formance. When we ablate eigenvectors to retain only the top-k principal components, SciQ accuracy degrades dramatically (Table 2). For instance, retaining only the top 10 eigen-directions reduces Pythia-1B's accuracy from 0.838 to 0.225, while OLMo-2-7B drops from 0.970 to 0.155. Interestingly, removing the top eigen-directions has minimal impact, suggesting that information is distributed across the full spectrum rather than concentrated solely in dominant directions. This finding validates our use of full-spectrum metrics like RankMe and α_{ReQ} rather than top-k proxies, and underscores that effective language understanding requires the entire representational manifold—not just its principal components. The necessity of preserving full spectral information aligns with the "compression-seeking" phase's anisotropic consolidation, which selectively strengthens certain directions while maintaining distributed representations across the manifold.

Geometry of Post-Training: Alignment vs Exploration. Different post-training recipes induce distinct shifts in LLM representation geometry, explaining the model's behavioral changes. Supervised Fine-Tuning (SFT) drives an "entropy-seeking" dynamic, expanding the representational manifold for specific instruction-response examples. This manifold expansion can be seen as evidence for the lazy-regime learning described by Ren and Sutherland (2024) during SFT, and points to a near-diagonal empirical NTK that results in an instance-level learning dynamics. Consequently, this dynamic improves in-distribution performance but risks overfitting due to higher representational capacity. In contrast, Reinforcement Learning from Verifiable Rewards (RLVR) promotes a "compression-seeking" dynamic, refining representations towards reward-aligned directions. This geometric compression may explain how RLVR amplifies and refines existing capabilities, as observed by Zhao et al. (2025), potentially by constraining representations to a more structured subspace while reducing its exploration ability, as shown by Yue et al. (2025). In summary, SFT/DPO-induced rank expansion may foster preference memorization and exploratory behavior, while RLVR-induced consolidation amplifies model-capabilties towards reward-oriented, less diverse generation (c.f. Figure 6C).

Limitations and Future Work Tracing a model's geometry, whether "entropy-seeking" or "compression-seeking" , could inform more effective interventions for LLM development and evaluation, such as the selection of optimal pretraining checkpoints for targeted fine-tuning or designing training strategies that deliberately navigate these geometric phases. Our findings have several limitations: (i) computational constraints limited our analysis to models up to 12B parameters, though the phases persist across scales from 160M to 12B; (ii) spectral metric computation requires $\sim$10K samples and scales quadratically with hidden dimension (iii) our theoretical analysis assumes simplified linear feature extractors, leaving the extension to full transformer architectures as future work; (iv) we focused on English-language models trained with standard objectives, and whether similar phases emerge in multilingual or alternatively-trained models remains unexplored. Furthermore, our findings are primarily correlational; establishing causal connections between geometric dynamics and emergent capabilities requires additional investigation.

6. Conclusion

We show that LLMs undergo non-monotonic representation geometry changes, often masked by steadily decreasing training loss. By employing spectral metrics of feature covariates (RankMe and α_{ReQ}), we delineate three distinct pretraining phases: "warmup" , "entropy-seeking" (correlating with n-gram memorization), and "compression-seeking" (correlating with long-context generalization). We further demonstrate that post-training recipes induce specific geometric changes: SFT/DPO exhibit "entropy-seeking" dynamics, whereas RLVR exhibit "compression-seeking" dynamics. These results provide a quantitative framework for guiding future advancements in LLM development.

Impact Statement The goal of our work is to advance the understanding of internal representations of LLMs. Although there are potential downstream societal consequences of this technology, we feel there are no direct consequences that must be specifically highlighted here.

Acknowledgments

The authors would like to thank Koustuv Sinha for insightful discussions that helped shape the scope of the project and Jacob Mitchell Springer for helping setup the OLMo-2 supervised finetuning pipeline. We are grateful to the OLMo team, particularly Nathan Lambert, Dirk Groeneveld, and Bailey Kuehl, for providing access to the OLMo-2 checkpoints (especially OLMo-2-1B) that enabled this research. The authors are also grateful to Daniel Levenstein, Johannes von Oswald, Jonathan Cornford, Mandana Samiei, Tejas Vaidhya, and Zahraa Chorghay for their comments and feedback. A.G. was supported by Vanier Canada Graduate Scholarship. G.L. was supported by NSERC (Discovery Grant RGPIN2018-04821), the Canada Research Chair in Neural Computations and Interfacing, CIFAR (Canada AI Chair), as well as IVADO and the Canada First Research Excellence Fund. B.A.R. was supported by NSERC (Discovery Grant: RGPIN-2020-05105; Discovery Accelerator Supplement: RGPAS-2020-00031) and CIFAR (Canada AI Chair; Learning in Machines and Brains Fellowship). The authors also acknowledge the material support of NVIDIA in the form of computational resources, as well as the compute resources, software and technical help provided by Mila (mila.quebec).

References

Kumar K Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-req: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. *Advances in Neural Information Processing Systems*, 35:17626–17638, 2022.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Nora Belrose, Quintin Pope, Lucia Quirke, Alex Mallen, and Xiaoli Fern. Neural networks learn statistics of increasing complexity. *arXiv preprint arXiv:2402.04362*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

Davis Brown, Charles Godfrey, Nicholas Konz, Jonathan Tu, and Henry Kvinge. Understanding the inner-workings of language models through representation dissimilarity. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

Branton DeMoss, Silvia Sapora, Jakob Foerster, Nick Hawes, and Ingmar Posner. The complexity dynamics of grokking. *arXiv preprint arXiv:2412.09810*, 2024.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback, 2023.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, 2022.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning*, pages 10929–10974. PMLR, 2023.

Arna Ghosh, Arnab Kumar Mondal, Kumar Krishna Agrawal, and Blake Richards. Investigating power laws in deep representation learning. *arXiv preprint arXiv:2202.05808*, 2022.