# Biregular Sparse Initialization Shifts the Rate and Shape of Compositional Escape in Sequential Arithmetic Curricula

**Clément Castellon** [1] **Arindam Biswas** [2]

## Abstract

Training a model on one task and then a second, related task can leave the second stalled short of mastery. We study a sharp case: a 2-layer transformer that has grokked modular addition, then trained on a 50/50 addition–multiplication curriculum, plateaus near 43% multiplication accuracy and stays there. We show that a single change at initialization (with the training procedure left untouched) can break this ceiling. Comparing sparse masks drawn from Ramanujan graphs (biregular optimal spectral expanders) against matched unstructured masks, we test three preregistered hypotheses, all of which the data substantially refine. Ramanujan initialization substantially raises the chance of reaching full multiplication accuracy, an advantage most pronounced at the highest sparsity and that unstructured masks never achieve. It also sharpens outcomes, replacing ambiguous partial learning with either clean failure or full success. The spectral margin that defines these graphs is near zero exactly where the benefit peaks, so we attribute the effect not to expansion but to degree uniformity, which leaves no unit isolated or hub-like. Embedding dynamics prove necessary for escape, though why biregular structure aids them remains open.

## 1. Introduction

*Curriculum learning* trains a model on tasks in a deliberate order, often easier-to-harder, rather than all at once (Bengio et al., 2009). A *sequential* curriculum is the extreme case: the model masters one task before the next is introduced. This raises a basic question—does having already learned one skill help or hinder learning the next?

Transformers trained on modular arithmetic offer a clean testbed. They exhibit *grokking*: after a long phase in which the model merely memorizes the training data, validation accuracy abruptly jumps to perfect generalization (Power et al., 2022). When a model that has grokked addition is then trained on a curriculum mixing addition and multiplication, a different phenomenon emerges—multiplication validation accuracy stalls around 0.43 and persists there for the full training budget. We call this the *sequential curriculum ceiling* and study it here.

Prior interventions on compositional generalization have focused on training dynamics: curriculum design and catastrophic-forgetting regularization (Kirkpatrick et al., 2017; Bengio et al., 2009). We ask instead whether a *static initialization* can shift this ceiling, leaving the training procedure unchanged.

Ramanujan graphs are bipartite, biregular graphs—every vertex on each side has the same fixed degree—whose second-largest singular value saturates the Alon–Boppana bound, certifying near-optimal spectral expansion (Hoory et al., 2006). Two properties make such masks a plausible lever for a sequential curriculum. First, strong expansion implies efficient information mixing across the sparse mask. Second, and central to this paper, degree-regularity guarantees that *every* input and output coordinate participates in the same number of active connections, so no coordinate is left isolated or turned into a hub that one task's gradients can monopolize. Biswas et al. (2024) showed these masks improve pretraining efficiency; here we ask whether they affect compositional generalization under sequential curriculum, where two task gradients compete for the same weights.

We preregistered three hypotheses before running experiments. **H1**: Ramanujan masks increase structural disjointness between task circuits. **H2**: Ramanujan initialization accelerates addition grokking. **H3**: The curriculum ceiling is unaffected, because interference arises from the shared embedding table.

All three are falsified or substantially refined by the data.

H1 is not supported as stated: the available M2 analysis points toward greater circuit overlap in sparse networks rather than increased disjointness; we therefore treat cir-

[1]ERC ModERN, Sorbonne Université [2]Polynom Research, Paris, France. Correspondence to: Clément Castellon <clement.castellon@gmail.com>.

*Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea.* *Copyright 2026 by the author(s).*

cuit overlap as a secondary diagnostic rather than the main mechanism claim.

H2 is partially confirmed but misattributed: sparsity accelerates grokking, but biregular structure adds nothing over matched unstructured sparsity.

H3 is falsified—most clearly at the highest sparsity, where Ramanujan escape significantly exceeds unstructured (Section 3.3). The embedding ablation refines rather than rescues H3: embedding dynamics are necessary for escape (freezing them collapses every condition to the ceiling), but the curriculum ceiling is not embedding-determined alone—biregular initialization shifts it when embeddings are free to co-adapt. Structure and embedding dynamics are thus jointly required, a dissociation H3 did not anticipate.

## 2. Setup

**Preliminaries.** We use the following terms throughout. *Sparsity* $s$ is the fraction of zeroed weights in the masked layers; unstructured masks keep each entry i.i.d. with probability $1 - s$. A run *escapes* the ceiling if its multiplication validation accuracy reaches $\geq 0.95$; the *escape rate* is the fraction of seeds in a condition that escape. A *partial escape* is an endpoint in $[0.50, 0.95)$: the model has learned some multiplication structure but not full generalization. The *ceiling* is the maximum multiplication validation accuracy reached after the phase transition. A mask is *biregular* if every row and every column has a fixed nonzero degree.

**Model and task.** We use a 2-layer causal transformer with $d_{\text{model}} = 128$, 4 attention heads, $d_{\text{ff}} = 512$, pre-norm residuals, and no dropout ($\sim$422K parameters). The vocabulary consists of integers 0–96, two operator tokens (+, $\times$), and an equality token. Inputs are $(a, \text{op}, b, =)$; targets are $(a \text{ op } b) \bmod 97$. We use a 40%/60% train/validation split, batch size 512, AdamW with $\text{lr} = 10^{-3}$ and weight decay 0.1. With $p = 97$ and two operands there are $97^2 = 9409$ distinct $(a, b)$ pairs per operation; a 40% split gives 3763 training and 5646 validation examples. With batch size 512 and `drop_last`, the model sees 3584 training examples per epoch (the remaining 179 resampled across epochs under shuffling). The deliberately small training fraction is what induces grokking and is not an inversion of a conventional split.

**Conditions.** We compare seven initialization schemes: *Dense* (standard normal, $\sigma = 0.02$) and biregular *Ramanujan-s* vs. *Unstructured-s* at sparsity $s \in \{0.90, 0.95, 0.977\}$. Ramanujan masks satisfy the Alon–Boppana bound on the second singular value. Unstructured masks are Bernoulli i.i.d. masks with entrywise keep probability $1 - s$, without post-processing to enforce nonzero row or column degree. Both are applied to attention projections $(Q, K, V, O)$ and the FFN down-projection; embeddings and FFN up-projection remain dense. Dense uses all 30 exp007 seeds. Each Ramanujan and unstructured condition uses the same matched subset of 15 exp007 seeds at each sparsity level.

**Measurements.** **M1** – single-task addition training for 25K steps. We record the grokking step (earliest point with 3 consecutive validation evaluations $\geq 0.9$) and whether grokking is *genuine* (train accuracy reaches 0.95 before validation accuracy reaches 0.50). **M2** – Jaccard overlap between the top-30% magnitude circuit extracted from addition-trained and multiplication-trained checkpoints, restricted to non-structural-zero weights. **M3** – sequential curriculum. *Phase 1*: addition only, trained until grokked. *Phase 2*: addition and multiplication batches interleaved 50/50 for 100K steps. The *multiplication ceiling* is the maximum multiplication validation accuracy reached after the phase transition. **M4 (ablation)** – exp008: phase-2 training with embeddings frozen vs. free, starting from the same phase-1 checkpoint. Conditions: dense, Ramanujan-0.977, unstructured-0.977.

## 3. Results

### 3.1. Sparsity Accelerates Grokking; Biregular Structure Does Not

All sparse conditions substantially accelerate addition grokking relative to the dense baseline (median step $\sim$3500 for dense). Median grokking steps are 1000, 1750, and 1250 for Ramanujan and 1600, 2050, and 1600 for unstructured at sparsity 0.90, 0.95, and 0.977 respectively. Mann-Whitney tests comparing Ramanujan vs. unstructured at matched sparsity find no significant differences (all $p > 0.10$), indicating the acceleration is driven by sparsity, not by the biregular construction.

This result refines H2: the claim that Ramanujan *specifically* accelerates grokking is not supported. Sparse initialization, regardless of structural properties, compresses the memorization phase.

### 3.2. Circuit Overlap Does Not Support the Disjointness Hypothesis

We also computed M2, the Jaccard overlap between top-30% magnitude circuits extracted from separately trained addition and multiplication checkpoints. Contrary to H1, the M2 analysis does not support increased disjointness under Ramanujan initialization. We treat this as a secondary diagnostic because the measurement does not cleanly distinguish biregular structure from sparsity itself.

This measurement is sensitive to a threshold-collapse artifact. If structural zeros are included when selecting the

| Condition | $n$ | Esc $\geq 0.95$ | Mean | Median |
|---|---|---|---|---|
| Dense | 30 | 4/30 | 0.45 | 0.43 |
| Ramanujan 0.90 | 15 | 8/15 | 0.76 | 0.95 |
| Ramanujan 0.95 | 15 | 8/15 | 0.83 | 1.00 |
| Ramanujan 0.977 | 15 | 11/15 | 0.92 | 1.00 |
| Unstructured 0.90 | 15 | 4/15 | 0.62 | 0.45 |
| Unstructured 0.95 | 15 | 5/15 | 0.72 | 0.67 |
| Unstructured 0.977 | 15 | 4/15 | 0.79 | 0.87 |

*Table 1.* Multiplication curriculum ceiling per condition. Escape rate ($\geq 0.95$) is nondecreasing for Ramanujan but is flat for unstructured across all three sparsity levels. Dense $n = 30$ from prior experiment (exp007).

top-30% weights by magnitude, the threshold degenerates on sparse networks and the resulting overlap is spuriously inflated. We therefore apply the stored mask buffer before thresholding and compute the top-30% mask only over structurally active weights.

### 3.3. Biregular Structure Shifts Escape Rate; Unstructured Does Not

Table 1 summarizes multiplication ceiling outcomes. For Ramanujan, escape rate ($\geq 0.95$) is nondecreasing with sparsity: 8/15, 8/15, 11/15 at 0.90, 0.95, 0.977. For unstructured, the escape rate is flat across the same range: 4/15, 5/15, 4/15. The Ramanujan advantage at sparsity 0.977 is statistically significant (Fisher exact test: $p = 0.013$ one-sided, $p = 0.027$ two-sided). At lower sparsities the direction is consistent but individually underpowered at $n = 15$. A paired Wilcoxon signed-rank test on matched-seed ceiling values confirms the distributional separation at 0.977 ($p = 0.008$). We designate the comparison at sparsity 0.977 as the single primary, preregistered test; it therefore requires no multiplicity correction. The comparisons at 0.90 and 0.95 are secondary and exploratory; we report them unadjusted and do not claim individual significance.

This falsifies H3 in a dose-dependent, structurally-specific manner. The ceiling is not determined solely by the shared embedding table, and the effect depends on structure beyond sparsity alone.

### 3.4. Biregular Structure Shifts Outcomes Toward Bimodality

An unexpected finding concerns the *shape* of the outcome distribution at sparsity 0.977. Unstructured networks place 67% of runs in the partial-escape range (0.50–0.95): the model learns something about multiplication but cannot complete full generalization. Ramanujan networks place only 27% of runs in this range; the remaining runs either fail cleanly ($< 0.50$) or escape completely ($\geq 0.95$). We report this as an exploratory, post-hoc observation at $n = 15$ per condition; a preregistered replication is the appropriate confirmation.

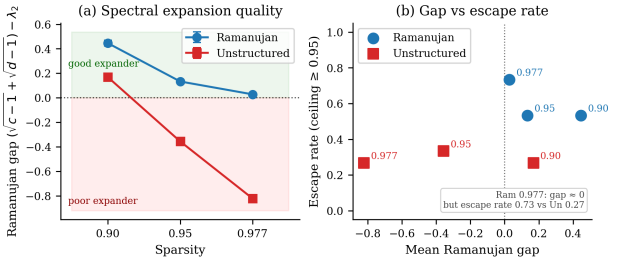


*Figure 1.* (a) Ramanujan margin ($\sqrt{c-1}+\sqrt{d-1}-\lambda_2$) by condition. Ramanujan networks satisfy the biregular Ramanujan criterion by construction; for unstructured masks, which are not biregular, the same quantity is shown only as a diagnostic using matched average degrees. (b) Scatter of mean Ramanujan margin vs. escape rate. At sparsity 0.977, the Ramanujan margin is 0.03 — near zero — yet Ramanujan escape rate (0.73) substantially exceeds unstructured (0.27). Variation in this margin does not explain escape within or across conditions.

This bimodality difference is not captured by mean or escape rate alone. A structural explanation is available: at sparsity 0.977, Bernoulli masks produce isolated rows or columns with non-negligible probability. For example, an i.i.d. $128 \times 128$ mask has an expected $128 \cdot 0.977^{128} \approx 6.5$ zero-output rows, and the masked attention projections aggregate many such opportunities for disconnection. In our runs, unstructured-0.977 masks contain a mean of 49 isolated row/column incidences across the masked attention projections of a transformer block. Biregular construction removes this failure mode by fixing nonzero row and column degrees. These isolated coordinates can sustain partial signal for some circuits while preventing full generalization, producing the ambiguous intermediate outcomes observed in unstructured networks.

## 4. Ramanujan Margin Is Not the Operative Mechanism

Figure 1 shows the Ramanujan margin ($\sqrt{c-1}+\sqrt{d-1}-\lambda_2$) by condition. Ramanujan masks satisfy the bound at all sparsity levels; for unstructured masks, the same diagnostic quantity becomes negative at 0.95 and 0.977 when evaluated using matched average degrees. However, the margin for Ramanujan collapses monotonically with sparsity: 1.73 at 0.90, 0.45 at 0.95, 0.03 at 0.977. At the sparsity level where the compositional advantage is largest, the margin to the Ramanujan bound is nearly zero. Thus, variation in this margin does not explain the escape-rate advantage. Within-condition correlation between per-run $\lambda_2$ and escape outcome is weak ($\rho \approx 0.2$–$0.3$).

The Ramanujan-margin diagnostic is therefore unlikely to

be the operative mechanism. We propose *degree uniformity* as the leading candidate: biregular masks guarantee fixed row and column degrees, so every input and output coordinate participates in the same number of active connections. This removes hubs and isolated coordinates, producing more uniform exposure to task gradients at the unit level rather than guaranteeing equal gradients for individual weights. In a sequential curriculum where phase 2 mixes two task gradients, hub neurons in an unstructured network can be captured by whichever task currently produces larger gradient magnitudes. Biregular structure eliminates hubs by construction, producing more balanced task-gradient exposure throughout phase 2.

**Embedding ablation (exp008).** Starting from pre-grokked Phase-1 checkpoints, we ran Phase 2 with embeddings either free or frozen, across dense, Ramanujan-0.977, and unstructured-0.977 conditions (n=15 matched seeds, with identical initialization within each free/frozen pair). Two results follow.

First, the structural advantage is preserved under a fully consolidated start rather than attenuated: Ramanujan-free escapes 13/15 versus unstructured-free 4/15 (Fisher exact $p = 0.003$), reproducing the exp007 separation (11/15 vs. 4/15), with dense-free at 1/15; the ordering Ramanujan $>>$ unstructured $>>$ dense matches exp007 throughout.

Second, the advantage is entirely embedding-mediated: with embeddings frozen, all three conditions collapse to the ceiling and escape 0/15 (mean ceilings 0.25, 0.41, 0.41 for dense, Ramanujan, unstructured), with Ramanujan-frozen and unstructured-frozen statistically indistinguishable (Fisher $p = 1.0$). Biregular weight structure does no embedding-independent work; its benefit appears only when embeddings co-adapt during Phase 2.

These results support H3's premise that embedding dynamics matter for phase-2 escape, while refuting its stronger prediction that the ceiling is unaffected by weight initialization. They further localize the structural advantage: it acts entirely through embedding co-adaptation, since frozen-embedding escape is identical across initializations. Why biregular structure specifically aids that co-adaptation—degree uniformity versus spectral properties—remains open and motivates the hub-controlled ablation named in Section 5.

## 5. Discussion

**Rate and shape shifts, not regime shifts.** The headline finding is a rate shift: Ramanujan initialization increases the probability of an outcome also reachable, rarely, by dense and unstructured initialization. The secondary finding is a shape shift: the distribution of outcomes changes from broadly ambiguous (unstructured) to cleanly bimodal (Ramanujan). Together, these suggest that biregular structure changes the geometry of the phase-2 loss landscape rather than simply biasing optimization toward a known region.

**Separation of sparsity and structural effects.** Prior sparse initialization literature conflates effects attributable to capacity reduction (fewer parameters) and structural properties (pattern of non-zero entries). Our design begins to separate them by comparing Ramanujan and unstructured masks at matched sparsity. Grokking speed is a capacity effect (same across methods). Compositional escape is a structural effect (Ramanujan consistently above unstructured; unstructured flat).

**Dead rows as confound.** At sparsity 0.977, Bernoulli masks produce isolated rows or columns with non-negligible probability; in our runs, unstructured-0.977 masks contain a mean of 49 isolated row/column incidences across the masked attention projections of a transformer block, whereas biregular construction fixes nonzero row and column degrees. This is a structural confound that conflates "random sparse" with "randomly disconnected"; future comparisons should control for these isolated coordinates explicitly, for example by post-processing unstructured masks to eliminate zero-degree rows and columns.

**Limitations and scope.** The setup is deliberately small (422K parameters, $p = 97$, arithmetic only). We treat modular arithmetic as a model organism: its narrowness is what permits the clean separation of sparsity from structure that is this paper's central methodological contribution, but it also means our findings are not yet shown to transfer to larger models or non-arithmetic compositional tasks, and we make no such claim. The mechanism likewise remains partially open: the embedding ablation shows embedding dynamics matter, and the spectral margin is ruled out (Section 4), but degree uniformity as the operative property is supported here by structural argument rather than direct ablation. A hub-controlled mask comparison (Bernoulli post-processed to fixed degrees vs. unconstrained, matched density) would close this gap and is the clear next experiment.

**Extensions.** For a longer version we plan: (1) a pattern-vs.-initialization ablation (Ramanujan mask with random dense-initialized values vs. biregular weights) to separate the structural and dynamical contributions; (2) linear mode connectivity analysis, forking runs from the same phase-1 checkpoint to estimate basin commitment time; (3) a task-agnostic circuit progress measure (epiplexity, $\varepsilon = H_{\text{emp}} - L_{\text{val}}$) to probe whether escape represents a latent mechanism becoming expressible or a new mechanism forming.

## 6. Related Work

Grokking on modular arithmetic was introduced by Power et al. (2022) and mechanistically characterized by Nanda et al. (2023), who showed that continuous Fourier progress measures track mechanism formation before validation accuracy improves. Merrill et al. (2023) frame grokking as competition between a dense memorizing subnetwork and a sparse generalizing one—motivating our focus on circuit support. Our setting differs: we compare task-circuit overlap across initialization structures rather than memorizing-vs.-generalizing subnetworks within a single training trajectory. Structured sparse initialization using deterministic Ramanujan graphs was studied by Biswas et al. (2024) as a deterministic alternative to lottery-ticket pruning (Frankle & Carbin, 2019). Data-free pruning-at-init methods such as SynFlow (Tanaka et al., 2020) share the goal but do not guarantee Ramanujan-style spectral control or degree uniformity. Compositional generalization under sequential or continual learning has been approached through curriculum design (Bengio et al., 2009) and catastrophic-forgetting regularization such as elastic weight consolidation (Kirkpatrick et al., 2017); our result identifies initialization structure as an underexplored lever that requires no modification to the training procedure.

## References


Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1–8. ACM, 2009. doi: 10.1145/1553374.1553380.

Biswas, A., Kalra, S. A., Mitra, P., and Basu, B. Sparse neural architectures and deterministic ramanujan graphs. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*, 2024. URL https://openreview.net/forum?id=MehPKPYbLA.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Hoory, S., Linial, N., and Wigderson, A. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006. doi: 10.1090/S0273-0979-06-01126-8.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1611835114. URL http://arxiv.org/abs/1612.00796. arXiv:1612.00796 [cs].

Merrill, W., Tsilivis, N., and Shukla, A. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. In *ICLR Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023. URL https://arxiv.org/abs/2303.11873.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability, October 2023. URL http://arxiv.org/abs/2301.05217. arXiv:2301.05217 [cs].

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets, January 2022. URL http://arxiv.org/abs/2201.02177. arXiv:2201.02177 [cs].

Tanaka, H., Kunin, D., Yamins, D. L. K., and Ganguli, S. Pruning neural networks without any data by iteratively conserving synaptic flow. In *NeurIPS*, 2020. URL https://arxiv.org/abs/2006.05467.