Structure and Scale in Simplicial Sequence Modelling

Matthew Farrugia-Roberts MATTHEW@FAR.IN.NET
Department of Computer Science, University of Oxford

Abstract

Modern large-scale deep learning exhibits two striking empirical phenomena: behavioural scaling laws (predictable performance gains with increasing scale) and emergent mechanisms (structured internal representations and circuits in deep neural networks). We hypothesise that these two phenomena are connected: that predictable changes in behaviour are the result of predictable changes in internal computational structure. In this paper, we report preliminary evidence of such a connection. We find a correlation between scaling patterns in performance and representations in small transformers trained to predict the outputs of a hidden Markov model, for which residual activations are known to linearly encode a belief distribution over latent states in a probability simplex.

Keywords: Science of deep learning, scaling laws, developmental interpretability.

1. Introduction

Through modern advancements in software [13, 30], hardware [3], and neural network architecture [5, 23, 62], it has become possible to train increasingly large neural networks on increasingly large amounts of data [e.g., 8, 42, 43, 51, 52]. The resulting models have driven remarkable growth in applications of AI throughout society [56]. While these applications have captured the world's attention, from the perspective of a *natural scientist,* the most interesting empirical phenomena in recent AI history are the following:

1. **Examples of interpretable internal structure:** Sometimes, trained neural networks store interpretable internal representations of relevant variables [e.g., 29, 50, 40, 18, 32, 37, 57, 54, 60, 9, 20, 22, 4], and operate on these representations using interpretable circuits [e.g., 40, 35, 36, 66, 17, 41, 65, 47, 61].

2. **Principled variation in behaviour with scale:** The behaviour of trained neural networks follows predictable patterns, such as performance power laws, as a function of the scale of training (the number of data points, model parameters, or optimisation steps) over many orders of magnitude [e.g., 24, 55, 28, 25, 6].

We hypothesise that these two phenomena are each reflections of a third underlying phenomenon:

3. **Principled variation in structure with scale:** the *degree and kind of computational structure present inside* trained neural networks follow predictable patterns as a function of scale.

Hypothetically, it is these changes in structure that drive changes in behaviour—in contrast to models of scaling that view neural networks as tabular function approximators [27, 34, 58]. Moreover, examples of interpretable mechanisms and representations appear when the underlying changes in structure happen to surface as particularly interpretable—a salient special case against a broader backdrop of unexplained frontier model cognition [38, 39].

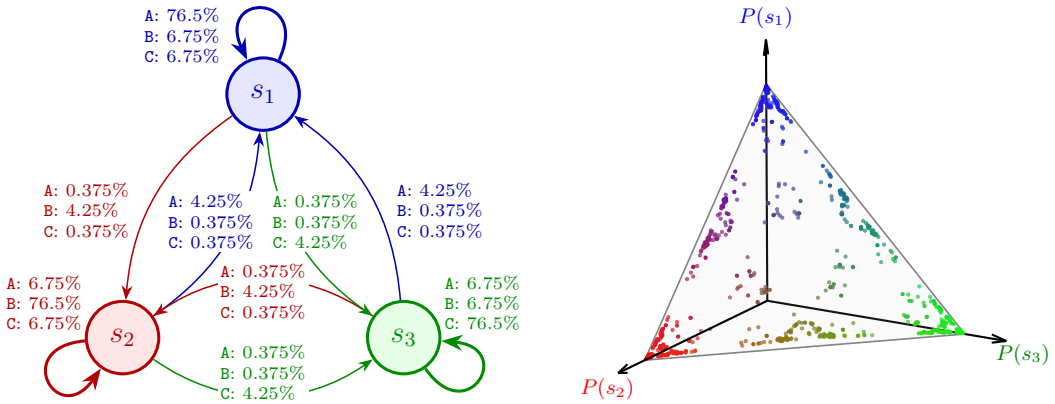

Figure 1: **Simplicial sequence modelling.** *Left:* Hidden Markov model sequence generator. Edge annotations denote probability of transitioning between latent states s_1, s_2, s_3 while emitting symbols $\texttt{A}, \texttt{B}, \texttt{C}$. *Right:* Probability simplex of belief distributions over latent states s_1, s_2, s_3. Scattered points are Bayesian belief distributions from a sample of sequences.

In this paper, we investigate this hypothesis by studying a known example of structured representations from the mechanistic interpretability literature, namely the simplicial sequence modelling setting of Shai et al. [57]. In this setting, we show that with increasing scale (training steps), the internal representations become increasingly refined, roughly following a power law.

2. Background

We generated a data set of observation sequences of length 10 sampled from an edge-emitting hidden Markov model with three latent states (Figure 1, left). The specific hidden Markov model is the "mess3" process studied by Marzen and Crutchfield [33] with parameters $\alpha = 0.85$, $x = 0.05$.

Optimally predicting the next observation given 0–9 previous observations requires computing the Bayesian posterior belief distribution over latent states and averaging next-symbol probabilities over this uncertainty. These belief distributions reside in the 2-simplex of distributions over three latent states (Figure 1, right).

Shai et al. [57] trained transformers on this data distribution and found that their transformers linearly encode Bayes-like belief distributions in final layer activations. They found that these representations were less strongly encoded at select mid-training checkpoints. We extend this work into an in-depth mechanistic scaling experiment as follows.

3. Experiments

We trained transformers to predict mess3 sequences using cross-entropy loss. We used a fixed architecture with four residual transformer blocks, each with one attention head of width 8 and one MLP layer of width 256 for a total of approximately 142 thousand parameters. We trained with stochastic mini-batch gradient descent using learning rate 0.01 and batches of 64 sequences. We use a transformer training implementation in JAX [7] derived from Farrugia-Roberts [21]. Each training run took approximately 3.7 hours per 10 million training steps on a TPU v4-2 device.

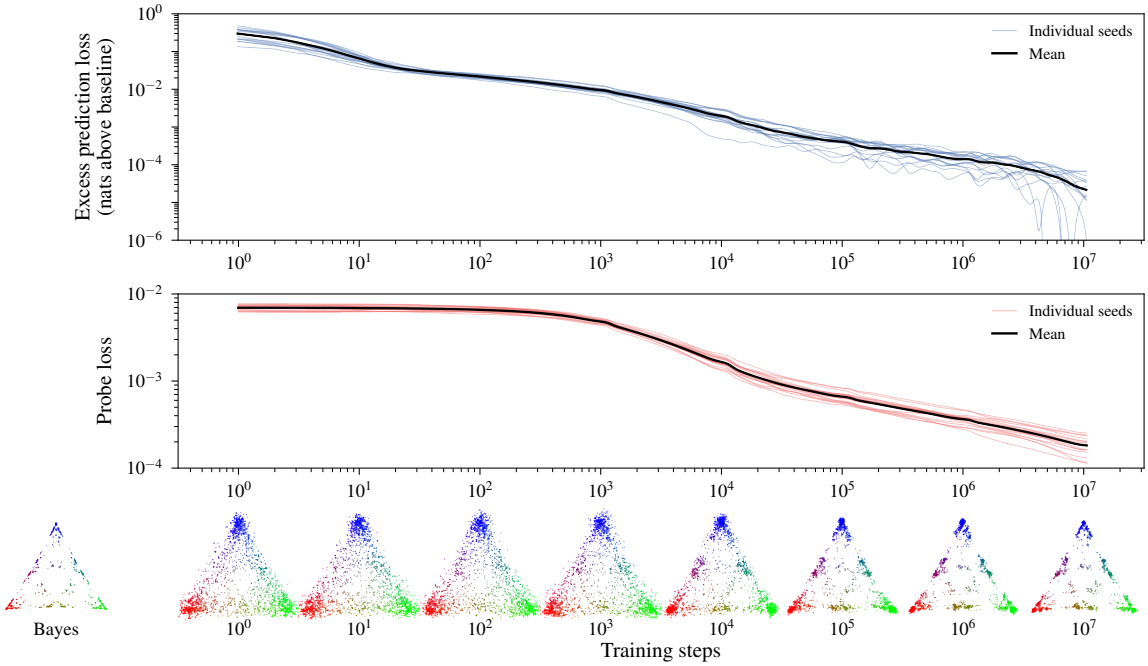

Figure 2: **Performance/representation scaling of simplicial sequence models.** *Top:* Per-step held-out test cross-entropy minus per-seed irreducible cross-entropy. Mean (black) of 16 seeds (blue), light log-Gaussian smoothing (see Appendix A). *Middle:* Mean squared error predicting Bayesian belief distributions from final-layer activations. Mean (black) of 16 seeds (red), light log-Gaussian smoothing. *Bottom:* Bayesian posterior beliefs ("Bayes") and reconstructed transformer beliefs at select training steps (1 seed).

Key variables. We scale the number of training steps while holding architecture, number of parameters, and data generator constant. Compute scales linearly with training steps, as does the number of unique training sequences (we generate fresh sequences for every batch). We continually monitor the following observables (for the first 1,000 steps, and thereafter at ≈100 steps per OOM).

1. **Excess prediction loss:** Mean next-token prediction cross-entropy loss across a fixed held-out batch of 1024 sequences, minus Bayes-optimal cross-entropy on that batch.

2. **Probe loss:** Mean squared error of a linear probe [2] reconstructing ground-truth Bayesian belief distributions from the transformer's final layer activations. (Mean taken across a fixed held-out batch of 1024 sequences, independent of sequences used to train the probe.)

Representations improve smoothly with scale. We trained 16 transformers for 10 million training steps (10 times longer than Shai et al.). Figure 2 shows the resulting trends in excess prediction loss and probe loss, along with a visualisation of belief distributions reconstructed from activations via trained probes at select steps. We see that excess prediction loss continually improves with increasing compute (roughly as a power law). Continuously measuring probe loss reveals that it quickly also enters a regime of continual improvement (again, roughly as a power law). The belief reconstruction visualisations show that the fractal geometry is more crisply represented with scale.

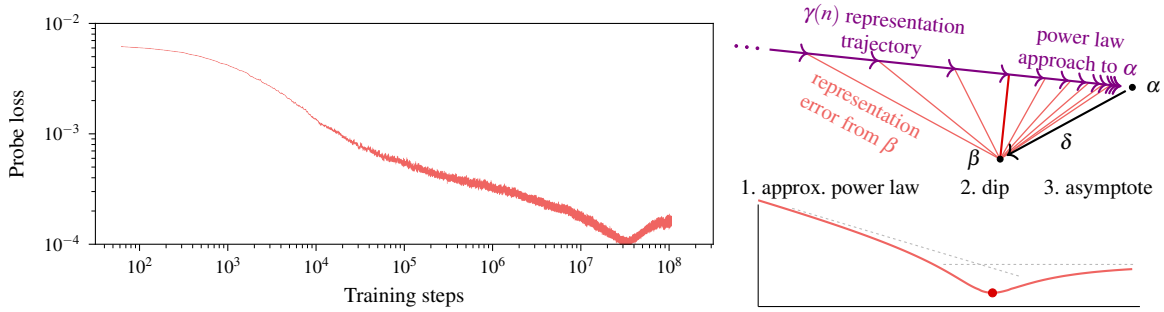

Figure 3: **Long-run representation scaling in a simplicial sequence model.** *Left:* Mean squared error predicting Bayesian belief distributions from final-layer activations (every 64 steps, no smoothing). *Right:* Toy model of dynamics in representation subspace.

Asymptotic versus ideal representations. We trained one transformer for 100 million steps (100 times longer than Shai et al.). Figure 3 shows that between 10 million and 100 million training steps, the probe loss is non-monotonic, turning around and then apparently beginning to plateau.

If the transformer's representations were asymptotically approaching the ideal Bayesian belief distributions, then we would expect monotonic scaling to continue indefinitely. However, failure to converge to the Bayesian ideal does not preclude convergence to an alternative, realisable representation. Indeed, finite-sized transformers cannot express the Bayesian posterior exactly [47], so they might asymptote to a realisable representation. Below we provide one toy model of representation scaling that recovers the power-law–dip–plateau pattern.

Let $\mathbb{R}^d$ be the joint space of final-layer transformer activations within some residual subspace for each sequence in a large fixed batch. Let $\gamma : \mathbb{N} \to \mathbb{R}^d$ represent the trajectory of a transformer's activations in this subspace over training, so that after $n \in \mathbb{N}$ training steps, the transformer's activations are $\gamma(n) \in \mathbb{R}^d$. Suppose these structures converge to some particular activation vector $\alpha \in \mathbb{R}^d$ with increasing training steps, that is, $\gamma(n) \to \alpha$. Moreover, suppose this convergence proceeds as a power law in squared error,

$$\|\gamma(n) - \alpha\|_2^2 = An^{-E} \tag{1}$$

for some $A, E > 0$. Let $\beta \in \mathbb{R}^d$ represent an encoding of the corresponding idealised Bayesian belief distributions. If $\beta = \alpha$ then we have that the transformer's representations converge to encoding Bayesian belief distributions with power law squared error. However, we might suppose instead that $\beta = \alpha + \delta$ for some small distance $\delta \in \mathbb{R}^d$. Then, how does squared representation error (measured against the ideal β) scale? We have by (1)

$$\|\beta - \gamma(n)\|_2^2 = \|\alpha - \gamma(n) + \delta\|_2^2 = An^{-E} - 2(\gamma(n) - \alpha) \cdot \delta + \|\delta\|_2^2. \tag{2}$$

For n small enough that $\|\delta\|_2 \ll \sqrt{An^{-E}}$, we have $\|\beta - \gamma(n)\|_2^2 \approx An^{-E}$ by Cauchy–Schwarz. That is, representation error exhibits an approximate power law early on. As $n \to \infty$, we have a plateau, $\|\beta - \gamma(n)\|_2^2 \to \|\delta\|_2^2$. In between, if $(\gamma(n) - \alpha) \cdot \delta > \frac{1}{2}An^{-E}$, the trajectory $\gamma(n)$ bypasses β as it slowly progresses towards α, and we have a dip.

Figure 3 (right) illustrates this dynamic. This model is simple enough that we were able to derive it after seeing the non-monotonicity (at about 50 million steps) and qualitatively predict the approximate plateau in Figure 3 (left) before it emerged.

4. Future work

More careful analysis is needed to determine if these trends in performance and representation error are robustly described by power laws. If there are power laws, each appears to have multiple regimes with different exponents. All we claim for now is that scaling is smooth, continually improving, and *roughly* power-law-shaped, for the first ≈30 million training steps.

Moreover, we establish only a correlation between representation error and prediction error scaling. It is not yet clear whether the smooth improvements in representations *cause* the smooth improvement in performance. Future work should develop methodologies for studying the connection between these observations.

The toy model at the end of Section 3 is merely illustrative. If we make the simplifying assumption of a constant angle subtended by δ and $\gamma(n) - \alpha$ and fit the remaining parameters, the model fails to quantitatively match the observed trends. This suggests that the representations may not be approaching α as a strict power law, though the nascent plateau suggests some kind of convergence.

Finally, it remains to explore trends in representations with increasing parameters (number of layers, number of embedding dimensions), such as those studied by Kaplan et al. [28] for behavioural scaling laws. Additional parameters may facilitate more refined internal representations.

5. Related work

A related example of principled variation in behaviour in deep learning comes from the literature on behavioural "phase transitions" in neural networks as a function of training time or data composition. Examples can be found across deep reinforcement learning [e.g., 14, 15, 31, 1, 19], synthetic sequence modelling [e.g., 11, 59, 53, 63, 44, 10, 16, 45], and language modelling [e.g., 49, 26, 64]. Moreover, there are examples of interpretable internal representations and mechanisms arising in *structural* phase transitions, including induction heads [17, 41, 63, 16], toy models of superposition [18, 12], and grokking [48, 35, 36]. A promising approach to reconciling the principles connecting training dynamics and data composition is developmental interpretability, particularly via singular learning theory [see, e.g., 12, 10, 64, 46]. So far, the implications of *scale* for the formation of internal structure have been under-explored.

6. Conclusion

The two most striking empirical phenomena of the modern deep learning era are scaling laws and examples of emergent computational structure. Neither of these phenomena is yet clearly understood, undermining experts' ability to predict the near-future trajectory and implications of AI technology, and society's ability to steer that trajectory.

We have proposed that these two phenomena are both reflections of some deeper principles driving variations in the emergence and refinement of computational structure in response to variations in scale. If so, understanding scaling laws and emergent mechanisms alike will require first understanding the principles relating structure and scale.

In this direction, we have taken a small first step by studying the variation in a known internal representation mechanism while scaling compute and data. In this simplified setting, we identified an example of a correlation between behavioural improvements with scale and the measurable and non-monotonic refinement of the linear encoding of Bayesian belief distributions over latent states. This is a hint that there is more to behavioural scaling patterns than tabular function approximation.

Acknowledgements

We thank Thomas Bush, Pranav Mahajan, Daniel Murfet, Louis Thomson, and Joan Velja for helpful conversations. Claude Opus 4.6 and 4.7 assisted with experimentation and plot generation. TPUs provided by Google's TPU Research Cloud.

References

[1] Karim Abdel Sadek, Matthew Farrugia-Roberts, Usman Anwar, Hannah Erlebach, Christian Schroeder de Witt, David Krueger, and Michael Dennis. Mitigating goal misgeneralization via minimax regret. *Reinforcement Learning Journal*, 2025. Cited on page 5.

[2] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. Preprint arXiv:1610.01644, 2018. Cited on page 3.

[3] Dario Amodei and Danny Hernandez. AI and compute. OpenAI blog, 2018. Cited on page 1.

[4] Jingmin An, Wei Liu, Qian Wang, and Fang Fang. Time travel engine: A shared latent chronological manifold enables historical navigation in large language models. Preprint arXiv:2601.06437, 2026. Cited on page 1.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. Published as a conference paper at ICLR 2015. Preprint arXiv:1409.0473, 2015. Cited on page 1.

[6] Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. Preprint arXiv:2404.10102, 2024. Cited on page 1.

[7] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. GitHub repository, 2018. URL http://github.com/jax-ml/jax. Cited on page 2.

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, pages 1877–1901, 2020. Cited on page 1.

[9] Thomas Bush, Stephen Chung, Usman Anwar, Adrià Garriga-Alonso, and David Krueger. Interpreting emergent planning in model-free reinforcement learning. In *International Conference on Learning Representations*, 2025. Cited on page 1.

[10] Liam Carroll, Jesse Hoogland, Matthew Farrugia-Roberts, and Daniel Murfet. Dynamics of transient structure in in-context linear regression transformers. Preprint arXiv:2501.17745, 2025. Cited on page 5.

[11] Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *International Conference on Learning Representations*, 2024. Cited on page 5.

[12] Zhongtian Chen, Edmund Lau, Jake Mendel, Susan Wei, and Daniel Murfet. Dynamical versus Bayesian phase transitions in a toy model of superposition. Preprint arXiv:2310.06301, 2023. Cited on page 5.

[13] Dan C. Cireşan, Ueli Meier, Jonathan Masci, Luca M. Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, volume 2, pages 1237–1242, 2011. Cited on page 1.

[14] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1282–1289, 2019. Cited on page 5.

[15] Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 2048–2056, 2020. Cited on page 5.

[16] Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, Eran Malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning markov chains. In *Advances in Neural Information Processing Systems 37*, pages 64273–64311, 2024. Cited on page 5.

[17] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. Cited on pages 1 and 5.

[18] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. Cited on pages 1 and 5.

[19] Chris Elliott, Einar Urdshals, David Quarel, Matthew Farrugia-Roberts, and Daniel Murfet. Stagewise reinforcement learning and the geometry of the regret landscape. Preprint arXiv:2601.07524, 2026. Cited on page 5.

[20] Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, 2025. Cited on page 1.

[21] Matthew Farrugia-Roberts. "Hi, JAX!": An introduction to JAX for deep learning research. GitHub repository, 2026. URL https://github.com/matomatical/hijax. Cited on page 2.

[22] Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. Preprint arXiv:2601.04480, 2026. Cited on page 1.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. Cited on page 1.

[24] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. Preprint arXiv:1712.00409, 2017. Cited on page 1.

[25] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems 35*, pages 30016–30030, 2022. Cited on page 1.

[26] Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape degeneracy and stagewise development in transformers. *Transactions on Machine Learning Research*, 2025. Cited on page 5.

[27] Marcus Hutter. Learning curve theory. Preprint arXiv:2102.04074, 2021. Cited on page 1.

[28] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. Preprint arXiv:2001.08361, 2020. Cited on pages 1 and 5.

[29] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. In *International Conference on Learning Representations (Workshop Track)*, 2016. Cited on page 1.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012. Cited on page 1.

[31] Lauro Langosco, Jack Koch, Lee D. Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12004–12019, 2022. Cited on page 5.

[32] 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*, 2023. Cited on page 1.

[33] Sarah E. Marzen and James P. Crutchfield. Nearly maximally predictive features and their dimensions. *Physical Review E*, 95(5):051301, 2017. Cited on page 2.

[34] Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems 36*, pages 28699–28722, 2023. Cited on page 1.

[35] Neel Nanda and Tom Lieberum. A mechanistic interpretability analysis of grokking. AI Alignment Forum, August 2022. Cited on pages 1 and 5.

[36] 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. Cited on pages 1 and 5.

[37] 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: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30. Association for Computational Linguistics, 2023. Cited on page 1.

[38] Neel Nanda, Josh Engels, Arthur Conmy, Senthooran Rajamanoharan, Bilal Chughtai, Callum McDougall, János Kramár, and Lewis Smith. A pragmatic vision for interpretability. AI Alignment Forum, December 2025. Cited on page 1.

[39] Chris Olah. The dark matter of neural networks? *Transformer Circuits Thread*, July 2024. Cited on page 1.

[40] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024.001, March 2020. Cited on page 1.

[41] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. Cited on pages 1 and 5.

[42] OpenAI. GPT-4 system card. Technical report, OpenAI, 2023. Cited on page 1.

[43] OpenAI. GPT-5 system card. Technical report, OpenAI, 2025. Cited on page 1.

[44] Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the Bayesian prism. In *International Conference on Learning Representations*, 2024. Cited on page 5.

[45] Core Francisco Park, Ekdeep Singh Lubana, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. In *International Conference on Learning Representations*, pages 66381–66433, 2025. Cited on page 5.

[46] Simon Pepin Lehalleur, Jesse Hoogland, Matthew Farrugia-Roberts, Susan Wei, Alexander Gietelink Oldenziel, George Wang, Liam Carroll, and Daniel Murfet. You are what you eat– AI alignment requires understanding how data shapes structure and generalisation. Preprint arXiv:2502.05475, 2025. Cited on page 5.

[47] Mateusz Piotrowski, Paul M. Riechers, Daniel Filan, and Adam S. Shai. Constrained belief updates explain geometric structures in transformer representations. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 49399–49419, 2025. Cited on pages 1 and 4.

[48] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. Preprint arXiv:2201.02177, 2022. Cited on page 5.

[49] Tian Qin, Naomi Saphra, and David Alvarez-Melis. Sometimes I am a tree: Data drives unstable hierarchical generalization in LMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11722–11740, 2025. Cited on page 5.

[50] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. Preprint arXiv:1704.01444, 2017. Cited on page 1.

[51] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. Cited on page 1.

[52] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. Cited on page 1.

[53] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems 36*, pages 14228–14246, 2023. Cited on page 5.

[54] Paul M. Riechers, Thomas J. Elliott, and Adam S. Shai. Neural networks leverage nominally quantum and post-quantum representations. Preprint arXiv:2507.07432, 2025. Cited on page 1.

[55] Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*, 2020. Cited on page 1.

[56] Sha Sajadieh, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Lapo Santarlasci, Juan Pava, Nestor Maslej, Russ Altman, Erik Brynjolfsson, Carla Brodley, Jack Clark, Virginia Dignum, Vipin Kumar, James Landay, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Elham Tabassi, Russell Wald, Toby Walsh, and Dan Weld. The AI index 2026 annual report. Technical report, Institute for Human-Centered AI, Stanford University, April 2026. Cited on page 1.

[57] Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. In *Advances in Neural Information Processing Systems 37*, pages 75012–75034, 2024. Cited on pages 1, 2, 3, and 4.

[58] Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. Preprint arXiv:2004.10802, 2020. Cited on page 1.

[59] Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems 36*, pages 27801–27819, 2023. Cited on page 5.

[60] Mohammad Taufeeque, Philip Quirke, Maximilian Li, Chris Cundy, Aaron David Tucker, Adam Gleave, and Adrià Garriga-Alonso. Planning in a recurrent neural network that plays Sokoban. Preprint arXiv:2407.15421, 2024. Cited on page 1.

[61] Mohammad Taufeeque, Aaron David Tucker, Adam Gleave, and Adrià Garriga-Alonso. Path channels and plan extension kernels: a mechanistic description of planning in a Sokoban RNN. In *International Conference on Learning Representations*, 2026. Cited on page 1.

[62] 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 30*, pages 5998–6008, 2017. Cited on page 1.

[63] George Wang, Matthew Farrugia-Roberts, Jesse Hoogland, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape geometry reveals stagewise development of transformers. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024. Cited on page 5.

[64] George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. Differentiation and specialization of attention heads via the refined local learning coefficient. In *International Conference on Learning Representations*, 2025. Cited on page 5.

[65] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. Cited on page 1.

[66] 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 36*, pages 27223–27250, 2023. Cited on page 1.

Appendix A. Plots without smoothing

Figure 2 incorporates Gaussian smoothing in log-step space with a kernel of width $\sigma = 1/20$ decades. Figure 4 shows the unsmoothed equivalent.

The buildup of samples towards the right side of each log-space decade reflects our measurement schedule. We capture observables for each of the first 1,000 steps, and thereafter at approximately 1,000 linearly-spaced steps per decade.

Moreover, this plot emphasises that the excess prediction loss occasionally drops below zero, saturating the log scale. We use a symmetric log scale to show that (1) the loss descends only modestly below the baseline and (2) it eventually returns positive.

Negative excess prediction loss indicates that the transformers are making better predictions than the Bayesian posterior predictive distribution on the fixed batch of 1024 evaluation sequences. We note that though this is impossible *in expectation over the data generating process,* it is entirely possible for the transformer to out-predict Bayes on any given sequence.

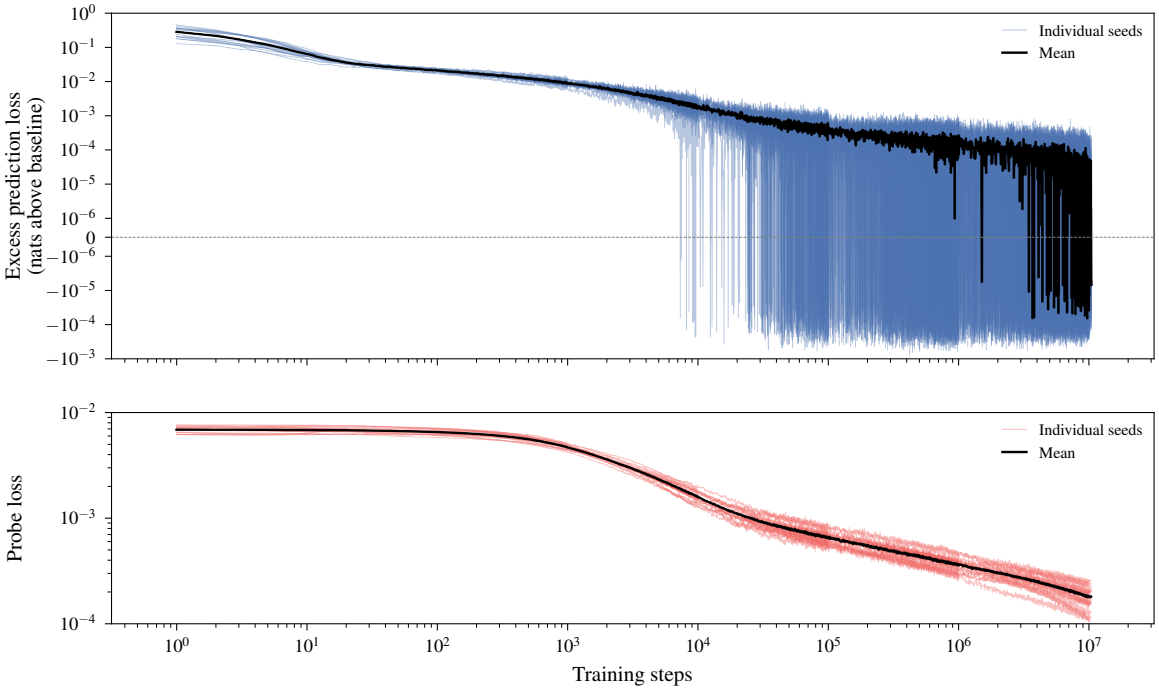

Figure 4: **Raw performance/representation scaling of simplicial sequence models.** *Top:* Per-step held-out test cross-entropy minus per-seed irreducible cross-entropy. Mean (black) of 16 seeds (blue), no smoothing. Vertical axis uses a symmetric log scale to display negative entries below the singularity. *Bottom:* Mean squared error predicting Bayesian belief distributions from final-layer activations. Mean (black) of 16 seeds (red), no smoothing.