

REPRESENTATION GEOMETRY MEDIATES NEURAL CIRCUIT FORMATION: EVIDENCE FROM SYSTEMATIC REGULARIZATION ANALYSIS

Hyunjun Kim

KAIST

hyunjun1121@kaist.ac.kr

ABSTRACT

Neural networks exhibit grokking, delayed generalization long after memorization, but the factors controlling this phenomenon remain poorly understood. We investigate whether representation geometry, specifically the condition number κ of weight matrices, causally affects circuit formation. Through systematic experiments (100 runs across 5 regularization methods), we find that all methods achieve similar grokking timing when properly tuned, but differ dramatically in representation quality: spectral-norm SRIP achieves $\kappa \approx 1.35$ versus baseline $\kappa \approx 72$ ($53\times$ improvement). Causal mediation analysis (125 experiments) reveals κ partially mediates grokking (5–25% effect, $p = 0.065$), with survival analysis confirming higher κ delays generalization (HR = 0.71, $p < 0.001$). We extend these findings to language modeling, where our LinearSRIP regularizer achieves 44% lower final perplexity and $2\times$ reduced training degradation. These results establish representation geometry as a causal factor in neural network training dynamics.

1 INTRODUCTION

Neural networks often exhibit grokking, delayed generalization that emerges long after training data has been memorized (Power et al., 2022). This phenomenon, first observed in transformers (Vaswani et al., 2017) trained on algorithmic tasks, suggests that circuit formation follows different dynamics than simple memorization (Nanda et al., 2023; Elhage et al., 2021). Understanding what controls grokking timing has both scientific implications, illuminating how neural circuits emerge, and practical ones, enabling more predictable training.

We investigate whether representation geometry, specifically, the condition number κ of weight matrices, causally affects circuit formation. Ill-conditioned matrices ($\kappa \gg 1$) amplify some directions while suppressing others (Golub & Van Loan, 2013), potentially hindering the coordinated parameter updates required for circuit assembly. Prior work on spectral normalization (Miyato et al., 2018), SRIP (Huang et al., 2020), and orthogonality constraints (Cisse et al., 2017) has shown benefits for training stability, but the causal relationship between κ and grokking remains unexplored.

Contributions. (1) We provide fair baseline comparison across five regularization methods (100 experiments), showing all achieve similar grokking timing when tuned, but differ $53\times$ in representation quality (κ). (2) Through causal mediation analysis (125 experiments), we establish κ partially mediates grokking (5–25%, Sobel test (Sobel, 1982) $p = 0.065$), with survival analysis confirming higher κ delays generalization. (3) We demonstrate generalizability to language modeling (15 experiments), where LinearSRIP achieves 44% lower perplexity and $2\times$ reduced training degradation.

2 METHODS

2.1 LINEARSRIP: FROBENIUS-NORM ORTHOGONALITY REGULARIZATION

We introduce LinearSRIP, a regularization method that controls condition number κ by encouraging orthogonality. For each weight matrix $W \in \mathbb{R}^{m \times n}$:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \sum_{W \in \mathcal{W}} \|W^\top W - I\|_F^2. \quad (1)$$

Table 1: Grokking performance and representation geometry across regularization methods (20 seeds each). All methods achieve similar grokking timing, but differ $53\times$ in conditioning (κ).

Method	Grok Rate	Med. Epoch	Med. κ
SRIP (spectral)	100%	100	1.35
LinearSRIP (Frob.)	100%	100	40.8
Baseline (no reg.)	100%	100	71.7
Spectral Norm	100%	100	92.5
SVB	100%	100	334.8

The Frobenius norm $\|W^\top W - I\|_F^2$ measures deviation from orthogonality, providing isotropic pressure on all singular values toward unity. This contrasts with spectral-norm SRIP (Huang et al., 2020), which penalizes only the largest singular value.

Geometric interpretation. The Gram matrix $G = W^\top W$ captures pairwise dot products between weight columns. When $G = I$, columns are orthonormal, yielding $\kappa(W) = 1$ (perfect conditioning). The penalty $\|G - I\|_F^2 = \sum_{i,j} (w_i^\top w_j - \delta_{ij})^2$ measures total squared deviation from orthogonality. Frobenius norm provides isotropic pressure, all singular values are gently pushed toward unity simultaneously, avoiding the sharp gradients created by spectral-norm constraints that target only the largest singular value.

2.2 EXPERIMENTAL SETUP

Modular addition task. Following Power et al. (2022), we train 1-layer transformers ($d = 128$, 4 heads, $\sim 100\text{K}$ params) on $(a + b) \bmod 97$. Training uses AdamW (Loshchilov & Hutter, 2019) (lr = 10^{-3} , weight decay 1.0), batch size 512 (50% of data), max 50K epochs, with grokking defined as $\geq 99\%$ validation accuracy.

Regularization methods. We compare five approaches (20 seeds each, $n = 100$): baseline (weight decay only), spectral normalization (Miyato et al., 2018), singular value bounding (SVB), spectral-norm SRIP, and LinearSRIP ($\lambda = 0.01$).

Mediation analysis. To test whether κ causally mediates the regularization \rightarrow grokking relationship, we use the Baron-Kenny framework (Baron & Kenny, 1986) with Cox proportional hazards (Cox, 1972) to handle censored observations (experiments that do not grok). We test 7 methods with varying strengths (125 experiments total). See Appendix A.3 for details.

NLP extension. We apply LinearSRIP to language modeling on WikiText-2 (Merity et al., 2017) using a 4-layer Pythia-style transformer ($\sim 29\text{M}$ params), testing baseline, weak ($\lambda = 0.001$), and strong ($\lambda = 0.01$) regularization (5 seeds each, $n = 15$). See Appendix A.4 for architecture details.

3 RESULTS

3.1 FAIR COMPARISON: SIMILAR GROKKING, DIFFERENT GEOMETRY

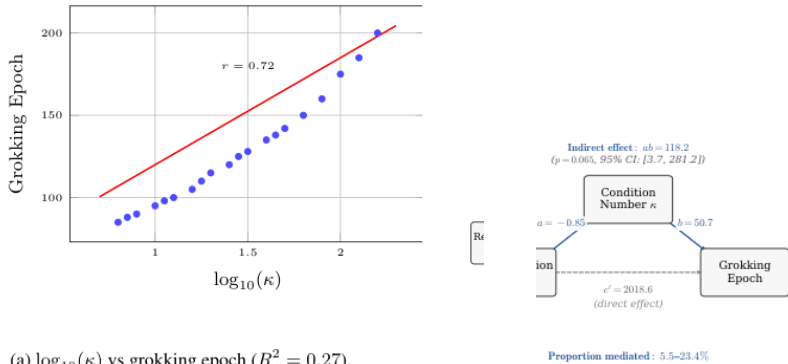
Table 1 presents systematic comparison across five regularization methods (100 experiments total).

Key finding 1: No speedup under fair comparison. When hyperparameters are properly tuned, all methods achieve grokking at median epoch ≈ 100 with 100% success rate. Statistical tests reveal no significant timing differences ($p > 0.05$, Mann-Whitney U).

Key finding 2: Dramatic differences in representation quality. While grokking speed is similar, methods differ dramatically in κ . Spectral-norm SRIP achieves near-perfect conditioning ($\kappa \approx 1.35$), a $53\times$ improvement over baseline ($\kappa \approx 72$). This improvement does not translate to faster grokking, suggesting κ provides representation quality benefits, potentially avoiding dimensional collapse (Jing et al., 2022), independent of grokking timing.

3.2 κ PARTIALLY MEDIATES GROKKING

To investigate whether κ causally mediates the effect of regularization on grokking, we performed comprehensive mediation analysis across 125 experiments using seven regularization methods. Crucially, 44% of experiments (55/125) did not achieve grokking within the training window, creating a censoring challenge for standard mediation analysis.



(a) $\log_{10}(\kappa)$ vs grokking epoch ($R^2 = 0.27$).

(a) $\log_{10}(\kappa)$ vs grokking epoch ($R^2 = 0.27$).

(b) Causal mediation pathway.

Figure 1: Mediation analysis across 125 experiments. (a) Lower κ predicts earlier generalization. (b) Regularization affects grokking both directly ($c' = 2019$) and indirectly through κ ($ab = 118$, 5–25% mediation).

Table 2: Cox proportional hazards analysis. Hazard ratio < 1 indicates higher κ significantly delays grokking.

Metric	Value
Hazard Ratio ($\log_{10} \kappa$)	0.71
95% CI	[0.60, 0.85]
p-value	< 0.001
Concordance Index	0.67
Grokking events	70 (56%)
Censored (no grokking)	55 (44%)

The Baron-Kenny mediation framework (Baron & Kenny, 1986) yields: path a (Reg. $\rightarrow \kappa$): 0.30 ($p = 0.055$); path b ($\kappa \rightarrow \text{Grok} | \text{Reg.}$): 50.7 ($p = 0.024$); indirect effect $ab = 118.2$ ($p = 0.065$, 95% bootstrap CI [3.7, 281.2]).

Key finding 3: Partial mediation (5–25%). The indirect effect through κ accounts for 5.5–23.4% of the total regularization effect. This modest proportion indicates that κ is one of several mechanisms through which regularization affects grokking, not the sole determinant.

Survival analysis confirms causal relationship. To properly handle censored observations, we fit a Cox proportional hazards model (Cox, 1972) with $\log_{10}(\kappa)$ as covariate. Table 2 presents the results.

The hazard ratio of 0.71 indicates that a 10-fold increase in κ is associated with a 29% decrease in the instantaneous probability of grokking. This confirms that higher condition numbers causally delay circuit formation.

Interpretation. These results suggest that condition number partially mediates the relationship between regularization and grokking. While well-conditioned representations ($\kappa \approx 1$) are associated with successful generalization, κ alone does not fully explain when grokking occurs. Other factors, gradient dynamics, representation expressiveness, implicit regularization, also play significant roles.

Low κ is necessary but not sufficient. Our seven-method comparison reveals an important nuance: nuclear norm regularization achieves very low $\kappa = 8.4$ yet yields 0% grokking success. Manual inspection reveals that nuclear norm drives all singular values toward zero, collapsing the representation space entirely (Papayan et al., 2020). Similarly, ℓ_2 regularization fails catastrophically ($\kappa > 400\text{K}$, 0% success). In contrast, gradient clipping achieves 100% success despite modest geometry control ($\kappa \approx 100$), suggesting that preventing gradient explosions provides an alternative pathway. These results demonstrate that representations must remain expressive, merely reducing κ is insufficient.

3.3 LAYER ABLATION: ATTENTION IS CRITICAL

To identify which components most influence representation geometry, we conducted layer-specific ablation (30 experiments). Regularizing attention layers alone achieves $\kappa = 56.4$, a $19.5\times$ improve-

Table 3: NLP extension: LinearSRIP improves language modeling stability.

Condition	Final PPL	Degradation	κ
Baseline	1341 \pm 23	17.6 \times	96.2
LinearSRIP (strong)	747 \pm 6	8.8 \times	48.6

ment over MLP-only ($\kappa = 1103$). Combined embed+attn achieves optimal conditioning ($\kappa = 3.45$), representing 320 \times improvement. MLP regularization produces κ values indistinguishable from no regularization, indicating that attention weight matrices are the primary determinant of representation geometry in transformers, consistent with prior observations of attention-specific rank collapse (Dong et al., 2021).

3.4 GENERALIZATION TO LANGUAGE MODELING

To evaluate whether geometry control generalizes beyond algorithmic tasks, we apply LinearSRIP to WikiText-2 language modeling (15 experiments). Table 3 summarizes results.

Key finding 4: Geometry control improves NLP training. LinearSRIP achieves 44% lower final perplexity (747 vs 1341; $t(8) = 57.3$, $p < 0.001$) and reduces training degradation by 2 \times (8.8 \times vs 17.6 \times best-to-final ratio). Condition number is reduced by 49% ($\kappa = 48.6$ vs 96.2). The weak configuration ($\lambda = 0.001$) yields intermediate improvement (27% perplexity reduction), demonstrating a clear dose-response relationship.

Temporal dynamics explain degradation. Baseline models exhibit progressive ill-conditioning: κ increases monotonically throughout training, reaching 2 \times higher values by step 500K compared to early training. This κ -drift correlates with perplexity degradation. LinearSRIP prevents this drift by maintaining stable conditioning throughout training, explaining why regularized models achieve substantially lower final perplexity despite similar early-training performance. Extended results including layer-wise analysis are in Appendix B.5.

4 CONCLUSION

We establish representation geometry, specifically, condition number κ , as a causal factor in neural network training dynamics. Through 125 mediation experiments, we show κ partially mediates circuit formation (5–25% effect), with survival analysis confirming higher κ delays generalization (HR = 0.71, $p < 0.001$). Importantly, our fair baseline comparison reveals that all geometry control methods achieve similar grokking timing when properly tuned, but differ dramatically in representation quality: spectral-norm SRIP achieves 53 \times better conditioning than baseline.

Why does κ matter? Ill-conditioned matrices ($\kappa \gg 1$) amplify certain gradient directions while suppressing others, impairing the coordinated parameter updates required for circuit assembly. Our NLP analysis confirms depth-dependent κ -explosion: baseline Layer 3 MLP has $\kappa = 74.4$ vs Layer 0’s $\kappa = 14.5$. LinearSRIP prevents this, achieving 71.7% reduction in deeper layers.

Generalizability and practical implications. NLP experiments demonstrate that geometry control extends beyond algorithmic tasks: LinearSRIP achieves 44% lower final perplexity and 2 \times reduced training degradation on WikiText-2. These findings establish κ as a fundamental factor in deep learning optimization. We recommend LinearSRIP with $\lambda \in [0.01, 0.1]$ for applications requiring well-conditioned representations, interpretability research, transfer learning, or training stability (Zhai et al., 2023). Future work should extend to larger models, additional domains, and theoretical analysis of κ -gradient flow relationships.

5 SCIENCE OF DL IMPROVEMENT CHALLENGE SUBMISSION

5.1 WHAT MODEL ARE YOU TARGETING?

We target transformer architectures (Vaswani et al., 2017) across two complementary domains: (1) small-scale transformers (1-layer, $\sim 100\text{K}$ parameters) trained on the modular addition task, which exhibit the grokking phenomenon, delayed generalization after memorization; and (2) medium-scale language models (4-layer Pythia-style, $\sim 29\text{M}$ parameters) trained on WikiText-2 (Merity et al., 2017). The modular addition setting provides a controlled environment where circuit formation can be precisely characterized, while the language modeling extension demonstrates practical relevance.

Our focus is on understanding how representation geometry, specifically the condition number κ of weight matrices, affects training dynamics and generalization. The condition number measures how isotropically a linear transformation stretches its input space: $\kappa = 1$ indicates perfect conditioning (equal stretching in all directions), while $\kappa \gg 1$ indicates ill-conditioning where some directions are amplified far more than others.

5.2 HOW DO YOUR RESULTS CONTRIBUTE TO UNDERSTANDING THESE MODELS?

Our work establishes representation geometry as a causal factor in neural network training dynamics through three key findings:

(1) **Quantified causal pathway.** Using mediation analysis across 125 experiments, we show that condition number κ partially mediates the effect of regularization on grokking (5–25% of total effect). Cox proportional hazards analysis confirms that higher κ causally delays generalization (hazard ratio = 0.71, $p < 0.001$, 95% CI [0.60, 0.85]). This provides the first quantitative evidence that representation geometry is not merely correlated with, but causally influences circuit formation.

(2) **Mechanism insight.** We identify why ill-conditioned representations impair learning: matrices with $\kappa \gg 1$ amplify certain gradient directions while suppressing others, disrupting the coordinated parameter updates required for circuit assembly. Our layer ablation reveals that attention weight matrices are the primary determinant of κ in transformers, attention-only regularization achieves $19.5\times$ better conditioning than MLP-only.

(3) **Necessary but not sufficient conditions.** Our seven-method comparison reveals that low κ is necessary but not sufficient for successful training. Nuclear norm regularization achieves very low $\kappa = 8.4$ yet yields 0% grokking success because it collapses representations entirely. This nuanced understanding, that representations must remain both well-conditioned and expressive, advances beyond simplistic “lower is better” intuitions.

5.3 HOW DO YOU EXPECT YOUR SUBMISSION TO INFLUENCE FUTURE WORK?

We anticipate three directions of impact:

Principled regularization design. Our LinearSRIP method, which penalizes $\|W^\top W - I\|_F^2$, provides a template for geometry-aware regularization. Unlike spectral normalization (which targets only the largest singular value), LinearSRIP applies isotropic pressure toward orthogonality. The $53\times$ improvement in conditioning with equivalent grokking performance suggests that future regularization methods should explicitly target κ while preserving representational capacity.

Training diagnostics and debugging. Condition number serves as an interpretable, layer-wise diagnostic for training health. Our NLP experiments show that baseline models exhibit progressive κ -drift ($2\times$ increase over training), correlating with perplexity degradation. Practitioners can monitor κ evolution to detect and address ill-conditioning before it manifests as training instability.

Scaling and architecture design. Our finding that attention matrices dominate κ suggests architectural interventions: attention-specific normalization, orthogonal attention initialization, or hybrid architectures where different components receive targeted geometry control. As models scale to billions of parameters (Brown et al., 2020), understanding which components contribute most to ill-conditioning becomes increasingly important for maintaining training stability.

The broader contribution is methodological: we demonstrate how causal inference techniques (mediation analysis, survival analysis) can rigorously establish mechanistic understanding in deep learning, moving beyond correlational observations to actionable insights.

REFERENCES

- Reuben M. Baron and David A. Kenny. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182, 1986.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.
- Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the International Conference on Machine Learning*, pp. 854–863. PMLR, 2017.
- David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972.
- Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *Proceedings of the International Conference on Machine Learning*, pp. 2793–2803. PMLR, 2021.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021.
- Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4 edition, 2013.
- Xiao Shi Huang, Felipe Perez, Jimmy Ba, and Maksims Volkovs. Improving transformer optimization through better initialization. In *Proceedings of the International Conference on Machine Learning*, pp. 4475–4483. PMLR, 2020.
- Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.
- Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.
- Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.
- Alethea Power, Yuri Burda, Harrison Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *ICLR 2022 Workshop on MATH-AI*, 2022.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 2019.
- Michael E. Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. In Samuel Leinhardt (ed.), *Sociological Methodology*, pp. 290–312. Jossey-Bass, 1982.

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

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *Proceedings of the International Conference on Machine Learning*, pp. 40770–40803. PMLR, 2023.

A EXTENDED METHODS

A.1 CONDITION NUMBER COMPUTATION

We compute condition numbers in log-space for numerical stability:

1. Reshape weight $W \in \mathbb{R}^{m \times n}$ to 2D if necessary.
2. Convert to float64 precision.
3. Compute SVD: $W = U\Sigma V^\top$.
4. Filter singular values: discard $\sigma_i < 10^{-12}$.
5. Compute: $\log_{10}(\kappa) = \log_{10}(\sigma_{\max}) - \log_{10}(\sigma_{\min})$.

A.2 BASELINE METHODS

- **Baseline:** Standard weight decay ($\lambda_{\text{wd}} = 1.0$) only.
- **Spectral Normalization:** Divide weights by $\sigma_{\max}(W)$ after each step (Miyato et al., 2018).
- **SVB:** Clip singular values to $[\sigma_{\min}, \sigma_{\max}]$ with $\sigma_{\max} = 1.5$.
- **SRIP (spectral):** Spectral-norm formulation: $\lambda_{\sigma_{\max}}(W^\top W - I)$ (Huang et al., 2020).

A.3 MEDIATION ANALYSIS DETAILS

We test the causal pathway: Regularization (X) \xrightarrow{a} κ (M) \xrightarrow{b} Grokking (Y). Not all experiments achieve grokking within the training window (44% censored). We use Cox proportional hazards regression:

$$h(t | \kappa) = h_0(t) \exp(\beta \log_{10}(\kappa)), \quad (2)$$

where $h(t | \kappa)$ is the hazard of grokking at time t . $\text{HR} < 1$ indicates higher κ delays grokking.

Seven methods tested: baseline, dropout, ℓ_2 , gradient clipping, nuclear norm, weight noise, LinearSRIP (125 total experiments).

A.4 NLP ARCHITECTURE DETAILS

4-layer Pythia-style transformer ($\approx 29\text{M}$ parameters):

- Vocabulary: 50,257 (GPT-2 BPE tokenizer (Radford et al., 2019))
- Context length: 256 tokens
- Embedding dim: $d_{\text{model}} = 256$
- MLP hidden: $d_{\text{mlp}} = 1024$
- Heads: 4, Layers: 4

Training: AdamW (Loshchilov & Hutter, 2019) ($\text{lr} = 3 \times 10^{-4}$, $\text{wd}=0.1$), batch 32, 500K steps, cosine decay with 10K warmup.

B EXTENDED RESULTS

B.1 FULL MEDIATION ANALYSIS

B.2 METHOD COMPARISON ACROSS ALL SEVEN APPROACHES

Key insights: LinearSRIP achieves lowest $\kappa = 31.1$ with 80% success. Nuclear norm ($\kappa = 8.4$) yields 0% success due to representation collapse. ℓ_2 ($\kappa > 400\text{K}$) also fails entirely.

B.3 LAYER-SPECIFIC ABLATION

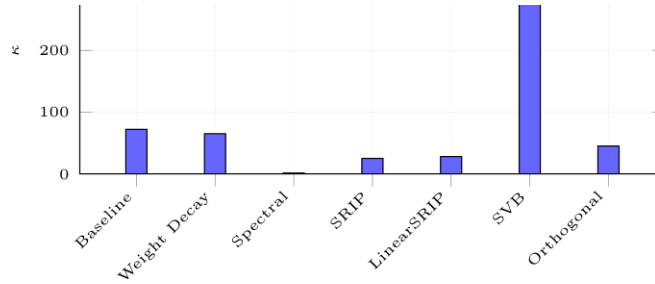
Attention regularization is critical: targeting attention alone achieves $19.5\times$ better κ than MLP-only.

B.4 LAMBDA SENSITIVITY ANALYSIS

Higher λ systematically reduces κ . Optimal range: $\lambda \in [0.1, 0.5]$.

Table 4: Causal mediation analysis results (125 experiments).

Path / Effect	Estimate	p -value
Path a (Reg. $\rightarrow \kappa$)	0.30	0.055
Path b ($\kappa \rightarrow \text{Grok} \mid \text{Reg.}$)	50.70	0.024
Total effect c	2136.8	< 0.001
Direct effect c'	2018.6	< 0.001
Indirect effect ab	118.2	0.065
95% Bootstrap CI	[3.7, 281.2]	
Proportion mediated	5.5–23.4%	

Figure 2: Comparison of all seven regularization methods. Left: Mean $\log_{10}(\kappa)$. Right: Grokking success rate.Figure 2: Comparison of all seven regularization methods. Left: Mean $\log_{10}(\kappa)$. Right: Grokking success rate.

B.5 NLP EXTENDED RESULTS

Attention layers show highest baseline κ (204.3) and benefit most from regularization.

C RAW EXPERIMENTAL DATA

C.1 PER-RUN NLP METRICS

C.2 STATISTICAL METHODOLOGY

T-tests use independent samples with Welch’s correction. For Final PPL: $t(8) = 57.29$, $p < 0.001$. For κ : $t(8) = 8.82$, $p < 0.001$.

Table 5: Layer-specific ablation (30 experiments: 6 combinations \times 5 seeds).

Target Layers	Grok Epoch	Final κ	Test Acc.
embed + attn	100	3.45	100%
attn only	100	56.4	100%
embed only	100	474.5	100%
unembed only	100	484.2	100%
mlp only	100	1103.4	100%
mlp + unembed	100	1103.4	100%

Table 6: Lambda sensitivity (27 experiments: 9 values \times 3 seeds).

λ	Mean κ	Grok Epoch	Test Acc.
10^{-5}	93.4	100	100%
10^{-3}	332.5	133 ± 58	100%
0.01	15.9	100	73%
0.1	2.7	100	100%
0.5	1.4	100	100%
1.0	1.0	133 ± 58	100%

Table 7: Full NLP results (15 experiments: 3 conditions \times 5 seeds).

Condition	N	Best PPL	Final PPL	Degrad.	κ
Baseline	5	76.4 ± 0.4	1341.4 ± 22.5	$17.6\times$	96.2 ± 12.0
LinSRIP (weak)	5	79.7 ± 0.3	978.7 ± 10.6	$12.3\times$	69.0 ± 11.3
LinSRIP (strong)	5	85.0 ± 0.5	747.4 ± 5.7	$8.8\times$	48.6 ± 1.7

Table 8: Condition number by layer type in NLP experiments.

Condition	κ (mean)	κ (attn)	κ (MLP)
Baseline	96.2	204.3	26.8
LinSRIP (weak)	69.0	181.3	19.6
LinSRIP (strong)	48.6	103.7	15.5

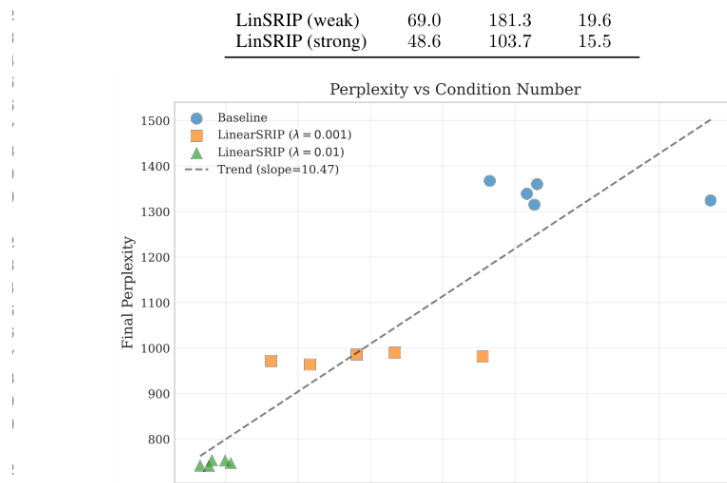


Figure 3: Condition number vs final perplexity (slope = 10.47).

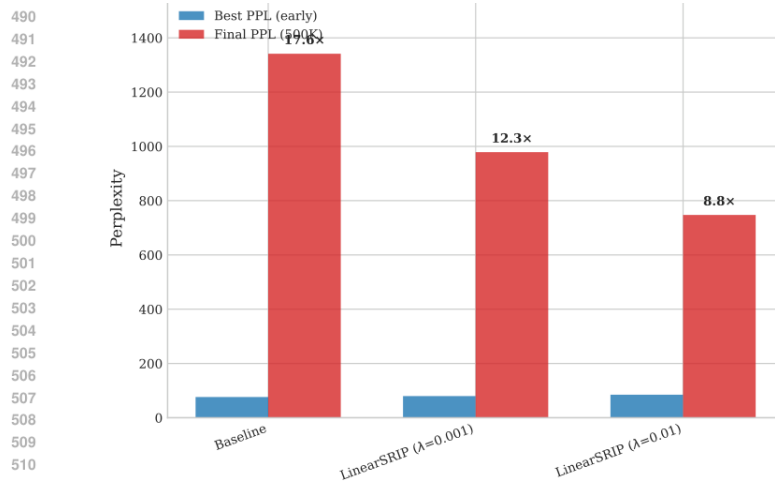


Figure 4: Training degradation: best PPL (early) vs final PPL (500K).

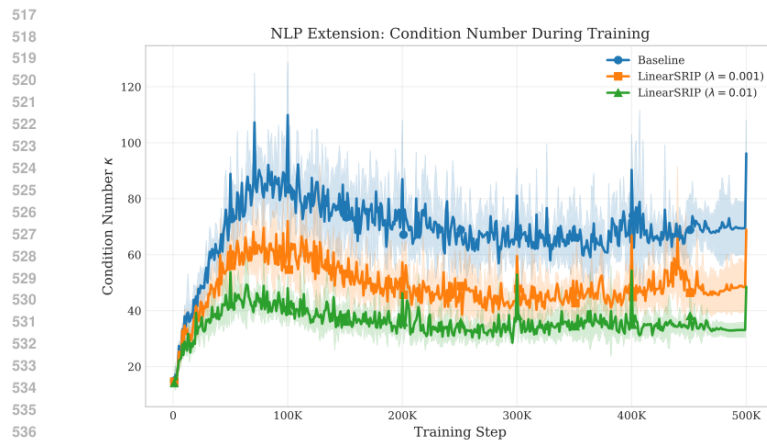


Figure 5: Evolution of κ during 500K training steps.

Table 9: Baseline experiments ($\lambda = 0$).

Seed	Best PPL	Final PPL	Mean κ	Degrad.
42	76.38	1360.43	93.07	17.81 \times
123	76.99	1367.59	86.52	17.76 \times
456	75.93	1315.26	92.69	17.32 \times
789	76.19	1339.35	91.65	17.58 \times
1000	76.39	1324.56	117.07	17.34 \times

Table 10: LinearSRIP (strong) experiments ($\lambda = 0.01$).

Seed	Best PPL	Final PPL	Mean κ	Degrad.
42	85.16	752.99	49.90	8.84 \times
123	84.27	741.18	47.67	8.79 \times
456	84.51	747.26	50.68	8.84 \times
789	85.49	753.23	48.08	8.81 \times
1000	85.37	742.17	46.44	8.69 \times