Rule-Dependent Neighbor Routing in Transformers Trained on Life-like Cellular Automata

Anonymous authors
Paper under double-blind review

Abstract

Mechanistic interpretability benefits from learned-model testbeds whose task semantics, local dependencies, and canonical sufficient statistics are exactly specified. We train small transformer encoders to predict one-step updates of the Conway, Seeds, and DayNight cellular automata on 16×16 toroidal grids. Across three token-ordering conditions, three seeds, and a fixed non-grid graph control, the models reach near-perfect accuracy and depend causally on true neighbors rather than matched non-neighbor controls. Comparing the trained models under a fixed architecture, data distribution, and analysis pipeline, we find rule-associated differences in the layer distribution of neighbor-attending heads: routing is concentrated early for Seeds, late for DayNight, and seed-sensitive for Conway, and a head's neighbor-attention mass correlates strongly with its single-head ablation damage. In Seeds, the predicate `dead & count == 2` is perfectly linearly decodable on dead cells, where it determines the output — and remains so after removing the explicit readout direction — but only partially on alive cells, while a uniformly faithful 0..8 neighbor count is not recovered under the same probes. Selected-head case studies show that the learned within-neighbor attention weighting is functionally important to the model's output, and DayNight additionally preserves complement-related representational geometry across layers. These results establish Life-like cellular automata as a controlled comparative testbed for learned routing mechanisms, while leaving the complete decision circuit and the causal role of specific rule properties open.[1]

1 Motivation

Cellular automata are useful mechanistic testbeds because the task semantics are local, discrete, and exactly specified (Gardner, 1970; Wolfram, 1983; Eppstein, 2010): the input–output function, the local dependency range, and a canonical high-level decomposition are all known, even though a rule does not fix a *unique* implementation (it can be realized by an explicit 0–8 count, by recognizing a few count-equality predicates, by a 512-entry lookup over local patterns, or by distributed boolean features). Complementary to compiled-model testbeds that hand-construct a transformer with known internal structure (Lindner et al., 2023), we study *learned* models with exact task-level ground truth. A one-step Life-like update decomposes into:

1. routing information from graph/grid neighbors,
2. computing a neighbor-count or rule-specific predicate,
3. applying a birth/survival decision conditioned on the current cell state.

The central question is not only whether a transformer can learn the mapping, but whether its internal organization reflects these algorithmic components — and whether *different rules induce different organizations*.

[1] Large language models were used as assistants for drafting and editing the manuscript and for implementing analysis and plotting scripts. The authors designed the study and experiments, verified every reported analysis, and are solely responsible for all scientific claims and the final text; the complete analysis code, the aggregate result tables, and the trained checkpoints needed to reproduce the case-study analyses are provided in the anonymous supplement.

Two properties make this setting unusually clean for mechanistic work:

- Every canonical high-level quantity (neighbor identity, count, predicate, transition type) is exactly computable, so probes and ablations have a well-defined reference — while recognizing that these are natural sufficient statistics rather than intermediate variables the model must explicitly realize.
- The model is small enough (16 heads) for exhaustive causal sweeps, avoiding the sampling and search compromises of large-model interpretability.

Figure 1 summarizes the task, the abstract algorithm it decomposes into, and the central finding that the learned neighbor-routing organization shifts with the rule.

Contributions.

- We introduce Life-like cellular automata as a mechanistic-interpretability testbed with exactly specified task semantics and canonical sufficient statistics for every local computation, and make the local update rule the experimental variable while architecture, data distribution, and analysis pipeline are held fixed (Section 3.3).
- We show, via exhaustive single-head ablations over all 16 heads of every run, that neighbor routing concentrates in a small set of heads whose causal depth is rule-dependent (early for Seeds, late for DayNight, seed-sensitive for Conway), and that a head's neighbor-attention mass correlates strongly with its single-head ablation damage (Section 4.4).
- We show that in Seeds the rule-sufficient predicate is linearly decodable exactly where it is output-relevant, using state-conditioned probes that control for the predicate being the output label (Section 4.3).
- We provide converging path-level and representational evidence: a per-rule QK/OV case study consistent with weighted neighbor-content aggregation (Section 4.7), and a complement-symmetry intervention showing DayNight preserves complement-related representational geometry that survives a cell-identity control (Section 4.8).

Throughout, we label each claim by support level and state its scope limits explicitly (Appendix A).

2 Related Work

Cellular automata are a standard setting for studying local computation, starting with Conway's Game of Life (Gardner, 1970) and the broader Life-like B/S rule family (Eppstein, 2010). Prior neural work has largely asked whether networks can learn CA dynamics or use CA-like dynamics as a generative substrate. Springer & Kenyon (2021) show that even one-step Game-of-Life prediction can be surprisingly hard for small convolutional networks, even though convolutional architectures can represent arbitrary CA exactly (Gilpin, 2019); this difficulty result has recently been revisited with alternative activation functions (Ahmed & Davis, 2026). Neural cellular automata instead train local update rules to produce target morphology or behavior (Mordvintsev et al., 2020). Recent transformer work studies behavioral learning and generalization in CA-like domains: LifeGPT trains a topology-agnostic generative transformer on Game-of-Life dynamics (Berkovich & Buehler, 2025), its follow-up AutomataGPT extends the same line to forecasting and rule inference across 100 two-dimensional CA rules (Berkovich et al., 2025), and Burtsev (2024) studies transformers trained on one-dimensional elementary cellular automata. These study behavioral prediction and rule inference; our goal is instead *mechanistic* — we use Life-like rules as a controlled analysis testbed with exact task-level ground truth, where every neighbor, count, predicate, and transition label is known exactly, and we make the *rule* the experimental variable.

This paper is also related to work on learned world models in synthetic tasks, especially Othello-GPT (Li et al., 2023) and the follow-up analysis showing its world state is linearly represented (Nanda et al., 2023b);

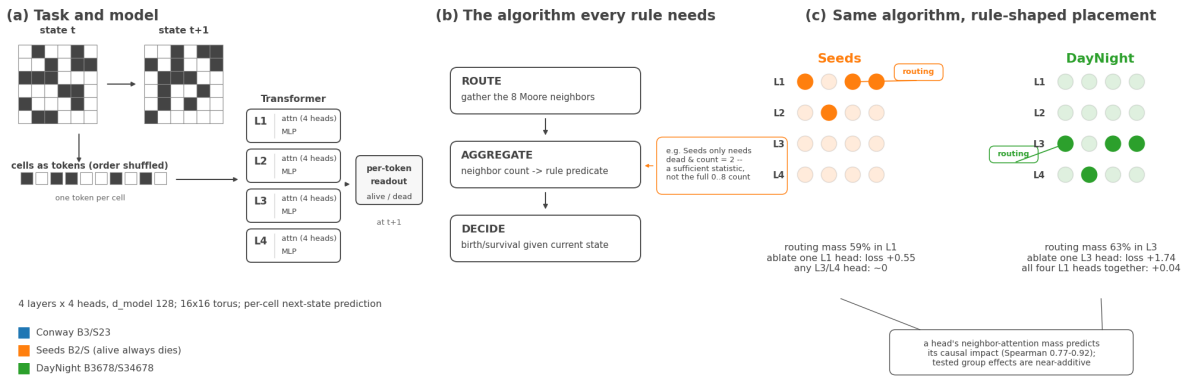

Figure 1: Overview of the testbed. A transformer encoder predicts the one-step Life-like update of each cell from its current board. The update decomposes into neighbor routing, predicate computation, and a state-conditioned birth/survival decision. Holding architecture and data fixed while varying the rule, we find that the depth of the learned neighbor-routing heads changes with the rule.

later work has strengthened and refined this line (Yuan & Søgaard, 2025; Singh et al., 2025). We use a similar probing logic, but hold the architecture, board distribution, and analysis pipeline fixed while the local update rule changes. This lets us ask whether different rules induce different internal organizations rather than only whether one latent world state can be decoded.

Mechanistic studies of algorithmic tasks provide the closest methodological precedent: modular-addition and grokking circuits (Power et al., 2022; Nanda et al., 2023a), induction heads (Olsson et al., 2022), and circuit-level analyses such as indirect object identification (Wang et al., 2023). Relatedly, the same algorithmic task can admit qualitatively different learned circuits across training runs (Zhong et al., 2023), which anticipates the seed-sensitivity we observe for Conway. We follow that tradition but choose a task family with explicit spatial ground truth and multiple related rules, making circuit organization comparable across rules without changing the architecture. Our evidence is correlational and ablational rather than a full causal-abstraction analysis: the causal-abstraction framework tests whether a network representation *realizes* a high-level variable through interchange interventions that swap the candidate representation between inputs and check the high-level counterfactual (Geiger et al., 2021); the exact local structure of our setting makes it well suited to such interventions, which we leave to future work (Section 7).

Finally, our per-head analysis builds on attention-head ablation and pruning work showing that a small number of specialized heads can carry much of a model's behavior (Michel et al., 2019; Voita et al., 2019), on transformer circuits work that treats attention heads as additive read/write components in the residual stream (Elhage et al., 2021), and on methodological guidance for activation-level interventions (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024; Meng et al., 2022). We use exhaustive single-head ablations because our models have only 16 heads, and pair these causal interventions with directly measured neighbor-attention mass. Probing methodology follows linear classifier probes (Alain & Bengio, 2016) with the standard caveats about probe expressivity and controls (Hewitt & Liang, 2019; Belinkov, 2022).

3 Experimental Setup

3.1 Tasks and data

We study three Life-like rules on toroidal grids:

- Conway: B3/S23
- Seeds: B2/S (alive cells always die)

- DayNight: B3678/S34678

Each training example is generated on the fly. For each board, the alive-cell density is sampled uniformly from $[0.05, 0.95]$, cells are sampled independently at that density, and the target is the exact one-step Life-like update. As a control for grid geometry, we also train on a fixed random graph: 256 nodes of degree 8, with the same birth/survival semantics applied to graph neighbors. Nodes are presented in a fixed order and identified through node-identity embeddings. Training and evaluation use the same sampled graph, so this setting is a **same-graph, non-grid control** against grid/tokenization artifacts; it does not test unseen-graph generalization.

3.2 Models and training

The main 16×16 models are 4-layer, 4-head transformer encoders (Vaswani et al., 2017) with `d_model=128`, `d_ff=512`, dropout 0, AdamW, learning rate `3e-4`, weight decay `0.01`, batch size 512, and a binary per-token next-state readout.

Each cell is one token. Every token embedding is the sum of a state embedding (alive/dead) and a learned cell-identity embedding; the cell-identity embedding is what makes cell identity recoverable regardless of token order, including under per-example shuffling. We study three **token-ordering / identity-encoding conditions** (all sharing the cell-identity embedding), which differ in token order and in whether an additional sequence-position embedding is used:

- `rowmajor`: deterministic row-major order; adds a sequence-position embedding;
- `fixed_permutation`: one fixed random permutation saved with the run; adds a sequence-position embedding;
- `shuffled_set`: a fresh random order per example; no sequence-position embedding, so cell identity is available only through the cell-ID embedding.

Because a cell's identity is fixed, in `rowmajor` and `fixed_permutation` the added sequence position is in one-to-one correspondence with the cell-ID embedding, so these are three encoding conditions on the same task rather than three independent tokenization inductive biases; we therefore describe them as token-ordering conditions and do not interpret them as a test of topology abstraction. Random-graph runs likewise use node-identity embeddings without a sequence-position embedding.

The 16×16 matrix covers rules × tokenizations × seeds 42/123/456, plus fixed random-graph runs for the same rules and seeds. All 16×16 runs analyzed in this paper trained for 50k steps. Evaluation uses fresh on-the-fly boards: each reported one-step metric is averaged over 100 batches of 512 boards (51,200 boards, i.e. ≈13.1M cell predictions per run), and the 3-step rollout metric over 20 batches (10,240 boards). Reported board accuracy is exact-board accuracy over all 256 cells, so it is much stricter than cell accuracy (e.g., a board accuracy of 0.999980 corresponds to exactly one incorrect board in 51,200).

3.3 Mechanistic analyses

The analysis suite includes eval / 3-step rollout accuracy, transition-conditioned metrics, layerwise linear probes for neighbor count, transition class, and rule-specific predicates, attention neighbor enrichment, layerwise neighbor ablation, edge-mass matched ablation, logit/loss causal metrics, and patching recovery. Standard probes use independent generated train/test batches (linear probing methodology and its caveats: Alain & Bengio, 2016; Hewitt & Liang, 2019; Belinkov, 2022). State-conditioned Seeds probes split cells by current state (alive/dead), use class-balanced subsampling where needed, and report balanced or macro accuracy.

The per-head analysis is exhaustive because each model has only 16 attention heads (cf. Michel et al., 2019; Voita et al., 2019). Model weights are frozen throughout. For every head we measure neighbor-attention mass and perform single-head ablation by removing that head's additive output contribution (zeroing its

`out_proj` columns, following the residual-stream view of heads as additive components: Elhage et al., 2021; the mean-ablation control instead substitutes the head's dataset-mean contribution, guarding against zero-ablation overstating effects: Zhang & Nanda, 2024; Parodi et al., 2026). The **primary causal metric** is the damage to output loss and cell accuracy on identical held-fixed boards. As a **secondary representation diagnostic**, we retrain linear predicate probes on the ablated model's activations for each condition; this measures whether the predicate remains linearly recoverable, not model behavior. Group ablations cover top-2, top-3, and all-L1 groups, plus a pairwise sweep over the top-5 neighbor-mass heads of every run ($\binom{5}{2} = 10$ pairs per run). These analyses test local redundancy/additivity around the most causally important heads but do not constitute a full pairwise interaction map over all 16 heads.

For the top routing head of one run per rule we additionally perform a QK/OV case study using a manually reimplemented forward pass (verified against the module forward to a maximum absolute logit difference below 10^{-3}), substituting only the target head's attention pattern while all other heads and all value/OV streams keep their original activations. The substitution conditions are: (i) the original pattern (baseline); (ii) *uniform-within-neighbors*: the per-query distribution over the 8 true neighbor tokens is flattened while preserving the per-query total neighbor mass and the off-neighbor pattern; (iii) *ideal-neighbor template*: uniform 1/8 over the true neighbors with total mass 1; (iv) *uniform-over-all* tokens; and (v) *row-shuffled*: the learned pattern with key positions permuted. On the OV side we fit a ridge regression predicting a *scalar* content target (the attention-weighted alive fraction of the attended tokens) from the head's d_{model}-dimensional per-token output contribution (its post-`out_proj` residual write), and report the single-output R^2 on a 30% held-out split. This is a path-level diagnostic, not a full circuit identification.

Three quantities used throughout Section 4.4 are defined as follows. **Neighbor-attention mass** of a head is the attention weight the query token assigns to its 8 true (Moore or graph) neighbor tokens, summed over the neighbor set and averaged over queries and batches; each attention row sums to 1, so mass lies in $[0, 1]$, and the neighbor set excludes the query cell itself. **Δloss** is the increase, relative to the un-ablated baseline, of the mean per-token binary cross-entropy (with logits) on identical held-fixed boards. The **causal head layer center** of a single run is the positive-part weighted center of mass over that run's 16 heads,

$$\text{center} = \frac{\sum_h \text{layer}(h)\ \max(\Delta\text{loss}_h, 0)}{\sum_h \max(\Delta\text{loss}_h, 0)},$$

computed *per run* so that the run — not the individual head, since heads nested in one run are not independent — is the unit of analysis. We then report the mean and a 95% CI over the 12 runs of each rule, matched-pair comparisons across rules (holding tokenization, seed, and topology fixed), and a main-effect variance decomposition over the balanced $3 \times 3 \times 3$ grid design (Section 4.4). All cross-rule layer-center, paired-comparison, and variance-decomposition results use **zero-ablation for all three rules**; a mean-ablation sweep on all three rules is reported as a sensitivity analysis and gives essentially the same centers, so the ordering does not depend on the ablation type. The 95% CIs are Student-t intervals over the per-rule run centers; the variance decomposition is a main-effect sum-of-squares partition on the balanced grid design.

3.4 32×32 scaling

The 32×32 scaling runs use fixed-permutation tokenization. Conway seeds 42/123/456 are continued to 200k steps; Seeds and DayNight seed42 are continued to 300k. Converged seed42 runs receive a reduced mechanism analysis. These runs test whether the earlier 50k failures were training-budget effects; they are not a seed-robust non-Conway 32×32 mechanism matrix.

4 Main Results

4.1 Models learn one-step Life-like updates

Table 1 reports one-step accuracy for all completed 16×16 runs. Every run reaches cell accuracy above 0.999999, and strict exact-board and 3-step rollout accuracy stay within 6×10^{-5} of perfect in the worst run.

The fixed random-graph control uses the same sampled graph for training and evaluation with node-identity embeddings, so it argues against purely grid-specific shortcuts but cannot exclude memorization of the

Table 1: One-step performance across the completed 16×16 matrix. Cell/board/rollout values are mean (min) over runs, 6 decimal places; board accuracy is strict exact-match over all 256 cells. Values shown as 1.000000 are exact at 6-decimal rounding but not necessarily exactly 1.

rule	setting	n runs	cell acc	exact-board acc	3-step rollout acc
Conway	grid, 3 tok. × 3 seeds	9	1.000000 (1.000000)	0.999993 (0.999980)	0.999999 (0.999998)
Conway	fixed random-graph control	3	1.000000 (1.000000)	0.999993 (0.999980)	0.999993 (0.999980)
Seeds	grid, 3 tok. × 3 seeds	9	1.000000 (1.000000)	0.999996 (0.999980)	1.000000 (0.999999)
Seeds	fixed random-graph control	3	1.000000 (1.000000)	1.000000 (1.000000)	1.000000 (1.000000)
DayNight	grid, 3 tok. × 3 seeds	9	1.000000 (1.000000)	0.999978 (0.999941)	0.999995 (0.999987)
DayNight	fixed random-graph control	3	1.000000 (1.000000)	1.000000 (1.000000)	0.999996 (0.999992)

fixed adjacency structure, and it does not test unseen-graph generalization. A forward-only long-rollout / density-OOD sanity check (horizon 10, shuffled-set checkpoints) stays near these ceilings and is reported in Appendix B.

4.2 True-neighbor causal dependence persists

Across tokenizations and the fixed random-graph control, attention enrichment, layerwise neighbor ablation, edge-mass matched ablation, and logit/loss causal metrics all show larger effects for true neighbors than for matched non-neighbor controls (Table 2). Ablating the block of true neighbors increases loss sharply, while ablating an equal-size block of matched non-neighbor (random) tokens leaves loss essentially unchanged (Δloss ≈ 0); the edge-mass intervention shows the same true-vs-far separation. The gap holds for every rule and, critically, persists on the fixed random graph, where there is no grid geometry to exploit. This reduces the risk that the models merely exploit a row/column tokenization artifact or a fixed grid geometry, though it does not establish abstract graph-adjacency generalization.

4.3 Seeds linearly exposes the rule's sufficient statistic

The rule predicate `seeds_birth_count2` (= `dead & count_eq_2`) reaches best balanced accuracy 1.0000 across all 12 completed 16×16 Seeds runs. The state-conditioned probes address the main caveat that this predicate is also the output label for Seeds. Averaged over 12 runs (Table 3):

Thus the same count-equality information is linearly exposed where the rule uses it and remains incomplete where it is output-irrelevant (L4 gap +0.33). A uniformly faithful 0..8 neighbor count is **not linearly recovered under our probes at any layer**, and the generic count probe decays in late layers (Figure 2). This supports the interpretation that Seeds representations are shaped by the rule's information demand: the model linearly exposes the sufficient statistic rather than observably maintaining a generic full count. The 32×32 seed42 sanity run shows the same pattern (dead-cell L4 = 1.0, alive-cell L4 = 0.74 and decaying). This is linear-decodability evidence with a state-conditioned control: it shows the sufficient statistic is *linearly exposed* where the rule uses it, but it does not establish that the model *causally uses* this particular direction in its computation (that would require an interchange intervention), nor is it proof that inaccessible nonlinear information is absent.

Fair-comparison and readout controls. Two controls sharpen this result (Seeds shuffled_set, seeds 42/123/456; dead cells at L4). First, to avoid comparing a binary predicate against a 9-class count, we train the *same* class-balanced binary `count_eq_`k probe for every $k \in \{0, \ldots, 8\}$: `count_eq_2` is highest at 1.000, the other sparse counts are also high (0.99/0.98/0.98 for $k = 0/1/3$), and the dense counts fall off (0.81–0.91 for $k \geq 4$). On equal footing the decision-relevant predicate is thus the sharpest, though low-count predicates are broadly decodable. Second, the L4 = 1.000 is *not* merely a re-read of the output decision: projecting out the L4 direction that best linearly predicts the model's own logit leaves `count_eq_2` at 1.000 (a prevalence-matched random-boolean predicate sits at chance, 0.50), so the predicate is encoded beyond the readout axis.

Table 2: True-neighbor vs. matched-control causal effects (peak-layer loss increase; random graph is mean over 3 seeds, grid is mean (min) over 9 runs). Ablating true neighbors damages the model while an equal-size matched control block does not, and the edge-mass true-vs-far intervention agrees. The separation persists on the fixed random graph, where no grid geometry is available.

	fixed random graph (mean / 3 seeds)					grid
rule	block-nbr	block-rnd	block gap	edge-true	edge-far	block gap, mean (min)
seeds	1.81	0.00	1.81	0.62	0.00	1.52 (1.22)
conway	2.83	0.00	2.83	0.76	0.00	2.31 (2.05)
daynight	3.90	0.00	3.90	1.50	0.00	3.73 (3.39)

Table 3: State-conditioned Seeds probes (balanced accuracy, mean over 12 runs). The count-equality predicate is exposed at L4 exactly where the rule uses it (dead cells) but stays incomplete where it is output-irrelevant (alive cells), and a uniformly faithful 0..8 count is never linearly recovered.

probe	scope	emb	L1	L2	L3	L4
count_eq_2	dead (output-relevant)	0.50	0.63	0.83	0.99	**1.000**
count_eq_2	alive (output-irrelevant)	0.50	0.66	0.68	0.73	**0.67** ↓
full 0..8 count	all	0.18	0.53	0.65	0.61	0.59 ↓

4.4 Neighbor routing lives in a few heads, at rule-dependent depth

With model weights frozen, we exhaustively zero-ablate every attention head in every 16×16 run (16 heads × 12 runs × 3 rules) on identical held-fixed boards, scoring each head primarily by output loss / cell-accuracy damage, with retrained linear probes as a secondary representation diagnostic (Section 3.3). The resulting head-level profile is rule-dependent (Table 4, Figure 3):

Routing is concentrated: averaged over runs, the effective number of causally important heads (the inverse participation ratio $1/\sum_h p_h^2$ of the positive-Δloss weights p_h, mean across runs) is ≈ 4 of 16 (Conway 4.0, DayNight 4.3, Seeds 4.4), and the top-3 heads account for $\approx 77\%$ of the total positive single-head ablation damage on average (Conway 78%, DayNight 77%, Seeds 77%). Three points matter:

1. **Observational mass correlates with causal impact.** A head's neighbor-attention mass correlates strongly with the loss increase when that head is ablated (pooled Spearman 0.77–0.92 across rules; per-run zero-ablation Spearman 0.90 ± 0.05 for Conway, 0.90 ± 0.05 for Seeds, 0.75 ± 0.09 for DayNight; Seeds mean-ablation gives 0.91 ± 0.03). For Seeds, the same neighbor-mass score also correlates with downstream predicate-probe damage (Spearman -0.914).

2. **Routing depth changes with the rule.** DayNight contrasts sharply with Seeds: 63% of DayNight neighbor routing mass lies in L3, and ablating all four L1 heads together increases loss by only +0.04, whereas Seeds routing is concentrated in L1. The earlier layer-level "Seeds early / DayNight late" result therefore upgrades to a head-level causal profile.

3. **Top-2 and top-3 head groups are approximately additive on average.** For the top-2 and top-3 groups of every rule, and the all-L1 group of Conway and Seeds, group/sum-of-singles ratios lie in 0.89–1.11. We exclude the DayNight all-L1 group from ratio-based summaries because its summed single-head effect is near zero (each L1 Δloss ≈ 0.0001, sum ≈ 0.0004), making the ratio numerically unstable; its absolute interaction $I_G = \Delta L_G - \sum_h \Delta L_h \approx 0.04$ is small relative to DayNight's dominant L3 effect (1.74) but superadditive relative to the individual L1 effects, so we do not claim additivity for DayNight L1. A pairwise sweep over the top-5 heads of every run (360 pairs) gives mean ratios of 1.04–1.08 per rule (Figure 9, right); a minority of pairs interact (8% subadditive below 0.8, 14% superadditive above 1.25; extremes 0.56/3.21). The tested top-head groups are thus approximately additive on average rather than strictly independent, while the DayNight all-L1 group and a minority of pairs interact; interactions beyond the top-5 heads remain unmeasured.

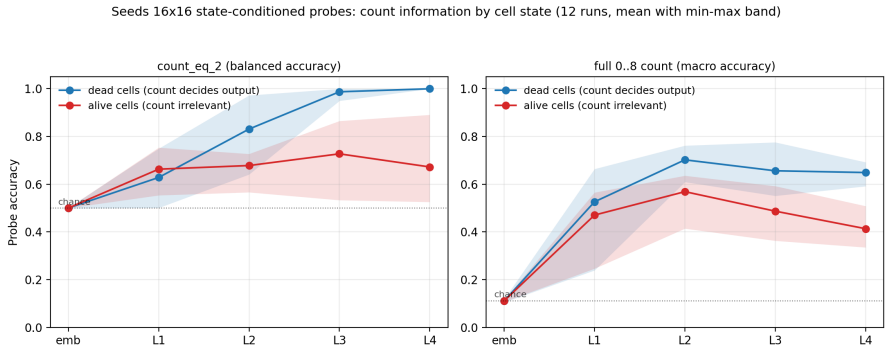

Figure 2: Seeds state-conditioned probes. The rule-sufficient predicate `dead & count_eq_2` becomes linearly exposed by the final layer on output-relevant (dead) cells, while remaining incomplete on output-irrelevant (alive) cells and a generic full neighbor count is not uniformly recovered.

Table 4: Rule-dependent head-level routing profile. The causal head layer center is the run-level mean with a 95% CI over the 12 runs of each rule (Section 3.3); the remaining columns are pooled descriptive statistics. Neighbor-attention mass correlates with single-head causal impact, and the center moves from early layers (Seeds) to late layers (DayNight).

rule	causal head layer center (mean [95% CI])	Spearman(mass, Δloss)	L1 mass share	L1 single-head mean Δloss	L3/L4
seeds	1.31 [1.13, 1.50]	0.918	58.5%	+0.55	≈ 0
conway	1.63 [1.13, 2.13]	0.898	44.9%	+0.72	small
daynight	**3.29 [3.10, 3.48]**	0.773	**3.0%**	+0.0001	L3 +1.74 / L4 +0.71

Run-level statistics. Because heads nested in one training run are not independent, we take the run as the unit of analysis: we compute a causal head layer center *per run* and summarize the 12 runs of each rule (Figure 4). The run-level means match the pooled values (Seeds 1.31 [95% CI $1.13, 1.50$], Conway 1.63 $[1.13, 2.13]$, DayNight 3.29 $[3.10, 3.48]$), so the ordering is not an artifact of pooling. A matched-pair comparison (holding tokenization, seed, and topology fixed) gives DayNight − Seeds = $+1.98$ $[+1.76, +2.19]$, with **all 12/12 pairs positive**; DayNight − Conway = $+1.66$ $[+1.19, +2.13]$, 11/12 positive; but Conway − Seeds = $+0.32$ $[-0.22, +0.86]$, only 6/12 positive, so Conway and Seeds are *not* cleanly separated — Conway's wide CI reflects its known seed sensitivity (Section 4.5). A main-effect variance decomposition on the balanced $3 \times 3 \times 3$ grid design attributes 73% of the layer-center variance to the rule, versus 6% to seed and 0.3% to tokenization (the rest residual/interactions). Within the tested grid design, rule identity is thus the dominant measured main effect on routing depth. The separation survives restricting to grid-only or fixed-graph-only runs (Figure 4).

As a sensitivity check, a mean-ablation sweep on *all three rules* gives essentially the same run-level centers as zero-ablation (Seeds 1.34 vs. 1.31, Conway 1.66 vs. 1.63, DayNight 3.29 vs. 3.29) and preserves the Seeds head ranking (Spearman 0.978). The Seeds-early / DayNight-late ordering therefore does not depend on the ablation choice, ruling out a zero-ablation-specific cross-rule artifact.

One interpretation is that Seeds can route and aggregate the shallow `dead & count == 2` statistic early, while DayNight's larger birth/survival sets and 0/1 complement symmetry may favor additional preprocessing before neighbor aggregation. We treat this as a hypothesis rather than a demonstrated path-level mechanism.

4.5 Conway: robust neighbor causality, seed-sensitive localization

Conway consistently shows true-neighbor causal effects, but its layer localization varies across seed, tokenization, topology, and per-head scans. Most runs have causal heads in L1–L2, but some shift later. We therefore claim robust neighbor causality for Conway, not a fixed Conway layer hierarchy.

Figure 3: Three-rule per-head profile. Left: share of neighbor-attention mass by layer (observational). Right: single-head zero-ablation damage by layer (causal). The causal center of routing moves from early layers in Seeds to late layers in DayNight, with Conway intermediate and seed-sensitive.

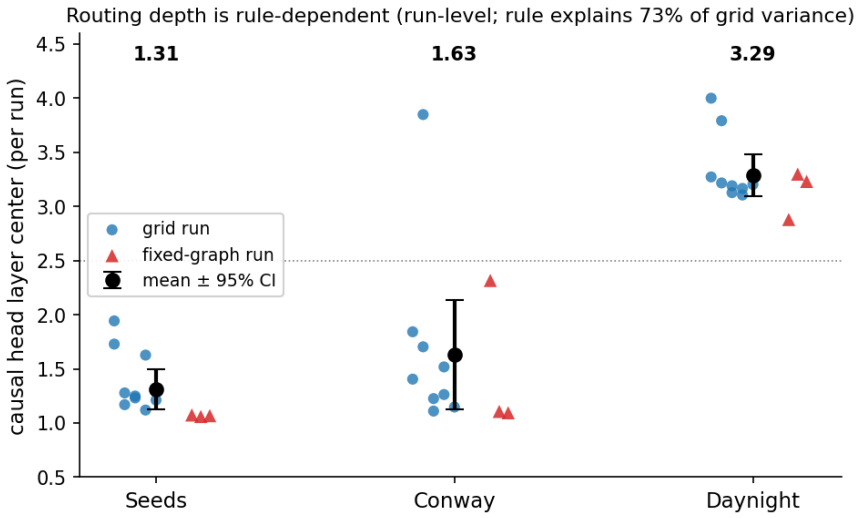

Figure 4: Run-level causal head layer center. Each point is one run (grid runs in blue, fixed-graph runs as red triangles); black markers are the per-rule mean ±95% CI over 12 runs. DayNight is separated from Seeds in all 12 matched pairs, while Conway overlaps Seeds and shows the wider spread expected from its seed sensitivity (the high Conway point is its anomalous late-routing run). A variance decomposition attributes 73% of grid layer-center variance to the rule.

4.6 32×32 scaling is training-budget sensitive

At a 50k-step budget, the 32×32 Seeds and DayNight runs appeared not to converge (DayNight board accuracy 0.0). Continued training changes this: Conway seeds 42/123/456 converge by 200k steps, and Seeds/DayNight seed42 converge by 300k steps, with final board accuracy 0.9825–0.9975. The training curves show a late exact-board-accuracy breakthrough (Figure 5), showing that longer training substantially alleviates the 50k-step optimization failure in the tested runs, so the 50k result was not sufficient evidence of a capacity limit. A reduced seed42 32×32 mechanism analysis preserves the 16×16 direction: Seeds remains early with `seeds_birth_count2` probe score 1.0, and DayNight remains later. Because non-Conway 32×32 runs currently use seed42 only, this remains a scaling sanity check rather than a seed-robust 32×32 claim.

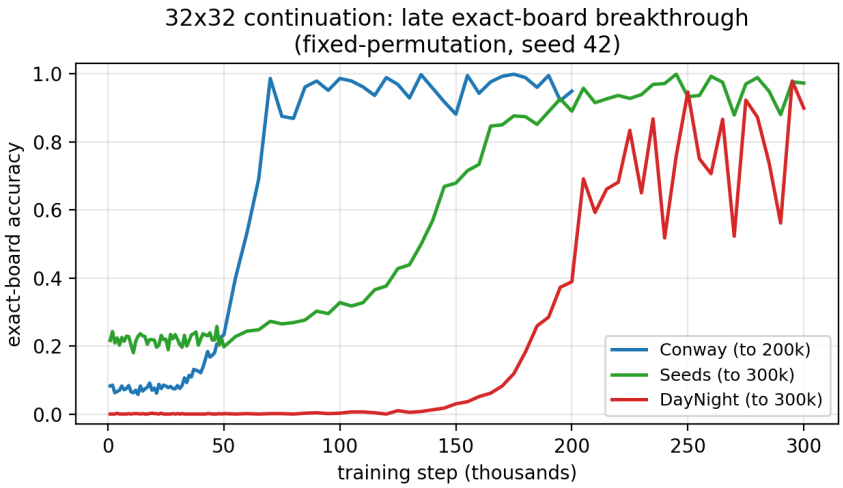

Figure 5: 32×32 continuation training curves (fixed-permutation, seed 42). Exact-board accuracy breaks through late in training — earliest for Conway, latest for DayNight — showing that longer training substantially alleviates the 50k-step optimization failure in the tested runs.

4.7 A QK/OV case study of top routing heads

We fix one preselected run per rule (the 16×16 shuffled_set seed42 run) and, within it, take the single head with the highest neighbor-attention mass (Conway L4H0, Seeds L1H3, DayNight L4H2), which is also among the top heads by ablation damage; no head is selected on the QK/OV outcome. For that head we separate the head's two streams with the pattern-substitution methodology of Section 3.3. On the **OV side**, the head's per-token output contribution is well explained by its attention-weighted alive content: a ridge regression onto the attention-weighted alive fraction of the attended tokens reaches $R^2 = 0.994$–0.997 for all three heads (a shuffled-label control gives $R^2 \approx 0$). This is, however, a *generic* architectural property rather than a signature of routing heads: computing the same R^2 for all 16 heads of each run places the selected head only at rank 3/16 (Conway), 6/16 (Seeds), and 2/16 (DayNight) — at or below the all-head median (0.995–1.000). We therefore read the OV R^2 as confirming the head-output aggregation identity (a head writes an attention-weighted combination of its value stream, and alive state is near-linearly encoded), not as head-specific evidence. The selected-head evidence is on the **QK side**, where attention-pattern substitution shows a consistent ordering: restricting the pattern to the true neighbor set retains far more function than uniform-over-all or row-shuffled controls (Seeds: loss 0.50 vs 0.80/1.16; DayNight: 0.27 vs 0.47/1.52; Conway: 0.34 vs 0.57/1.13; baseline ≈ 0),[2] but flattening the distribution *within* the neighbor set — while preserving the set, the per-query neighbor mass, and the off-neighbor pattern — is as damaging as the mass= 1 ideal template. In these heads, the learned within-neighbor weighting is therefore functionally important. These case studies support a weighted neighbor-content aggregation interpretation for the selected routing heads, consistent with the per-head direction specialization visible in the attention maps (Figure 6; four heads covering eight directions non-uniformly). Notably the pattern replicates on Conway's anomalous late-routing run. Scope: one run and one head per rule — case studies, not full path-level circuit identification.

4.8 DayNight maintains complement-aligned representations

DayNight is complement-symmetric: complementing the board commutes with the update. A symmetry intervention (3 seeds, 16×16 shuffled_set) shows the models reflect this *representationally*: matched-cell CKA between $h(x)$ and $h(1-x)$ stays above 0.98 at every DayNight layer for every seed (seed means

[2]These per-condition losses are one hardware realization (GPU). Board sampling uses device-local RNG, so the exact values differ on CPU (e.g., Seeds 0.57 vs 0.89/1.21) and, on the GPU path, in the last few significant figures across CUDA architectures; the across-condition ordering and the OV R^2 range are the hardware-robust, reproducible claims. The supplement ships both CPU and CUDA tables so a reviewer matches their backend.

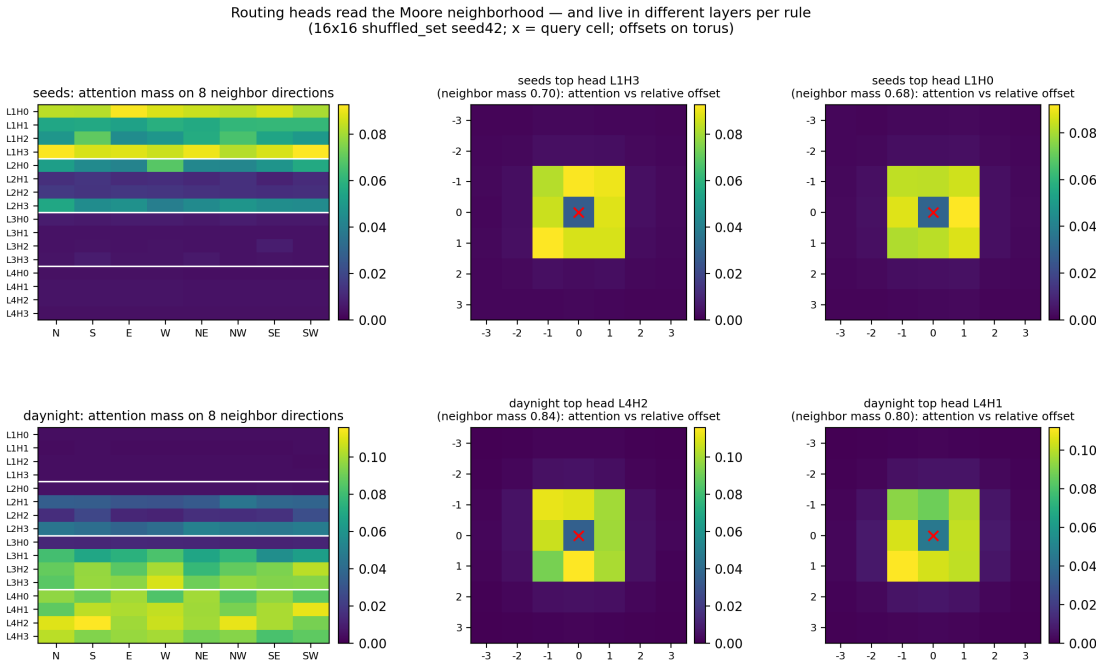

Figure 6: Qualitative attention evidence. Top routing heads place their attention mass on Moore-ring neighbor positions, with different heads specializing to different neighbor directions non-uniformly, consistent with the weighted neighbor-content aggregation of Section 4.7.

≥ 0.99); the thresholded predicted labels satisfy complement covariance on all evaluated cells ($\text{pred}(1-x) = 1 - \text{pred}(x)$ at agreement 1.0000), while Conway and Seeds — whose outputs are polarity-bound — diverge with depth (L4 CKA 0.40 and 0.45; Figure 7). Because learned cell-identity embeddings could inflate raw alignment, we repeat the measurement after subtracting each cell's mean representation (computed jointly over originals and complements): the contrast survives unchanged (DayNight ≥ 0.98 at every layer for every seed; Conway/Seeds L4 at 0.40/0.48), and the identical-architecture Conway/Seeds baselines make the same point independently. Two further baselines (seed42) confirm the alignment is complement-specific and not a low-rank artifact of CKA: permuting the complement pairing (pairing cell i of x with an unrelated cell's complement) or pairing $h(x)$ with an independent same-distribution board both collapse the matched-cell CKA to ≈ 0 at every layer for every rule, whereas the true per-cell complement pairing gives DayNight ≥ 0.98 throughout (Figure 8). DayNight thus preserves a complement-specific representational geometry — we say *aligned* rather than *invariant* because CKA is insensitive to sign flips of a low-variance polarity component. These controls establish that the alignment is specific to the matched complement relation rather than cell identity or generic same-distribution similarity. They do not identify a polarity-normalization mechanism or explain DayNight's later routing profile; the connection between complement alignment and deferred routing (Section 4.4) remains a hypothesis (Section 7). We note for completeness that a count-probe transfer test (train on originals, test on complements with complemented labels) is trivially ≈ 1 for any accurate model, because complements are themselves in-distribution boards; we report it only as a pipeline validity check. The mechanism implementing the alignment remains unidentified.

5 Summary of Claims and Evidence

We summarize the claims and their support in Table 5 (Appendix A). In brief:

Strong. Models learn one-step Life-like updates across completed 16×16 token-ordering conditions and the fixed non-grid graph control (Table 1); true-neighbor causal dependence persists across these conditions

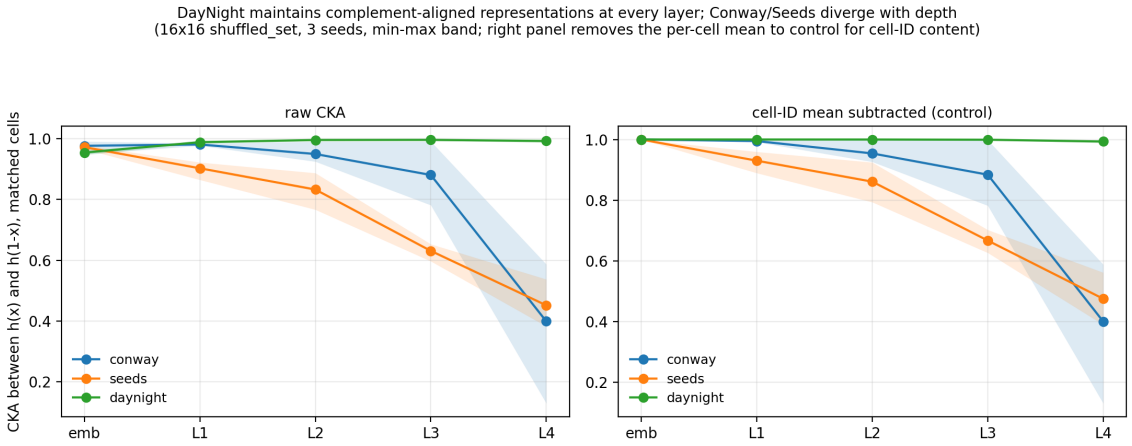

Figure 7: Complement-symmetry alignment. Left: matched-cell CKA between $h(x)$ and $h(1-x)$ per layer; DayNight stays complement-aligned at every layer while polarity-bound Conway/Seeds diverge with depth. Right: the same contrast after subtracting each cell's mean representation, ruling out that cell-identity embeddings inflate the alignment.

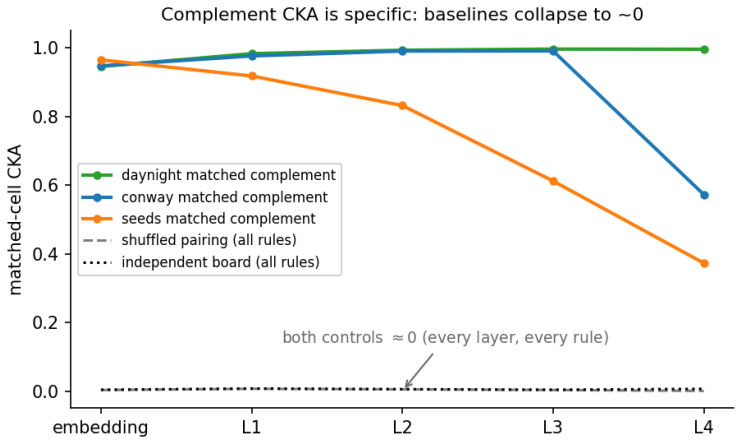

Figure 8: Complement-CKA specificity controls (seed42). Matched-cell CKA between $h(x)$ and $h(1-x)$ per layer (solid) stays high for DayNight and diverges for polarity-bound Conway/Seeds, while the two baselines — permuting the complement pairing (dashed) and pairing with an independent same-distribution board (dotted) — collapse to ≈ 0 at every layer for every rule. The alignment is thus specific to the matched complement relation, not a low-rank artifact of CKA.

and the fixed graph control; Seeds exposes a rule-specific sufficient statistic, supported by full predicate probes and state-conditioned controls; DayNight is reliably later than Seeds in both layerwise and head-level analyses; neighbor routing is implemented by a small set of heads whose depth changes with the rule, and top-2 and top-3 group ablations are approximately additive on average (with documented exceptions).

Moderate. Conway has robust neighbor causality, but its localization is seed/tokenization/topology-sensitive; per-rule case studies support a weighted neighbor-content aggregation interpretation for top routing heads, resting on the QK-side result that the learned within-neighbor weighting is functionally important (the OV $R^2 = 0.994$–0.997 is a generic head-output identity that holds for all 16 heads, not head-specific evidence); DayNight maintains complement-aligned representations at every layer (per-seed CKA ≥ 0.98, seed means ≥ 0.99, surviving a cell-ID mean-subtraction control; polarity-bound controls diverge); 32×32

convergence is achievable with enough training, and reduced seed42 32×32 mechanism checks are consistent with the 16×16 pattern, though non-Conway 32×32 evidence remains single-seed.

6 Limitations / What We Do Not Claim

- We do not claim unseen-graph generalization; the fixed random-graph control uses the same graph for training and evaluation.
- We do not claim path-level circuit identification beyond the per-rule QK/OV case studies of Section 4.7 (three heads total); head-level localization is covered across all runs, path-level evidence remains case-study scale.
- We do not claim a full pairwise head-interaction map; pairs are measured for the top-5 heads of each run only, and a minority of tested pairs do interact.
- The complement-alignment result (Section 4.8) identifies a representational property (alignment, not strict invariance), not the mechanism that implements it.
- We do not claim seed-robust non-Conway 32×32 scaling or a full 32×32 mechanism matrix.
- Long-rollout / OOD-density robustness is covered only by a lightweight 16×16 shuffled-set forward-evaluation sanity check (Appendix B); it is not a broad robustness claim.
- Linear probes have the usual limitation: failure of linear decodability is not proof that information is absent (Belinkov, 2022).
- External validity is limited to what we tested: three hand-selected rules (Conway, Seeds, DayNight), a single 4-layer/4-head architecture, and one training recipe. "Rule-dependent" therefore refers to these three rules, not the full Life-like B/S rule space; we do not claim the routing-depth ordering generalizes to other rules, architectures, or hyperparameters.

7 Future Work

The testbed opens several directions that build on the results above. A natural next step is to turn the per-rule QK/OV case study (Section 4.7) into full path-level circuits using path patching (Goldowsky-Dill et al., 2023) and automated circuit-discovery tooling (Conmy et al., 2023), and to test whether specific representations realize the canonical high-level variables (`count`, `count_eq_2`, birth/survival branch) through interchange interventions in the causal-abstraction framework (Geiger et al., 2021), which the exactly specified local structure of this setting makes straightforward to construct. A further step is to identify the components that produce DayNight's complement-related representational geometry (Section 4.8), which our intervention establishes as a representational property but does not mechanistically localize, and to test whether that alignment is causally linked to DayNight's later routing — the "polarity normalization before routing" hypothesis — e.g. by intervening to disrupt the alignment and measuring the change in routing depth or behavior. The exact ground truth of the setting also makes it well suited to studying how circuit organization varies with graph topology (via unseen-graph generalization tests), with scale (seed-robust 32×32 mechanism matrices), and under distribution shift (broader long-rollout and OOD-density evaluation). More broadly, because the rule is a controllable variable with known structure, the framework is a convenient substrate for validating and stress-testing mechanistic-interpretability methods against a fully specified algorithm.

Reproducibility Statement

An anonymous code and data supplement accompanies this submission. It contains the full model, training, evaluation, and analysis library, the analysis and plotting scripts, the summary tables and per-analysis methodology write-ups behind every table and figure, and the trained checkpoints for the two case-study analyses. With those bundled checkpoints, the QK/OV case study (Section 4.7) and the complement-symmetry intervention (Section 4.8) reproduce directly; the aggregate analyses over the full 16×16 matrix

are reproduced by retraining with the provided configs and scripts (checkpoints for the full matrix exceed the supplementary size limit). All analyses are forward-only (they load a trained checkpoint and compute metrics on fresh boards with a fixed data seed, so every ablation condition sees identical inputs). One-step metrics are averaged over 51,200 boards per run, and the model, data-generation, and evaluation settings are given in Section 3. Because board sampling uses device-local RNG, exact per-condition losses and linear-probe scores are hardware-specific (a fixed seed yields different boards, hence slightly different learned models, across GPUs); the reproducible contract is therefore the reported thresholds and orderings (near-perfect one-step accuracy; the true-neighbor vs. matched-control causal gaps; OV $R^2 \in [0.994, 0.997]$; per-layer DayNight CKA ≥ 0.98), with the shipped CSVs being one hardware realization.

Broader Impact Statement

This work studies small transformers on synthetic cellular-automaton prediction as an interpretability testbed. It uses no human or sensitive data and trains only small models. We are not aware of direct negative societal impacts; the primary intended use is methodological, as a controlled setting for validating and stress-testing mechanistic-interpretability methods.

References

Tashin Ahmed and Q. Tyrell Davis. It's much easier for neural networks to learn game of life dynamics with the right activation function: Polynomial kolmogorov-arnold networks. In *Proceedings of the 2026 Artificial Life Conference*, 2026.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Jaime A. Berkovich and Markus J. Buehler. Lifegpt: Topology-agnostic generative pretrained transformer model for cellular automata. *npj Artificial Intelligence*, 2025. arXiv:2409.12182 (2024).

Jaime A. Berkovich, Noah S. David, and Markus J. Buehler. Automatagpt: Forecasting and ruleset inference for two-dimensional cellular automata. *arXiv preprint arXiv:2506.17333*, 2025.

Mikhail Burtsev. Learning elementary cellular automata with transformers. *arXiv preprint arXiv:2412.01417*, 2024.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

David Eppstein. Growth and decay in life-like cellular automata. In Andrew Adamatzky (ed.), *Game of Life Cellular Automata*, pp. 71–97. Springer, 2010.

Martin Gardner. Mathematical games: The fantastic combinations of john conway's new solitaire game "life". *Scientific American*, 223(4):120–123, 1970.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

William Gilpin. Cellular automata as convolutional neural networks. *Physical Review E*, 100(3):032402, 2019.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations (ICLR)*, 2023.

David Lindner, János Kramár, Sebastian Farquhar, Matthew Rahtz, Thomas McGrath, and Vladimir Mikulik. Tracr: Compiled transformers as a laboratory for interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Alexander Mordvintsev, Ettore Randazzo, Eyvind Niklasson, and Michael Levin. Growing neural cellular automata. *Distill*, 2020. doi: 10.23915/distill.00023. Web-native publication; no PDF edition exists.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023a.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop*, 2023b.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Felipe Parodi, Jordan K. Matelsky, and Melanie Segado. Zero-ablation overstates register content dependence in dino vision transformers. In *CVPR Workshops*, 2026.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Aditya Singh, Zihang Wen, Srujananjali Medicherla, Adam Karvonen, and Can Rager. Automatically finding rule-based neurons in othellogpt. In *NeurIPS 2025 Workshop on Mechanistic Interpretability*, 2025.

Jacob M. Springer and Garrett T. Kenyon. It's hard for neural networks to learn the game of life. In *International Joint Conference on Neural Networks (IJCNN)*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. In *International Conference on Learning Representations (ICLR)*, 2023.

Stephen Wolfram. Statistical mechanics of cellular automata. *Reviews of Modern Physics*, 55(3):601–644, 1983.

Yifei Yuan and Anders Søgaard. Revisiting the othello world model hypothesis. In *International Conference on Learning Representations (ICLR)*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A Claim and evidence summary

Table 5: Claim hierarchy and evidence summary. Strong claims are supported across the completed 16×16 tokenization and fixed random-graph matrix. Moderate claims are supported but have explicit scope limits.

Claim	Main evidence	Support	Caveat
Models learn one-step Life-like updates.	Cell/board/rollout accuracy is near-perfect across Conway, Seeds, and DayNight on 16×16 grid tokenizations and fixed random graph.	Strong	Random graph is same-graph train/eval; long rollout and OOD-density are covered only by lightweight 16×16 shuffled-set sanity checks (Appendix B), not broad robustness.
True-neighbor causal dependence persists.	Neighbor attention enrichment, layerwise neighbor ablation, edge-mass matched ablation, and logit/loss metrics all favor true neighbors over matched non-neighbor controls.	Strong	Grid/tokenization artifact risk is reduced, not eliminated; unseen-graph transfer is untested.
Seeds exposes a rule-specific sufficient statistic.	`dead & count_eq_2` reaches best balanced accuracy 1.0000 across 12 completed 16×16 runs; state-conditioned probes show L4 `count_eq_2` accuracy 1.0000 on dead cells vs 0.6722 on alive cells; survives projecting out the readout direction (1.000) and is highest among all class-balanced `count_eq_`k probes.	Strong	Linear-probe evidence (decodability, not established causal use of the direction); low-count predicates are also broadly decodable; count_eq_k and readout-projection controls on shuffled-set, 3 seeds; absence of linear decodability is not proof of absent information.
Routing depth is rule-dependent.	Run-level (zero-ablation, all rules) causal head layer centers 1.31 [1.13,1.50] (Seeds), 1.63 [1.13,2.13] (Conway), 3.29 [3.10,3.48] (DayNight); DayNight−Seeds positive in 12/12 matched pairs; rule explains 73% of grid layer-center variance (seed 6%, tokenization 0.3%); neighbor-attention mass correlates with single-head causal impact (Spearman 0.77–0.92).	Strong	Head-level evidence only; Conway/Seeds not cleanly separated (Conway is seed-sensitive); QK/OV path-level localization remains case-study scale.
Top-2 and top-3 head-group ablations are approximately additive on average.	Top-2 and top-3 groups (all rules) and the all-L1 group of Conway/Seeds have group/sum-of-singles ratios 0.89–1.11. The DayNight all-L1 group is excluded from ratio summaries because the summed single-head effect is near zero; its absolute interaction is $\approx +0.04$, superadditive relative to the individual L1 effects. Top-5 pairwise sweep mean ratio 1.04–1.08.	Strong	A minority of tested pairs and the DayNight all-L1 group show interactions; the full pairwise map beyond the top-5 heads is not measured, and approximate additivity does not by itself establish a functional division of labor.
Weighted neighbor-content aggregation in top heads.	Per-rule QK/OV case study: QK-side attention-pattern substitution shows the learned within-neighbor weighting is functionally important. (OV $R^2 = 0.994$–0.997 is a generic head-output identity holding for all 16 heads, not head-specific evidence.)	Moderate	Three heads total (one run per rule); OV side not head-specific; not full path-level circuit identification.
DayNight maintains complement-aligned representations.	Matched-cell CKA between $h(x)$ and $h(1-x) \geq 0.98$ at every DayNight layer per seed (means ≥ 0.99), surviving a cell-ID mean-subtraction control; polarity-bound Conway/Seeds diverge (L4 0.40/0.48).	Moderate	Identifies a representational property (alignment, not invariance), not the implementing mechanism; controls on shuffled-set, 3 seeds.
Conway is neighbor-causal but localization is seed-sensitive.	True-neighbor causal gaps persist, but layer centers vary across seed/tokenization/topology and per-head scans.	Moderate	Do not claim a fixed Conway layer hierarchy.
32×32 convergence is achievable with enough training.	Continued fixed-permutation runs reach board accuracy at least 0.9825: Conway seeds 42/123/456 to 200k, Seeds/DayNight seed42 to 300k.	Moderate	Non-Conway 32×32 evidence is seed42 only; reduced 32×32 mechanism analysis is not a full matrix.

B Long-rollout and density-OOD sanity check

A forward-only evaluation rolls the 16×16 shuffled-set checkpoints out to horizon 10, autoregressively (feeding the model its own predictions) and teacher-forced, from initial densities inside the training mixture and at its sparse/dense edges. Teacher-forced cell accuracy is at least 0.9999 at four decimals for all rules and conditions (worst raw value: DayNight train-mixture, 0.99991). Autoregressively, at horizon 10:

- Seeds is exactly solved in every tested condition;
- Conway keeps exact-board accuracy at or above 0.9993;
- DayNight is exact in sparse/mid/dense conditions and reaches cell accuracy 0.9994 / board accuracy 0.9821 on the training mixture.

Extreme densities partly collapse to trivial attractors (e.g., dense Conway boards die out), which inflates those cells of the table. The check shows that one-step models remain usable under rollout and density shift on 16×16 shuffled-set checkpoints; it is not a broad OOD-robustness claim.

C Additional per-head controls

Figure 9 shows the group-ablation redundancy summary (top-2/top-3/all-L1 group vs sum-of-singles) and the pairwise ratio distribution behind Section 4.4; Figure 10 compares the Seeds zero-ablation and mean-ablation head rankings.

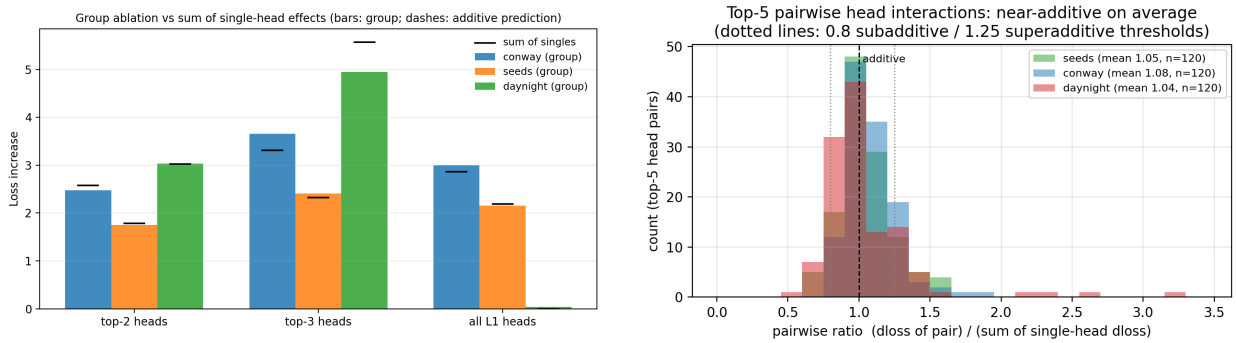

Figure 9: Left: group-ablation redundancy (group vs sum-of-singles). The DayNight all-L1 group is excluded from ratio summaries (its summed single-head effect is near zero); its absolute interaction ≈ +0.04 is superadditive relative to the individual L1 effects. Right: distribution of top-5 pairwise interaction ratios, concentrated near 1 (additive) with a minority of interacting pairs.

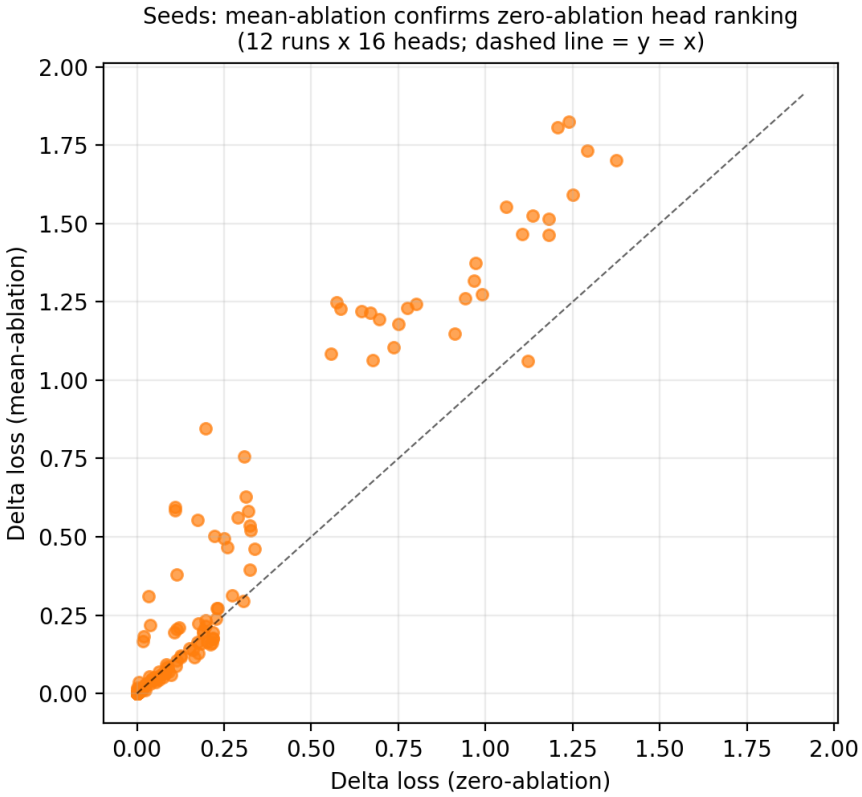

Figure 10: Seeds mean-ablation vs zero-ablation head ranking (Spearman 0.978), showing the head ranking is not an artifact of pushing activations to zero.