We limited ourselves to the simpler stationary ergodic setting in this paper to empirically verify important basic features of transformer representations. However, the framework presented in this work will naturally extend to non-stationary [29] and non-ergodic processes [6], which is important for the extension to real-world tasks. For example, in in-context learning the underlying process is inherently nonergodic. In the more general setting, we also expect transformers to represent belief-state geometry, although that geometry will then reflect the more general types of non-stationary and non-ergodic processes representative of real-world data like natural language [7, 8].

In the experiments we ran, we generated training data using HMMs. A natural question is what to expect with training data generated from non-HMM sources. It is important to note that any dataset made of sequences of tokens used to train a transformer can be represented as being generated from an HMM, if one allows for non-ergodicity, non-stationarity, and infinite states. For example, in the case of training data that takes the form of parenthesis matched strings, one would usually conceive of this as being generated by a push-down automaton. However, this can equivalentaly be represented by an HMM consisting of an infinite chain of hidden states. Another interesting point that we did not focus on in our presentation for the sake of simplicity, is that even if one does not assume an HMM for the generating machine, the question of the computational structure of an optimal predictor is naturally an HMM, with hidden states given by distinct distributions over the future (whether the generator is an HMM or a Turing Machine, or something else) [33].

In this work we verified that belief states are linearly represented in the residual stream, even in cases where the belief states form intricate fractal geometries. Although this is quite suggestive that these representations are used to implement some form of Bayesian updating, we did not directly test this. In other words, we do not yet know if these LLMs develop something like a circuit for implementing Bayesian updates, as would be natural for MSP dynamics. So far, we have only established the representation of belief states and their geometry.

Computational mechanics is primarily (but not completely, see Ref. [21]) concerned with optimal prediction, but LLMs in practice are not perfectly optimal. Further work is needed to understand how near-optimality and non-optimality influence belief state representations. Progress in this direction could offer insights into evolving structures during training.

Moving beyond stationary ergodic HMM training data, we still generically expect fractal-like structure in the residual-stream activations. Why? The fractal structure is a result of Bayesian updating applied to the “non-deterministic” computational structure of HMMs [16]. With stationary processes, this same Bayesian map is folded into itself time and again, producing a very clean fractal. For natural language, even in the absence of stationarity, there will nevertheless be resonances of non-deterministic computational structure, which would imply a type of folding beliefs in a self-similar manner. However, a more rigorous version of these statements and the empirical validation is left for future work.

5 Conclusion

In this work, we introduced a theoretical framework that establishes a clear connection between the structure of training data and the geometric properties of activations in trained transformer neural networks. Through experiments, we validated that the geometry of belief states is linearly represented within the residual stream of the transformer architecture. This finding suggests that transformers construct predictive representations that surpass simple next-token prediction. Instead, these representations encode the complex geometry associated with belief updating over hidden states, capturing an inference process over the underlying structure of the data-generating process.

Our work takes a significant step towards concretizing the understanding of the computational structures that drive the behavior of large language models (LLMs) and how these structures relate to the data on which they are trained. This concretization is crucial, as the rapid advancements in LLMs have led to models with increasingly sophisticated behavioral capabilities. However, without a clear understanding of the underlying computational structures and their relationship to the training data, it becomes challenging to interpret, trust, and further improve these models.

References

[1] Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. *arXiv preprint arXiv:2403.06963*, 2024.

[2] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[3] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[4] J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8(January):17–24, 2012. doi: 10.1038/NPHYS2190.

[5] J. P. Crutchfield, C. J. Ellison, J. R. Mahoney, and R. G. James. Synchronization and control in intrinsic and designed computation: An information-theoretic analysis of competing models of stochastic computation. *CHAOS*, 20(3):037105, 2010. Santa Fe Institute Working Paper 10-08-015; arxiv.org:1007.5354 [cond-mat.stat-mech].

[6] James P Crutchfield and Sarah Marzen. Signatures of infinity: Nonergodicity and resource scaling in prediction, complexity, and learning. *Physical Review E*, 91(5):050106, 2015.

[7] Lukasz Debowski. *Information theory meets power laws: Stochastic processes and language models*. John Wiley & Sons, 2020.

[8] Łukasz Dębowski. A simplistic model of neural scaling laws: Multiperiodic santa fe processes. *arXiv preprint arXiv:2302.09049*, 2023.

[9] Joshua Engels, Isaac Liao, Eric J Michaud, Wes Gurnee, and Max Tegmark. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*, 2024.

[10] Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*, 2024.

[11] Rick Goldstein. Handcrafting a network to predict next token probabilities for the random-random-xor process. https://apartresearch.com/event/compmech, June 2024. Research submission to the Computational Mechanics Hackathon! research sprint hosted by Apart, PIBBSS, and Simplex.

[12] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[13] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[14] Jiachen Hu, Qinghua Liu, and Chi Jin. On limitation of transformer for learning hmms. *arXiv preprint arXiv:2406.04089*, 2024.

[15] A. M. Jurgens and J. P. Crutchfield. Shannon entropy rate of hidden Markov processes. *Journal of Statistical Physics*, 183(2):32, 2021.

[16] Alexandra M Jurgens and James P Crutchfield. Divergent predictive states: The statistical complexity dimension of stationary, ergodic hidden markov processes. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(8), 2021.

[17] 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. *arXiv preprint arXiv:2210.13382*, 2022.

[18] Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022.

[19] Wolfgang Löhr. Predictive models and generative complexity. *Journal of Systems Science and Complexity*, 25:30–45, 2012.

[20] S. E. Marzen and J. P. Crutchfield. Nearly maximally predictive features and their dimensions. *Phys. Rev. E*, 95(5):051301(R), 2017. doi: 10.1103/PhysRevE.95.051301. SFI Working Paper 17-02-007; arxiv.org:1702.08565 [cond-mat.stat-mech].

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

[22] Neel Nanda and Joseph Bloom. Transformerlens. `https://github.com/TransformerLensOrg/TransformerLens`, 2022.

[23] Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

[24] Pedro A Ortega, Jane X Wang, Mark Rowland, Tim Genewein, Zeb Kurth-Nelson, Razvan Pascanu, Nicolas Heess, Joel Veness, Alex Pritzel, Pablo Sprechmann, et al. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*, 2019.

[25] Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. *arXiv preprint arXiv:2311.04897*, 2023.

[26] Keenan Pepper. RNNs represent belief state geometry in their hidden states. https://apartresearch.com/event/compmech, June 2024. Research submission to the Computational Mechanics Hackathon! research sprint hosted by Apart, PIBBSS, and Simplex.

[27] P. M. Riechers and J. P. Crutchfield. Spectral simplicity of apparent complexity, Part I: The nondiagonalizable metadynamics of prediction. *Chaos*, 28:033115, 2018. doi: 10.1063/1.4985199.

[28] P. M. Riechers and J. P. Crutchfield. Spectral simplicity of apparent complexity, Part II: Exact complexities and complexity spectra. *Chaos*, 28:033116, 2018. doi: 10.1063/1.4986248.

[29] Paul M Riechers and James P Crutchfield. Fraudulent white noise: Flat power spectra belie arbitrarily complex processes. *Physical Review Research*, 3(1):013170, 2021.

[30] Joshua B Ruebeck, Ryan G James, John R Mahoney, and James P Crutchfield. Prediction and generation of binary markov processes: Can a finite-state fox catch a markov mouse? *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(1), 2018.

[31] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001.

[32] Ilya Sutskever. Building AGI, alignment, future models, spies, Microsoft, Taiwan, & enlightenment. Podcast on Dwarkesh Patel Podcast, March 2023. URL `https://www.youtube.com/watch?v=YEUclZdj_Sc`. Accessed: 2024-05-22.

[33] D. R. Upper. *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*. PhD thesis, University of California, Berkeley, 1997. Published by University Microfilms Intl, Ann Arbor, Michigan.