologies are visually recoverable at intermediate mixture ratios. Second, residual-stream activation patching and graph-difference steering causally intervene on this graph-family signal: late-layer patching almost fully transfers the clean graph preference, while linear steering moves predictions in the intended direction and fails under norm-matched and label-shuffled controls. An explicit transition-cache baseline reproduces the fitted asymmetry and held-out-edge transfer, however, so these results do not by themselves distinguish global inference from local copying. Taken together, our findings are consistent with a dual-mechanism account in which structure inference and induction circuits operate in parallel, but do not uniquely identify it.

1 Introduction & Related Works

Since the advent of LLMs, in-context learning (ICL) has remained an area of research that remains poorly understood. Over the past five years, numerous investigations into ICL have yielded interesting results: alignment (Anwar et al., 2024; Lin et al., 2023), jailbreaking (Polyakov and Kuznetsov, 2026), and demonstration selection (Qin et al., 2024) are just a few areas that have prioritized it as a subfield of study (Dong et al., 2024).

Within the subfield of mechanistic interpretability (MechInterp), there is yet to be a concrete consensus on this topic. Originally posed as a debate between inference over latent structure or shallow

pattern-matching circuits by Olsson et al. (2022), ICL has seen numerous recent works in MechInterp from theoretical accounts framing ICL as implicit Bayesian inference over latent concepts (Xie et al., 2022), to mechanistic studies of induction head formation and diversity (Singh et al., 2024), to causal evidence that ICL decomposes into separable task schema and input-output binding mechanisms (Kim, 2025).

Recent work from Park et al. (2025) provided striking evidence for the former: given random-walk traces on an unknown graph as a toy task, Llama-3.1-8B exhibits a *sharp phase transition* in neighbor-prediction accuracy. This transition results in simultaneously reorganizing its residual-stream geometry to mirror the graph’s adjacency structure. If the phase transition reflects latent-structure inference, then LLMs maintain implicit probabilistic world models that can be probed mechanistically. If it reflects induction circuits, the same behavioral signature requires only local copying heads and carries no implications about global representations.

Contributions. Our contributions are a first step to triangulate the answer to this ongoing debate. (1) We replace the flat log-prior of Bigelow et al. (2025) with a complexity-weighted structure-specific prior, report its fitted parameters and model-selection scores, and test the same fit on a transition-cache baseline. (2) We expose Llama-3.1-8B to interleaved walks from two competing graphs and show topology-biased evidence accumulation. (3) PCA of residual-stream activations reveals that both graph topologies are visually recoverable at intermediate mixture ratios; principal-angle comparisons quantify alignment with the pure-graph references, while a twelve-point energy curve shows that the aggregate geometric trend is weak. (4) Activation patching and steering causally link this representational structure to next-token predictions; a matched cache intervention clarifies

* Equal contribution.

that held-out-edge transfer alone does not rule out local copying.

Ultimately, our conclusion is that the question of whether LLMs *believe* or *copy* may be a false dichotomy, and the architecture of that coexistence is what mechanistic interpretability must now explain.

2 Background

Recently published works in ICL have taken two simultaneous approaches to investigating this problem. [Park et al. \(2025\)](#) used a toy task of random walks over a sixteen word grid where nodes are non-semantically related words and edges are connections between word pairs adjacent in the grid. That paper showed that Llama-3.1-8B undergoes a sharp phase transition in neighbor-hit accuracy as context length grows, and that layer-level PCA of node-token representations progressively recovers the true graph topology. This was interpreted as evidence for an implicit Bayesian world model over graph structure. Following the publication of [Park et al. \(2025\)](#), two blog posts entered the debate. [Arditi \(2026\)](#) identified specific attention heads in Llama-3.1-8B that implement induction, arguing that the phase transition in graph ICL is fully explained by these heads accumulating local transition statistics, with no need to posit global structure inference. [Ransome \(2026\)](#) replicated these findings and extended the analysis to additional graph topologies. We note that this second post appeared concurrently with our own work; we refer readers to both for complementary perspectives on the mechanistic debate.

Alongside the ongoing debate around work by [Park et al. \(2025\)](#), there was a theoretical Bayesian dynamics approach to ICL published by [Bigelow et al. \(2025\)](#). These authors fit a sigmoid-shaped parametric function over log-odds evidence accumulation to LLM accuracy curves, treating the model as maintaining a latent binary belief over two hypotheses about the data source. They found evidence for a dual-mechanism account in which both Bayesian updating and induction circuits contribute.

2.1 An Operational Test: Locality Versus Generalization

A recurring objection to this line of work is that the two accounts can seem indistinguishable in principle: a graph is nothing more than its set of edges,

so any mechanism that tracks transition statistics is, in some sense, tracking graph structure. We adopt locality as a concrete way to probe the distinction. Without an intervention, a transition cache can only reproduce statistics for transitions observed in the visible context, whereas a mechanism that has inferred global topology may generalize to unseen edges. This criterion motivates three diagnostics in our setting: complexity sensitivity in the fitted prior term (Equation 4), separable representations under graph mixtures, and causal effects on held-out edges. None is decisive under the present protocol. In particular, the clean donor context used for activation patching may contain transitions absent from the corrupt context, allowing a local state-transfer intervention to import them. We therefore treat the three diagnostics as complementary and run matched transition-cache controls for the behavioral fit and held-out-edge intervention in Sections 4.2 and 6.4.

Concretely, our transition-cache baseline is a simple non-parametric model with no learned parameters and no access to model internals. For a given random-walk sequence, it maintains a table of observed (source node, next node) transition counts, updated online as the walk unfolds; predictions are formed by normalizing the outgoing count row for the current node, with Laplace smoothing $\alpha = 0.1$ added to every possible edge to avoid zero probabilities for unseen transitions. When we “patch” the cache (Section 6.4), we replace its corrupt-context count row for the final token with the clean-context count row, mirroring the residual-stream activation patch at the representation level. The cache is intentionally incapable of representing anything beyond locally observed transitions, so any behavior it reproduces is by construction explainable without appeal to global structure inference.

3 Behavioral Model

The central question is whether the LLM’s behavior during ICL looks more like a Bayesian reasoner accumulating evidence about latent structure, or a pattern-matcher copying recent tokens. To probe this, we fit a belief-dynamics model to observed accuracy curves and ask: do the recovered parameters tell a story consistent with genuine structural inference?

We emphasize that this section is a purely behavioral, curve-fitting exercise: fitting a sigmoid

to accuracy-vs-context curves does not by itself establish that the LLM maintains an explicit internal probability distribution over hypotheses. We use terms like “belief” and “skepticism” as shorthand for the fitted log-odds parameters, not as claims about internal representations; that mechanistic question is instead addressed directly by the representational and causal experiments in Sections 5 and 6.1.

3.1 Single-Hypothesis Baseline

We first derive a single-hypothesis baseline that motivates the general K -hypothesis framework below. Let $S \in \{0, 1\}$ be a latent binary variable labeling which accuracy regime is observed: $S = 1$ denotes the high-accuracy regime in which behavior tracks the in-context graph structure; $S = 0$ denotes the regime consistent with reliance on pretrained associations. The log-prior over S is $b \in \mathbb{R}$, with $b < 0$ encoding initial skepticism toward the arbitrary in-context graph. As the LLM observes more walk steps, evidence accumulates sub-linearly:

$$\log \frac{p(\mathbf{x} \mid S = 1)}{p(\mathbf{x} \mid S = 0)} \approx \gamma N^{1-\alpha} \quad (1)$$

where $\gamma > 0$ controls evidence strength per token and $\alpha \in (0, 1)$ captures diminishing returns from correlated walk steps: as $\alpha \rightarrow 1$, successive tokens contribute vanishingly little additional log-odds, so evidence effectively saturates after a modest number of steps rather than accumulating linearly forever; as $\alpha \rightarrow 0$, the model recovers ordinary linear (fully additive) evidence accumulation in N . Combining prior and likelihood via Bayes’ rule gives the predicted neighbor-hit accuracy at context length N :

$$\hat{p}(N) = p_0 + (q - p_0) \sigma(b + \gamma N^{1-\alpha}) \quad (2)$$

where p_0 is the pre-transition neighbor-hit rate (estimated empirically from $N \leq 100$ tokens), $q \in (p_0, 1]$ is the graph-mode success rate, and σ is the sigmoid. The phase transition inflection point is $N^* = (-b/\gamma)^{1/(1-\alpha)}$, corresponding to the context length at which the fitted log-odds cross zero, i.e., where the accuracy curve transitions from the low regime toward the high regime.

Parameters $\theta = (b, \gamma, \alpha, q)$ are fit by minimizing MSE between $\hat{p}(N)$ and observed accuracy curves, equivalent to MLE under an additive Gaussian noise model on observed accuracies. We use L-BFGS-B with 16 random restarts and

box constraints $b \in [-30, 30]$, $\gamma \in [10^{-6}, 50]$, $\alpha \in [0, 0.99]$, $q \in (p_0, 1]$. The bounds enforce domain constraints directly; the lowest-loss restart is kept and validation and test MSE are reported afterward.

3.2 General Framework

Equations 1 and 2 define the single-hypothesis baseline. We extend it to an LLM presented with a context generated by one of K competing hypotheses $\mathcal{H} = \{H_1, \dots, H_K\}$ about the latent data-generating structure. We fit, for each hypothesis H_k , a sigmoid curve to the model’s per-hypothesis accuracy as a function of context length, as if evidence for H_k were accumulating in log-odds form; this is a descriptive parameterization of the behavior, not an assumption about how the LLM internally represents the hypotheses. The curve’s log-odds offset is governed by a term $b_k \in \mathbb{R}$, with $b_k < 0$ encoding an initial bias against H_k . Evidence in favor of H_k accumulates sub-linearly with context length N , giving a predicted accuracy at context length N :

$$\hat{p}_k(N) = p_{0,k} + (q_k - p_{0,k}) \sigma(b_k + \gamma_k N^{1-\alpha_k}) \quad (3)$$

where $p_{0,k}$ is the pre-transition accuracy under H_k , q_k is the post-transition accuracy, $\gamma_k > 0$ controls evidence strength, and $\alpha_k \in (0, 1)$ captures diminishing returns from correlated observations. The inflection point $N_k^* = (-b_k/\gamma_k)^{1/(1-\alpha_k)}$ marks the context length at which the fitted curve for H_k crosses from the low- to high-accuracy regime.

A key question is what determines b_k , the LLM’s initial bias toward or against each hypothesis. We propose that b_k is governed by the *complexity* of H_k : a more complex hypothesis requires more evidence to overcome the prior. Concretely, we parameterize

$$b_k = b_0 - \lambda \cdot C(H_k), \quad (4)$$

where b_0 is a shared baseline log-odds, $\lambda \geq 0$ is a learned penalty weight, and $C(H_k)$ is an MDL-inspired complexity measure of hypothesis H_k . A positive λ is consistent with a complexity penalty, but it is not mechanism-specific: topology-dependent learning rates can induce the same fitted coefficient, as our transition-cache control shows below.

We note explicitly that the presence of the exponent α_k in Equation 3 means this model is not a pure additive log-odds (Bayesian) accumulator

in the strict sense: a genuinely additive account would have evidence scale linearly in N ($\alpha_k = 0$), and any $\alpha_k > 0$ introduces a sub-additive, saturating correction on top of the log-odds framing. Our fitted $\hat{\alpha}_k$ values are close to 1 in both vocabulary conditions (Table 1), meaning evidence accumulates very non-additively: successive walk steps contribute rapidly diminishing log-odds, and the curve’s shape past the transition is governed almost entirely by q_k and the transition location rather than by continued additive accumulation. Practically, this means the “prior strength” λ we estimate in Equation 4 should be read as a descriptive summary of how quickly each hypothesis’s saturating curve is offset from the other, not as a coefficient in a normative Bayesian updating rule. We view this as a limitation of treating the fit as evidence for genuine Bayesian inference, and flag $\alpha_k \rightarrow 1$ as an important caveat for interpreting λ as a complexity-sensitive prior.

3.3 Instantiation: Competing Graph Structures

We instantiate this framework using the graph random-walk task of Park et al. (2025). The LLM is presented with token sequences generated by random walks over an unknown graph G , and must predict the next node, a valid neighbor of the current node. The two competing hypotheses are H_{grid} and H_{ring} : a 4×4 grid (16 nodes, 24 edges, degree 2–4) and a 16-node ring (16 edges, uniform degree 2), each with nodes assigned distinct single-token English nouns. The observable accuracy $\hat{p}_k(N)$ is the neighbor-hit probability, the probability that the model’s next-token prediction is a valid graph neighbor of the current node under hypothesis H_k .

We run two vocabulary conditions. In the *disjoint* condition, the grid and ring draw their node labels from two non-overlapping sets of 16 nouns each, so no token is ever a node in both graphs. In the *neutral* (closed-vocabulary) condition used for the mixture-geometry analysis in Section 5, both graphs are defined over the *same* shared 16-token vocabulary, with edges differing between the two topologies; this isolates topology-driven representational change from vocabulary-driven change, at the cost of making the two graphs’ node identities indistinguishable from token identity alone.

The MDL complexity of each graph hypothesis is naturally given by the length of its edge-list

encoding:

$$C(G) = |E(G)| \cdot \lceil \log_2 |V| \rceil \text{ bits}, \quad (5)$$

Equation 5 yields $C(\text{grid}) = 96$ bits and $C(\text{ring}) = 64$ bits. The grid costs more to describe because it has more edges. If $\hat{\lambda} > 0$ with $\hat{b}_{\text{grid}} < \hat{b}_{\text{ring}}$, the LLM requires disproportionately more context to commit to the denser topology. Under the complexity-prior interpretation this should yield $\hat{\lambda} > 0$; the cache baseline tests whether local transition learning can produce the same result without an explicit MDL prior.

To create a genuine competition between hypotheses, we interleave walks from both graphs at a controlled mixture ratio $\rho \in [0, 1]$, where ρ is the probability that any given 100-token segment is drawn from the ring walk. The effective context length for graph k is $\rho_k \cdot N$ (a mean-field approximation that holds in expectation across walk realizations) giving

$$\hat{p}_k(\rho, N) = p_{0,k} + (q_k - p_{0,k}) \cdot \sigma(b_k + \gamma_k(\rho_k N)^{1-\alpha_k}). \quad (6)$$

We compare a **per-graph** parameterization (8 free parameters: $b_0, \lambda, \gamma_k, \alpha_k, q_k$ per graph) against a **mixture-bias** ablation (5 parameters) that shares a single sigmoid but linearly interpolates the prior: $b(\rho) = (1 - \rho)b_{\text{grid}} + \rho b_{\text{ring}}$. The mixture-bias version can capture prior asymmetry but not topology-specific evidence rates, making it a direct test of whether per-graph dynamics are needed beyond the prior alone. Model selection uses AIC and BIC, detailed in Section 3.4. We use Llama-3.1-8B (non-instruct) loaded via TransformerLens (Nanda and Bloom, 2022), with layer-26 residual-stream activations for representational analyses following Park et al. (2025).

3.4 Estimation Details

The joint objective minimizes MSE over all (ρ, k, N) triples:

$$\hat{\theta} = \arg \min_{\theta} \sum_{\rho, k} \sum_{N \in \mathcal{C}} [\hat{p}_{k,\text{obs}}(\rho, N) - \hat{p}_k(\rho, N; \theta)]^2 \quad (7)$$

We use L-BFGS-B with 24 random restarts and box constraints $b_0 \in [-15, 15]$, $\lambda \in [-2, 2]$, $\gamma_k \in [10^{-6}, 50]$, $\alpha_k \in [0, 0.99]$, $q_k \in (p_{0,k}, 1]$. Note that λ bounds include negative values; $\hat{\lambda} < 0$ would mean the LLM prefers more complex graphs,

which would falsify the complexity-prior hypothesis.

Model selection between the per-graph (8 parameters) and mixture-bias (5 parameters) versions uses the AIC and BIC definitions in Equations 8 and 9 under the Gaussian residual assumption:

$$\text{AIC} = n \cdot (\log(2\pi \cdot \text{MSE}) + 1) + 2k, \quad (8)$$

$$\text{BIC} = n \cdot (\log(2\pi \cdot \text{MSE}) + 1) + k \log n \quad (9)$$

where n is the number of training observations and k is the number of free parameters. The pre-transition accuracy $p_{0,k}$ is estimated per graph by averaging neighbor-hit accuracy over training walks at $N \leq 100$ tokens, then pooled to a single $p_{0,\text{grid}}$ and $p_{0,\text{ring}}$ per vocabulary condition for identifiability.

We note one limitation of the `overlap` fit: the optimizer saturated the lower bound of the b_0 search range ($\hat{b}_0 = -15.00$), indicating it wanted a more negative value than allowed. As a consequence, $\hat{\lambda}_{\text{overlap}}$ and the implied $\hat{b}_{\text{grid}} - \hat{b}_{\text{ring}}$ gap may be biased toward zero in this condition; widening the bounds or reparameterizing b_0 is a straightforward follow-up.

4 Experiment 1

4.1 Which Sigmoid Fits Best?

For each (condition, ρ) cell, we fit both the baseline and our model with the complexity-weighted prior to the training walks and evaluate on held-out sequences.

4.2 Behavioral Sigmoid Fits

The grid’s inflection point N^* shifts monotonically later as ring evidence increases, the graph-level competition effect captured by the per-graph model. Table 1 reports all parameters fitted with Equations 3–6 under the objective in Equation 7. The recovered penalty is positive in both vocabulary conditions and implies $\hat{b}_{\text{grid}} < \hat{b}_{\text{ring}}$. The per-graph parameterization improves AIC by 92.9 (overlap) and 91.1 (disjoint), and BIC by 77.3 and 75.5, respectively, over the mixture-bias ablation. Figure 1 shows representative observed curves and both fitted models. These are descriptive comparisons: a positive fitted λ is evidence of asymmetric dynamics, but is not by itself a mechanistic identification of a Bayesian prior (see the discussion of $\hat{\alpha}_k \rightarrow 1$ above, which further complicates a literal Bayesian reading of λ). Indeed,

a directed transition cache with Laplace smoothing 0.1 (defined in Section 2.1), run under the same disjoint-vocabulary generation protocol, recovers $\hat{\lambda} = 0.530$ (versus the LLM’s 0.542) and also prefers the per-graph model ($\Delta\text{AIC} = -81.1$, $\Delta\text{BIC} = -65.4$). Unequal topology-dependent learning rates are therefore sufficient to induce the fitted penalty even without an explicit complexity prior.

5 Experiment 2

5.1 Does the Residual Stream Encode Latent Graph Structure?

To test whether the behavioral signatures reflect genuine changes in internal representations, we probe activations directly, following the representational analysis of Park et al. (2025). For each node v , we average activations over all positions where $w_t = v$ at context length T , yielding a class-mean matrix. Projecting these class-mean vectors into PCA space lets us ask whether the low-dimensional geometry recovers the true graph topology, and, critically, whether both graph topologies are simultaneously recoverable at intermediate ρ .

To quantify structural alignment we report degree-normalized Dirichlet energy under the true graph Laplacian $L = D - A$, where A is the adjacency matrix and D is the diagonal degree matrix. We define $H_T \in \mathbb{R}^{|V| \times d}$ as the matrix of class-mean activations with rows $\mu_v(T)$ for each node v present in a trailing context window at length T , and $\bar{H}_T$ is the degree-weighted mean.

$$\begin{aligned} \mathcal{E}(T) &= \text{Tr}(H_T^\top L H_T) \\ &= \frac{1}{2} \sum_{i,j} A_{ij} \|\mu_i(T) - \mu_j(T)\|^2 \end{aligned} \quad (10)$$

We normalize Equation 10 by total degree-weighted variation:

$$\mathcal{E}_{\text{norm}}(T) = \frac{\text{Tr}(H_T^\top L H_T)}{\text{Tr}((H_T - \bar{H}_T)^\top D (H_T - \bar{H}_T))} \quad (11)$$

Lower $\mathcal{E}_{\text{norm}}$ in Equation 11 means adjacent nodes are closer together in activation space. It measures alignment with graph adjacency, but does not on its own distinguish a latent graph representation from geometry induced by transition co-occurrence. We therefore use it as a descriptive complement to the causal interventions rather than a mechanism-specific statistic.

Table 1: Joint per-graph sigmoid parameters and model-selection scores, including the transition-cache control. AIC/BIC entries give per-graph (PG) and mixture-bias (MB) values; lower is better. The overlap estimate of b_0 reaches its lower search bound.

Condition	b_0	λ	γ_g	α_g	q_g	γ_r	α_r	q_r	AIC PG	AIC MB	BIC PG / MB
Overlap	−15.000	0.461	50.000	0.965	0.928	35.289	0.952	0.933	−1889.0	−1796.1	−1847.4 / −1770.1
Disjoint	−6.844	0.542	50.000	0.966	0.935	33.124	0.953	0.930	−1709.8	−1618.7	−1668.1 / −1592.6
Cache (disjoint)	−7.177	0.530	50.000	0.969	0.951	33.201	0.958	0.954	−2102.4	−2021.3	−2060.7 / −1995.3

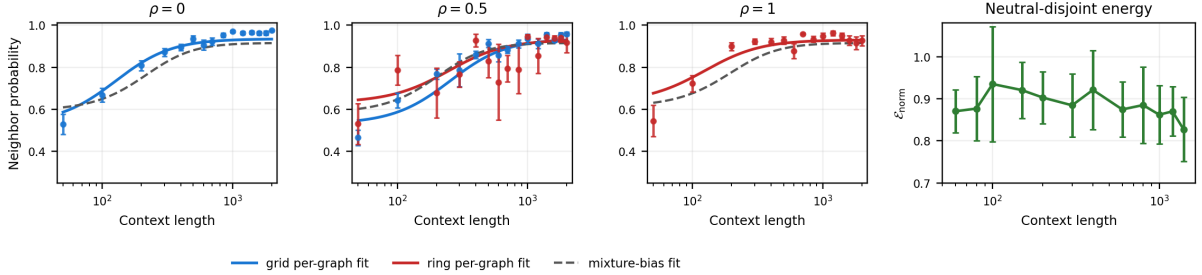


Figure 1: Saved-result diagnostics. Left: representative neutral-disjoint accuracy curves (points show mean $\pm$ SEM over walks), the per-graph fit (colored solid lines), and the shared mixture-bias ablation (gray dashed). Right: normalized Dirichlet energy at twelve context lengths (mean $\pm$ SD over eight walks).

5.2 Residual-Stream Geometry Results

The behavioral results suggest the LLM maintains a complexity-sensitive structural bias. But do those behavioral signatures have a correlate inside the model? Near N^* ($T = 200$), the PC1/PC2 plane shows only partial ring structure; by $T = 1400$ the ring topology is visually recoverable in the low-dimensional geometry. The right panel of Figure 1 shows that the cached eight-walk analysis provides twelve energy measurements, not only two endpoints. The series is noisy but modestly decreasing overall: 0.903 ± 0.062 at $T = 200$, 0.885 ± 0.090 at $T = 800$, and 0.828 ± 0.076 at $T = 1400$ (mean $\pm$ SD across walks). Six of eight per-walk slopes against $\log T$ are negative; the mean slope is -0.014 ± 0.005 (mean $\pm$ SE). This supports, at most, a weak downward trend and does not justify treating energy as decisive evidence.

The competing-structures regime provides a complementary visualization. Figure 2 shows class-mean PCA at $T = 1400$ across the full ρ -ladder in a secondary closed-vocabulary experiment where grid and ring share the same 16-token vocabulary (Section 3.4’s neutral condition). At $\rho = 0.5$, both graphs’ edge sets are partially, but not fully, recoverable when overlaid on different PC projections: as the quantitative principal-angle analysis below shows, the pooled mixture subspace at $\rho = 0.5$ sits at an intermediate distance from both pure references rather than cleanly matching either one, so this panel should be read as evidence of partial, not

full, joint recoverability.

We quantify this geometry by rerunning layer 26 for 16 seeds per cell and using the pure $\rho = 0$ grid and $\rho = 1$ ring subspaces as references. For each mixture, we form the paper-aligned class-mean matrix over positions $[1350, 1400]$, extract its top $k \in \{2, 3, 4\}$ PCs, and compute principal angles $\theta_i = \arccos \sigma_i(U_\rho U_{\text{ref}}^\top)$. Table 2 reports the mean of the four angles. In the controlled closed-vocabulary condition, alignment shifts smoothly from grid to ring: the grid/ring means are $31.6^\circ/46.7^\circ$ at $\rho = .2$, $41.6^\circ/44.8^\circ$ at $\rho = .5$, and $52.5^\circ/24.8^\circ$ at $\rho = .8$. At $\rho = .5$, the individual angles are $(15.3, 25.5, 36.0, 89.7)^\circ$ to grid and $(20.4, 26.7, 44.3, 87.9)^\circ$ to ring; bootstrap 95% intervals for their means are $[39.1, 46.0]^\circ$ and $[43.0, 47.7]^\circ$. Both mean angles at $\rho = 0.5$ (roughly $42\text{--}45^\circ$) sit well short of the 0° that would indicate full recovery of either pure reference and well short of the $59\text{--}60^\circ$ orthogonal-looking angle observed at the opposite pure endpoint, so we describe the grid at $\rho = 0.5$ as *partially* recoverable rather than as fully recovered. A full-prefix sensitivity analysis gives the same pattern. Thus the quantitative result supports partial alignment with both pure references, not two mutually orthogonal subspaces. The neutral-disjoint values are higher but confounded by the distinct token vocabularies, so we do not interpret them as structural orthogonality. A transition cache can also induce graph-aligned geometry; we therefore treat these

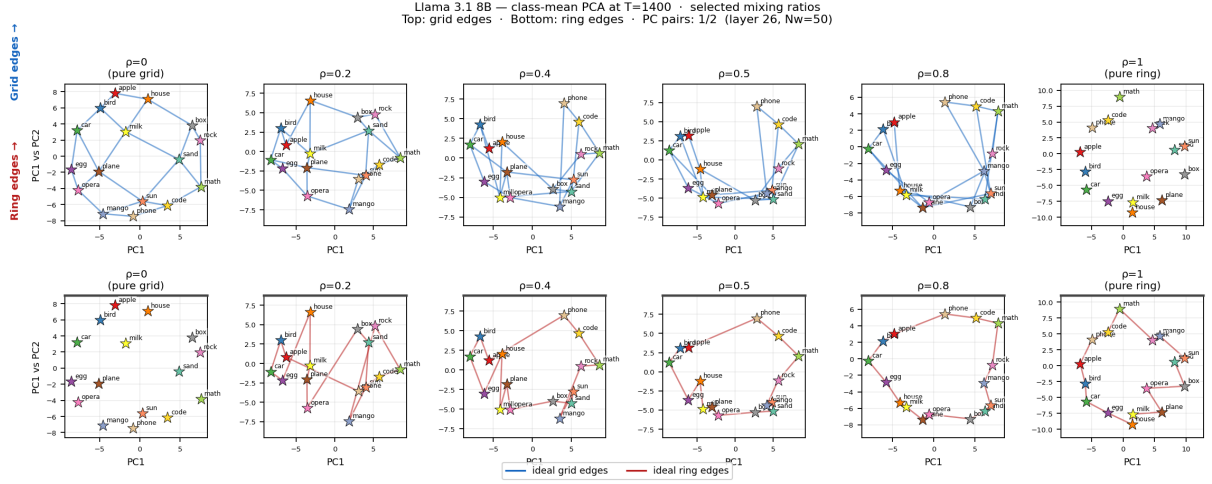


Figure 2: Class-mean PCA at $T = 1400$ across mixture ratios ρ . The top row overlays grid edges in blue; the bottom row overlays ring edges in red.

Table 2: Mean top-4 principal angle (degrees) from each pooled mixture subspace to the pure grid/ring references. Lower values indicate closer subspace alignment. End-point self-comparisons are zero by definition.

ρ	Closed vocab. G/R	Neutral-disjoint G/R
0.0	0.0/59.1	0.0/78.9
0.2	31.6/46.7	54.3/55.8
0.4	40.9/46.5	58.7/51.1
0.5	41.6/44.8	57.6/49.5
0.8	52.5/24.8	69.2/47.6
1.0	59.1/0.0	78.9/0.0

results as evidence of recoverable topology, not as evidence that rules out induction.

6 Experiment 3

6.1 Does Graph-Family Information Causally Control Next-Token Predictions?

The behavioral and PCA analyses are correlational: they show that outputs and representations are consistent with latent structure inference, but not that the relevant residual-stream information is used by the final prediction. We therefore run two causal interventions, following the activation-patching and activation-steering logic used in mechanistic interpretability and representation engineering (Meng et al., 2022; Turner et al., 2023; Zou et al., 2023).

For clean graph G_c , corrupt graph G_r , and final token x_t , we score each logit vector with a graph-family contrast

$$\Delta(x_t) = \frac{1}{|\mathcal{N}_{G_c}(x_t)|} \sum_{w \in \mathcal{N}_{G_c}(x_t)} z_w - \frac{1}{|\mathcal{N}_{G_r}(x_t)|} \sum_{w \in \mathcal{N}_{G_r}(x_t)} z_w \quad (12)$$

where $\mathcal{N}_G(x_t)$ denotes the set of valid next-node neighbors of x_t under graph G , and z_w is the next-token logit for word w . We generate matched clean/corrupt prompt pairs from the grid and ring that end at the same current node (both drawn from the disjoint-vocabulary condition, so grid and ring node tokens never overlap). For activation patching, we cache clean residual activations and rerun the corrupt prompt while replacing the final-position `hook_resid_post` activation after block ℓ . The normalized patch effect is

$$E_{\text{patch}}(\ell) = \frac{\Delta_{\text{patch}}(\ell) - \Delta_{\text{corrupt}}}{\Delta_{\text{clean}} - \Delta_{\text{corrupt}}}. \quad (13)$$

Thus 0 means no movement from the corrupt graph preference, and 1 means the patched corrupt run recovers the clean run’s graph preference.

For steering, we compute a layer-wise graph-difference vector on disjoint training contexts,

$$v_\ell = \mathbb{E}[h_\ell(x_t) \mid G_c = \text{grid}] - \mathbb{E}[h_\ell(x_t) \mid G_r = \text{ring}] \quad (14)$$

and add αv_ℓ to final-position residual activations in held-out ring contexts. We compare the real vector to two controls: a Gaussian random vector matched to the real vector’s norm and a shuffled-label vector computed after permuting graph labels in the steering-vector training set. Full protocol details, including the seen/held-out edge split used to test transition-cache explanations, are given below.

6.2 Prompt Pair Construction and Metrics

Clean and corrupt prompts are generated from different graph families but end at the same current

node. Because the graph hypotheses are undirected, we sample a random walk ending at the desired final node by generating a valid walk from that node and reversing it. The reversed walk has the same graph support, and the model is always evaluated at the final position, predicting the next graph word.

The primary score is the graph-family logit contrast in Equation 12. For patching, normalized effect is defined by Equation 13. For steering, normalized effect is computed analogously, replacing Δ_{patch} with the steered metric and using the target and source prompt metrics as endpoints. Rows with small denominators are marked unusable in the raw JSONL; none were excluded in the reported patching runs.

6.3 Seen and Held-Out Edge Split

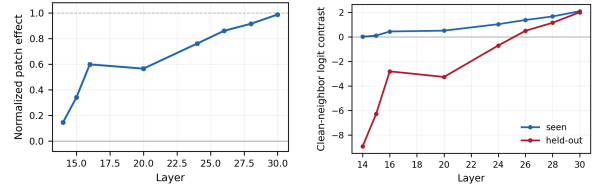
Following the locality diagnostic in Section 2.1, for each final token we split graph neighbors according to whether the edge incident to the final token was observed in the evaluation context. The “seen” set contains true graph neighbors of the final token whose edge appeared in either direction. The “held-out” set contains true graph neighbors whose edge did not appear. For clean/corrupt patching, this split is computed using the corrupt context, so the diagnostic asks whether a clean activation intervention helps graph-neighbor logits that the corrupt prompt did not locally observe.

6.4 Causal Activation Intervention Results

Using the clean/corrupt prompt-pair protocol from Section 6.1, we intervene on final-position residual-stream activations and score the resulting next-token distribution with the graph-family logit contrast in Eq. 12.

At context length $T = 1400$, final-token residual patching rises rapidly across the selected layer sweep, establishing that late residual-stream states causally control the graph-neighbor logit contrast rather than merely encoding decodable graph information. The held-out diagnostic is delayed relative to the seen-edge effect, crosses zero by layer 26, and reaches 2.0 at layer 30. Thus patched activations boost graph-consistent predictions for edges absent from the corrupt prompt. This does not imply that the donor state generalized to those edges: the clean prompt itself contains clean-graph transitions.

To calibrate that distinction, we run the directed transition-cache baseline defined in Section 2.1 on the same 200 prompt pairs. Its patching analogue



(a) Final-token patching at $T = 1400$. (b) Seen/held-out split at $T = 1400$.

Figure 3: Long-context activation-patching diagnostics. The selected-layer run shows late-layer recovery and held-out graph-neighbor logits becoming positive in late layers.

replaces the corrupt count row with the clean row. This replacement has normalized recovery 1.0 by construction and moves the held-out contrast from -6.069 ± 0.016 to 1.260 ± 0.099 (mean $\pm$ SE). Therefore held-out clean-to-corrupt transfer is compatible with a local cache. The LLM patching result establishes causal use of graph-family information, but this particular split does not identify how that information was computed.

Figure 3 shows the layerwise and held-out patching curves; selected aggregates appear in Table 3.

Steering provides a complementary lower-bandwidth causal intervention. Adding the grid-minus-ring direction from Equation 14 to held-out ring contexts recovers 0.449 ± 0.004 of the clean-corrupt graph contrast at $\alpha = 5$, while negative α reverses the effect and both random norm-matched and shuffled-label controls remain near zero. The effect also strengthens with layer (Figure 4). Because steering uses a single global direction rather than pair-specific activation replacement, it does not fully reproduce patching; held-out edge-specific logits remain substantially harder to steer. Together, these interventions support the existence of a manipulable graph-family representation that contributes causally to next-token prediction, without uniquely identifying a structure-inference mechanism.

6.4.1 Activation Patching

The patching run reported in the paper uses Llama-3.1-8B, 200 grid/ring clean-corrupt prompt pairs, context length $T = 1400$, final-position `hook_resid_post` activations, and the selected layer set $\{14, 15, 16, 20, 24, 26, 28, 30\}$.

6.4.2 Activation Steering

The steering run uses disjoint train and evaluation contexts. We compute grid-minus-ring

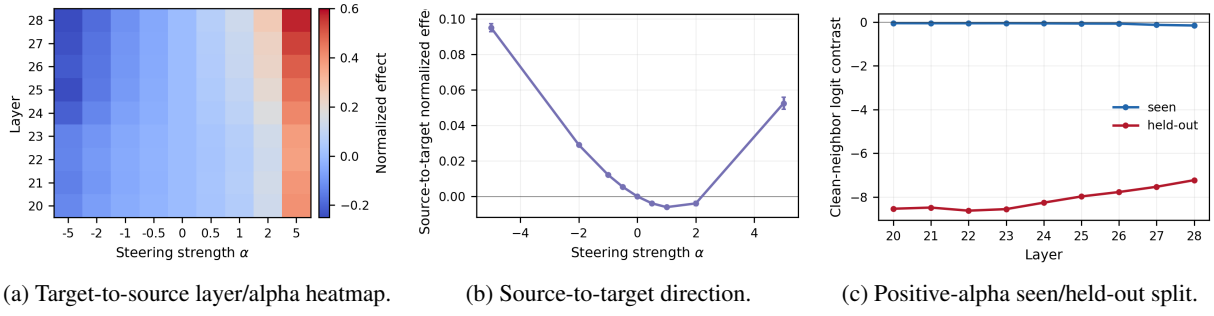


Figure 4: Additional steering diagnostics. Steering is strongest for target-to-source grid-minus-ring additions in late layers. The reverse direction is weaker, and held-out edge-specific logits remain difficult to move with a single global vector.

Table 3: Selected activation-patching aggregates. Effects are mean $\pm$ standard error over prompt pairs.

Context	Layer	n	Normalized effect	Held-out contrast
1400	16	200	0.598 ± 0.013	-2.814
1400	26	200	0.860 ± 0.008	0.492
1400	30	200	0.987 ± 0.001	2.000

vectors from 1000 training contexts per graph at $T = 1400$, then evaluate on 500 held-out grid/ring prompt pairs at layers 20–28 and $\alpha \in \{-5, -2, -1, -0.5, 0, 0.5, 1, 2, 5\}$. The raw JSONL contains duplicate rows from a prior append; all reported steering aggregates and paper figures deduplicate by the complete intervention key, keeping the first occurrence (`pair_id`, `layer`, `$\alpha$` , `control`, `evaluation direction`), yielding the expected 243,000 unique interventions.

6.4.3 Interpretation

Activation patching is a high-bandwidth, pair-specific intervention: it replaces the corrupt residual state with the clean residual state at one layer and position. Steering is a low-bandwidth, population-level intervention: it adds one global graph-difference vector to many contexts. The large gap between near-complete patching recovery and partial steering recovery is therefore not a contradiction. It suggests that the residual stream contains both a broad graph-family direction and finer edge- or context-specific information that a single steering vector does not capture.

7 Discussion & Conclusion

The behavioral, representational, and causal evidence converge on the same picture: Llama-3.1-8B is not well described as either a pure Bayesian structure learner or a pure induction-cache machine. Topology-specific sigmoids describe the behavioral

curves better than the shared mixture-bias ablation, graph-aligned geometry is recoverable from the residual stream, and patching and steering show that graph-family information is causally used by the output computation.

At the same time, the new controls limit what these results establish. The cache recovers a similar fitted complexity coefficient and reproduces held-out-edge transfer when its clean state replaces its corrupt state, while the principal-angle analysis favors interpolation between pure references over orthogonal coexistence. The present protocol therefore does not rule out induction. A dual-mechanism account remains a plausible interpretation, but not an identified conclusion.

We also acknowledge two scope limitations. First, all experiments use a single model family on a synthetic random-walk task; we do not know whether the behavioral, representational, and causal signatures reported here generalize to more realistic reasoning benchmarks, or other model scales and families, and we treat that generalization as an open empirical question rather than an assumption. Second, a natural follow-up is to repeat the patching, steering, and cache-control comparisons at a range of intermediate T to identify whether, and at what context length, local-copying explanations become insufficient on their own. Immediate next steps are a direct graph-identity probe, a degree-matched graph comparison, and a test of genuinely unseen edges without importing them from a clean donor context.

References

- Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, and 23 others. 2024. [Foundational challenges in assuring alignment and safety of large language models](#). *Preprint*, arXiv:2404.09932.
- Andy Arditi. 2026. [In-context learning of representations can be explained by induction circuits](#). LessWrong. Crosspost of ICLR 2026 Blogpost Track post.
- Eric Bigelow, Daniel Wurgaft, YingQiao Wang, Noah Goodman, Tomer Ullman, Hidenori Tanaka, and Ekdeep Singh Lubana. 2025. [Belief dynamics reveal the dual nature of in-context learning and activation steering](#). *Preprint*, arXiv:2511.00617.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A survey on in-context learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.
- Chaeha Kim. 2025. [Task schema and binding: A double dissociation study of in-context learning](#). *Preprint*, arXiv:2512.17325.
- Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. [The unlocking spell on base llms: Rethinking alignment via in-context learning](#). *Preprint*, arXiv:2312.01552.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.
- Neel Nanda and Joseph Bloom. 2022. Transformerlens. <https://github.com/TransformerLensOrg/TransformerLens>.
- 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, and 7 others. 2022. [In-context learning and induction heads](#). *Preprint*, arXiv:2209.11895.
- Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. 2025. [Iclr: In-context learning of representations](#). *Preprint*, arXiv:2501.00070.
- Alex Polyakov and Daniel Kuznetsov. 2026. [Involuntary in-context learning: Exploiting few-shot pattern completion to bypass safety alignment in gpt-5.4](#). *Preprint*, arXiv:2604.19461.
- Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. 2024. [In-context learning with iterative demonstration selection](#). *Preprint*, arXiv:2310.09881.
- Jack Ransome. 2026. [In context learning representations, byproduct or mechanism?](#)
- Aaditya K. Singh, Ted Moskovitz, Felix Hill, Stephanie C. Y. Chan, and Andrew M. Saxe. 2024. [What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation](#). *Preprint*, arXiv:2404.07129.
- Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. [Activation addition: Steering language models without optimization](#). *Preprint*, arXiv:2308.10248.
- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. [An explanation of in-context learning as implicit bayesian inference](#). *Preprint*, arXiv:2111.02080.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. [Representation engineering: A top-down approach to AI transparency](#). *Preprint*, arXiv:2310.01405.