Extended Abstract Track

Multi-task Learning yields Disentangled World Models: Impact and Implications

Editors: List of editors' names

Abstract

Intelligent perception and interaction with the world hinges on internal representations that capture its underlying geometry ("disentangled" or "abstract" representations). The ability to form these disentangled representations from high-dimensional, noisy observations is a hallmark of intelligence, observed in both biological and artificial systems. In this opinion paper we highlight unpublished experimental and theoretical results guaranteeing the emergence of disentangled representations in agents that optimally solve multi-task evidence aggregation classification tasks, canonical in the cognitive neuroscience literature. The key conceptual finding is that, by producing accurate multi-task classification estimates, a system implicitly represents a set of coordinates specifying a disentangled, topology-preserving representation of the underlying latent space. Since the theory relies only on the system accurately computing the classification probabilities, we are able to derive a closed-form solution for extracting disentangled representations from any multi-task classification system. The theory provides conditions for the emergence of these representations in terms of noise, number of tasks, and evidence aggregation time, and we experimentally validate the theoretical predictions on RNNs and GPT-2 style models solving such canonical evidence-aggregation decision-making neuroscience tasks. It turns out that GPT models are much better at disentangling factors, showcasing their unique ability in grokking world models. We here discuss implications of these findings, for machine learning and neuroscience alike.

Keywords: zero-shot generalization, disentanglement, representation learning, multi-task learning, interpretability, computational neuroscience, world models, continuous attractors

1. Introduction

Humans and animals can generalize to new settings effortlessly, leveraging a combination of past experiences and world models (Lake et al., 2015, 2016). Modern foundation models also display emergent out-of-distribution (OOD) generalization abilities, in the form of zero- or few-shot learning (Brown et al., 2020; Pham et al., 2021; Oquab et al., 2023).

One mechanism for generalization is through abstract, or *disentangled*, representations (Higgins et al., 2017; Kim and Mnih, 2018; Johnston and Fusi, 2023). These two concepts are interrelated yet somewhat distinct (see Appendix A and Ostojic and Fusi (2024)). An abstract representation of $x_1, \ldots, x_n$ represents each x_i linearly and approximately mutually orthogonally. Disentangled representations encode each x_i orthogonally, without the necessity of linearity. When a representation is abstract, a linear decoder (i.e. downstream neuron) trained to discriminate between two categories can readily generalize to stimuli not observed in training, due to the structure of the representation. Furthermore, the more disentangled the representation is, the lower the interference from other variables and hence the better the performance. This corresponds to decomposing a novel stimulus into its familiar attributes, and reasoning based on those. For instance, imagine you are at a grocery store, deciding whether a fruit is ripe or not. If the brain's internal representation of food

© 2023 .

attributes (ripeness, caloric content, etc.) is disentangled, then learning to perform this task for bananas would lead to zero-shot generalization to other fruit (e.g. mangos, Figure 1a).

Several brain areas including the amygdala, prefrontal cortex and hippocampus have been found to encode variables of interest in an abstract format (Saez et al., 2015; Bernardi et al., 2020; Boyle et al., 2022; Nogueira et al., 2023; Courellis et al., 2024). This raises the question of under which conditions do such representations emerge in biological and artificial agents alike. Here we argue that multi-task learning is crucial to get the kind of topology-preserving representations that yield generalization in biological systems, and that a parallel processing view of the brain, in line with the cortical architecture, is naturally conducive to that framework. To do so, we first summarize findings from recent unpublished work (anonymized link included in section 6, "Resources") that demonstrates the mathematical conditions for disentanglement in autoregressive architectures (including RNNs, LSTMs and GPT-2 style models) that can deal with noisy sequential real-world data, and then discuss the implications of the work for machine learning and neuroscience alike.

2. Problem formulation

In recent unpublished work (referred to as "highlighted work" and included with the submission) we consider canonical cognitive neuroscience tasks that involve evidence aggregation over time, mirroring decision-making under uncertainty. Tasks have a trial structure. In each trial, a ground truth vector $\mathbf{x}^* \in \mathbb{R}^D$ ($x_i^* \sim \mathit{Uniform}(-0.5, 0.5)$) is sampled (Figure 1b). Each element x_i^* of $\mathbf{x}^*$ corresponds to different options a decision-maker might have, or to different attributes of the same item. The target output for the trial $\mathbf{y}(\mathbf{x}^*) \in \{-1, +1\}^{N_{\text{task}}}$ is a vector of N_{task} +1s and -1s, depending on whether $\mathbf{x}^*$ is above or below each of N_{task} classification boundaries (Figure 1b). The boundaries are fixed, and reflect criteria based on which decisions will be made. Imagine for example that x_1 corresponds to food and x_2 to water reward. Depending on the agent's internal state, one could take precedence over the other, and the degree of preference is reflected in the slope of the line.

We train RNNs and GPT-2 style models to output the target labels $\mathbf{y}(\mathbf{x}^*)$ (Figure 1c). The models do not have access to the ground truth $\mathbf{x}^*$ but rather a noised-up, non-linearly transformed version of it. Specifically, the input is $\mathbf{X}(t) \in \mathbb{R}^D$ where $\mathbf{X}(t) = \mathbf{x}^* + \sigma\mathcal{N}(\mathbf{0}, \mathbf{I}_D)$, σ being the input noise standard deviation. The model should integrate noisy samples $\mathbf{X}(t)$ over time, viewed through an static, injective observation map (encoder) f, to estimate $\hat{\mathbf{Y}}_i(t) = \Pr\{\mathbf{y}_i(\mathbf{x}^*) = 1 | f(\mathbf{X}(1)), \ldots, f(\mathbf{X}(t))\}$.

3. Contributions

We here summarize the contributions of the hightlighted work:

- We prove that any optimal multi-task classifier is guaranteed to learn an abstract representation of the ground truth contained in the noisy measurements in its latent state, if the classification boundary normal vectors span the input space. Furthermore, the representations are guaranteed to be disentangled if $N_{\text{task}} \gg D$. Intriguingly, noise in the observations is necessary to guarantee the latent state would compute an optimal, disentangled representation of the ground truth (for detailed proofs, see Appendix B of unpublished paper).

Extended Abstract Track

- We confirm that RNNs trained to multitask develop abstract representations that generalize OOD as quantified by regression generalization (Johnston and Fusi, 2023) when $N_{\text{task}} \geq D$ (Figure 1f), and that the latent factors are approximately orthogonal when $N_{\text{task}} \gg D$ (disentanglement, Figure 1g). The computational substrate of these representations are continuous attractors (Amari, 1977) storing a ground truth estimate in a product space of the latent factors (Figure 1d). Furthermore, these representations preserve the topology of the real world in their structure, where a latent factor (e.g. x_1, x_2) corresponds to a direction in PC space of RNN hidden layer activity, and nearby trials get mapped to nearby trajectories in PC space (Figure 1d).
- We show that the setting is robust to a number of manipulations, including interleaved learning of linear and non-linear tasks and free reaction time decisions.
- We reproduce these findings in GPT-2 style models, which generalize better due to them learning orthogonal representations for lower N_{task}, confirming their appropriateness for constructing world models.
- Finally, we demonstrate the strong advantage of multi-task learning, which scales linearly with D and leads to representations that can be used for any task that involves the same latent variables, over previously proposed mechanisms of representation learning in the brain ("context-dependent computation") (Mante et al., 2013; Yang et al., 2019), which scale linearly with N_{task} and exponentially with D.

Despite being framed in the context of canonical decision-making neuroscience tasks, these results are general; they apply to any system aggregating noisy evidence over time.

4. Implications for representation learning

4.1. Topology-preserving representation learning

These results have implications for the learning of representations that inherit the topological structure of the world. It suggests that this naturally happens, as long as there are enough tasks to uniquely identify the location of the ground truth $\mathbf{x}^*$ when solving these classifications (see Appendix B of the unpublished paper). Crucially, the constraints from different tasks need to be placed simultaneously on the representation, which explains why representations emerging from context-dependent computation (Mante et al., 2013) are typically not disentangled. A prime example of such multiple imposed constraints from neurobiology is the fly head-direction system (Vafidis et al., 2022; Wilson, 2023), where a ring-like topology-preserving representation of head direction might be enforced exactly because of it's functional role in driving many downstream circuits for navigation.

4.2. Consistency across individuals

Potentially even more far reaching, this work implies guarantees about representational alignment across individuals or neural networks. It suggests that as long as we are faced and solve similar problems in the day-to-day world, we are bound to arrive at similar, disentangled representations of latent factors governing these decisions. This is reminiscent of the Platonic representation hypothesis (Huh et al., 2024), which suggests that the convergence in deep neural network representations is driven by a shared statistical model of

reality, like Plato's concept of an ideal reality. This could explain why for example modern LLMs come to encode high-level, human-interpretable concepts (Templeton et al., 2024).

4.3. Interplay between number of tasks and fine-grainness of representations

Intriguingly, this work reveals a fundamental interplay between richness of tasks performed and complexity/detail of the representation learned. If only a small number of tasks are performed, the resulting representations will be fundamentally limited to lie within the space spanned by these tasks. However, as more tasks are added, finer details could be discerned. Therefore, the theorem and experimental results provided are not a one-way-street from dimensionality D of the latent factors to how many tasks N_{task} are required to uncover such latents. Rather, in a complicated and high-dimensional world, the richness of the tasks at hand directly affects the dimensionality D of the latents that can be extracted, allowing for "ground truths" $\mathbf{x}^*$ at different levels of granularity to be explored. The richer the label information available, the more fine-grained the resulting world model will be.

5. Connections to neuroscience and machine learning

5.1. Correspondence to brain processes

The brain encodes variables of interest in a disentangled format, in processes as disparate as memory (Boyle et al., 2022), emotion (Saez et al., 2015), and decision making (Bongioanni et al., 2021). Furthermore, performance in tasks has been shown to degrade once said neural representations collapse (Saez et al., 2015), supporting the role of abstract representations in guiding generalizable behavior. Given that the cortical architecture is uniquely suited for parallel processing (Hawkins et al., 2019), the cortex is a prime candidate area for the construction of disentangled world models. Another such area is the thalamus; it is posited that thalamocortical loops operate in parallel, and combined with internal state-dependent mechanisms lead to state-dependent action selection (e.g. prioritizing water when thirsty), while evidence integration occurs in corticostriatal circuits (Rubin et al., 2020).

5.2. Relation to machine learning paradigms

The experiments in the highlighted work are inspired by canonical cognitive neuroscience tasks. Yet, the conclusions drawn are informative concerning the fundamental nature of generalization. For instance, why do foundation models generalize well in various domains? We suggest that parallel processing forces learning of generalizable world models, and our setting directly applies to settings where neural networks predict a rich representation of the world from partial observations. Some examples are predictive coding where high-dimensional next states have to be predicted (Gornet and Thomson, 2024), and self-supervised learning, where multiple missing image patches have to filled up at once (Dosovitskiy et al., 2020). In particular, our findings may explain why GPT-style architectures have been so successful at grokking world models, with impressive zero- and few-shot learning capabilities.

Finally, an alternative to multi-task learning that we explore is slow interleaved learning. This allows the weights of a neural network to be effectively conditioned to solve all the tasks simultaneously. The relation between multi-task and interleaved learning is a promising topic for future research.

6. Resources

We here include an anonymized link to the full unpublished paper, for reviewer convenience: https://drive.google.com/file/d/1Km-GjeuiKWbhktsEx8gxpAd7F4JnLQ_Y/view?usp=share_link

Acknowledgments

Acknowledgements will be included in the final version to preserve anonymity.

References

S Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2):77–87, 1977. doi: 10.1007/bf00337259. URL https://doi.org/10.1007/bf00337259.

Silvia Bernardi, Marcus K. Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C. Daniel Salzman. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4):954–967.e21, November 2020. doi: 10.1016/j.cell.2020.09.031. URL https://doi.org/10.1016/j.cell.2020.09.031.

A. Bongioanni, D. Folloni, L. Verhagen, J. Sallet, M. C. Klein-Flügge, and M. F. S. Rushworth. Activation and disruption of a neural mechanism for novel choice in monkeys. *Nature*, 591(7849):270–274, January 2021. doi: 10.1038/s41586-020-03115-5. URL https://doi.org/10.1038/s41586-020-03115-5.

Lara M. Boyle, Lorenzo Posani, Sarah Irfan, Steven A. Siegelbaum, and Stefano Fusi. Tuned geometries of hippocampal representations meet the demands of social memory, January 2022. URL https://doi.org/10.1101/2022.01.24.477361.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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, 2020. URL https://arxiv.org/abs/2005.14165.

Hristos S. Courellis, Juri Minxha, Araceli R. Cardenas, Daniel L. Kimmel, Chrystal M. Reed, Taufik A. Valiante, C. Daniel Salzman, Adam N. Mamelak, Stefano Fusi, and Ueli Rutishauser. Abstract representations emerge in human hippocampal neurons during inference. *Nature*, 632(8026):841–849, August 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07799-x. URL http://dx.doi.org/10.1038/s41586-024-07799-x.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL https://arxiv.org/abs/2010.11929.

James Gornet and Matt Thomson. Automated construction of cognitive maps with visual predictive coding. *Nature Machine Intelligence*, 6(7):820–833, July 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00863-1. URL http://dx.doi.org/10.1038/s42256-024-00863-1.

Jeff Hawkins, Marcus Lewis, Mirko Klukas, Scott Purdy, and Subutai Ahmad. A framework for intelligence and cortical function based on grid cells in the neocortex. *Frontiers in Neural Circuits*, 12, January 2019. ISSN 1662-5110. doi: 10.3389/fncir.2018.00121. URL http://dx.doi.org/10.3389/fncir.2018.00121.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations, 2018. URL https://arxiv.org/abs/1812.02230.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

W. Jeffrey Johnston and Stefano Fusi. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nature Communications*, 14(1), February 2023. doi: 10.1038/s41467-023-36583-0. URL https://doi.org/10.1038/s41467-023-36583-0.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, December 2015. doi: 10.1126/science.aab3050. URL https://doi.org/10.1126/science.aab3050.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, November 2016. doi: 10.1017/s0140525x16001837. URL https://doi.org/10.1017/s0140525x16001837.

V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, November 2013. doi: 10.1038/nature12742. URL https://doi.org/10.1038/nature12742.

Ramon Nogueira, Chris C. Rodgers, Randy M. Bruno, and Stefano Fusi. The geometry of cortical representations of touch in rodents. *Nature Neuroscience*, 26(2):239–250, January 2023. doi: 10.1038/s41593-022-01237-9. URL https://doi.org/10.1038/s41593-022-01237-9.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. URL https://arxiv.org/abs/2304.07193.

Srdjan Ostojic and Stefano Fusi. Computational role of structure in neural activity and connectivity. *Trends in Cognitive Sciences*, 28(7):677–690, July 2024. ISSN 1364-6613. doi: 10.1016/j.tics.2024.03.003. URL http://dx.doi.org/10.1016/j.tics.2024.03.003.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning, 2021. URL https://arxiv.org/abs/2111.10050.

Jonathan E. Rubin, Catalina Vich, Matthew Clapp, Kendra Noneman, and Timothy Verstynen. The credit assignment problem in cortico-basal ganglia-thalamic networks: A review, a problem and a possible solution. *European Journal of Neuroscience*, 53 (7):2234–2253, May 2020. ISSN 1460-9568. doi: 10.1111/ejn.14745. URL http://dx.doi.org/10.1111/ejn.14745.

A. Saez, M. Rigotti, S. Ostojic, S. Fusi, and C.D. Salzman. Abstract context representations in primate amygdala and prefrontal cortex. *Neuron*, 87(4):869–881, August 2015. doi: 10.1016/j.neuron.2015.07.024. URL https://doi.org/10.1016/j.neuron.2015.07.024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Pantelis Vafidis, David Owald, Tiziano D'Albis, and Richard Kempter. Learning accurate path integration in ring attractor models of the head direction system. *eLife*, 11:e69841, jun 2022. ISSN 2050-084X. doi: 10.7554/eLife.69841. URL https://doi.org/10.7554/eLife.69841.

Rachel I. Wilson. Neural networks for navigation: From connections to computations. *Annual Review of Neuroscience*, 46(1):403–423, July 2023. ISSN 1545-4126. doi: 10.1146/annurev-neuro-110920-032645. URL http://dx.doi.org/10.1146/annurev-neuro-110920-032645.

Guangyu Robert Yang, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many

cognitive tasks. *Nature Neuroscience*, 22(2):297–306, Feb 2019. ISSN 1546-1726. doi: 10.1038/s41593-018-0310-2. URL https://doi.org/10.1038/s41593-018-0310-2.

Appendix A. Definitions

In the main text we used the terms "abstraction" and "disentanglement". Since there exists some ambivalence about their meaning in the literature, we would like to strictly define them here. We will be using definitions adapted from Ostojic and Fusi (2024):

- Abstract representation of $x_1, \ldots, x_n$ represents each x_i linearly and approximately mutually orthogonally. Thus, abstractness ensures a simple linear map can decode each x_i regardless of variation in $x_{j \neq i}$.
- Disentangled representations of $x_1, \ldots, x_n$ encode each x_i orthogonally, without the necessity of linearity.

Note that under this definition, axis-alignment is not a requirement for disentanglement (also see Higgins et al. (2018)). The highlighted work suggests that the computer science and computational neuroscience communities should adopt this broader definition of disentanglement, because otherwise we might be missing cases where the factors are not axis-aligned, but they are still orthogonal and can still be isolated by a linear decoder. Our argument is that there is nothing special about individual factors being encoded by individual neurons. Rather, we think that allowing for mixed representations within the definition of disentanglement leads to a more holistic view of disentanglement. A contribution of the highlighted work, along with others (Johnston and Fusi, 2023), is to bring this argument to the forefront.

Appendix B. Supplementary figures

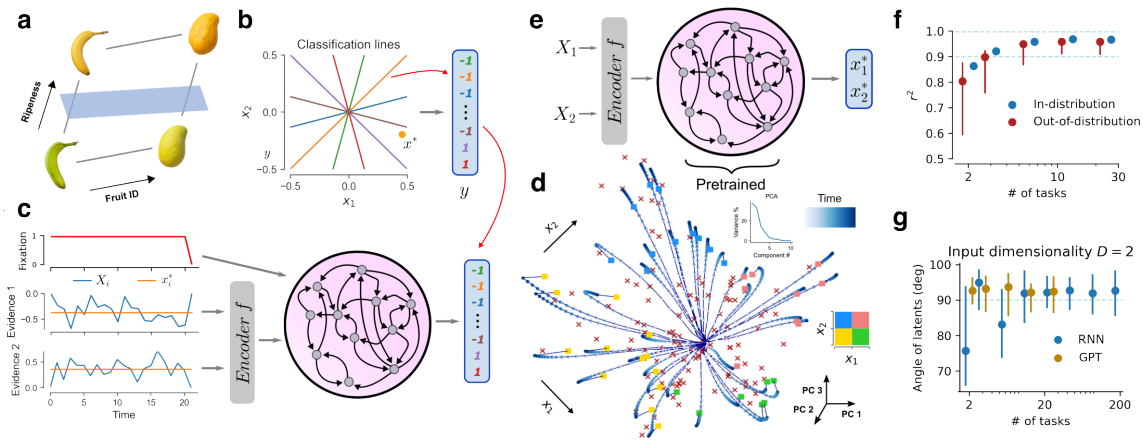

Figure 1: **Learning disentangled representations.** **(a)** A disentangled representation directly lends itself to OOD generalization: a downstream linear decoder that can differentiate ripe from unripe bananas can readily generalize to mangos, even though it has never been trained on mangos. **(b)** The task is to simultaneously report whether the ground truth $\mathbf{x}^*$ lies above $(+1)$ or below (-1) a number of classification lines. **(c)** RNNs are trained to report the outcome of all the binary classifications in **b** at the end of the trial (indicated by the fixation input turning 0). **(d)** Top 3 PCs of RNN activity. Each line is a trial, while color saturation indicates time. All trials start from the center and move outwards, towards the location of $\mathbf{x}^*$ in state space. The last timepoint in each trial (squares) is colored according to the quadrant this trial was drawn from. Red x's correspond to attractors. Input noise here is removed so that trajectories can be visualized easier. The network learns a two-dimensional continuous attractor that seems to provide a disentangled representation of the state space. **(e)** To evaluate OOD generalization, a linear decoder (see **a**) is trained to output the ground truth $\mathbf{x}^*$ at the end of the trial, while keeping network weights frozen. The decoder is trained in 3 out of 4 quadrants and tested OOD in the 4th quadrant. **(f)** ID and OOD generalization performance for networks trained in different number of tasks N_{task}. The 25, 50 and 75 percentiles of r^2 for each network size are reported. ID and OOD performance increase with number of tasks, and the generalization gap decreases, indicating that the networks have indeed learned abstract representations. **(g)** Angles between latent factor decoders. The angles approach 90 degrees as $N_{\text{task}} \gg D$ for RNNs, but they are already close to 90 degrees for $N_{\text{task}} \geq D$ for GPT-style models. The errors that remain for $N_{\text{task}} \geq 24$ for RNNs and for $N_{\text{task}} \geq 2$ for GPT can be attributed to variability in the linear decoder fits. Therefore, we conclude that the representations become disentangled (see Appendix A) for both models. RNNs are disentangled as $N_{\text{task}} \gg D$ as our theory predicts, but GPT style models disentangle as long as $N_{\text{task}} \geq D$, showcasing their unique ability in grokking world models.