Beyond Behavioural Evaluations for Assessing World Models

Anonymous Authors[1]

Abstract

To predict the future capabilities of agentic systems, it's useful to understand the extent to which foundation models have internal world models. Agents with robust internal world models generalise better to unseen and out-of-distribution data. Interpretability evaluations suggest that transformers trained on tasks like Othello have robust world models. However, behavioural evaluations question whether these agents truly have world models as robust as the interpretability research indicates. We argue that to claim that an ML model doesn't have a robust world model, practitioners must either use Interpretability evaluations or provide an argument for why their behavioural evaluations are fully elicitating the capabilities of the model. We hence propose a protocol for combining Evaluations and Elicitation to assess the world models of frontier AI systems.

1. Introduction

Foundation models are difficult to evaluate for the same reason that they are so useful: they perform a wide variety of tasks requiring different latent knowledge and we are not always sure how to analyse what knowledge they possess and how they leverage such knowledge (Wei et al., 2022a; Lubana et al., 2024; Brown et al., 2020; Bubeck et al., 2023).

A core question for understanding current AI systems and how transformative future systems may be is whether AI agents have an internal world model (Li et al., 2023; Liu et al., 2024) or whether they merely use shallow heuristics (Hao et al., 2023; Bender et al., 2021). We may say that an AI system has a **world model** if there is a correspondence between the model's representations and the external environment (Ha & Schmidhuber, 2018). Recent work in Interpretability has suggested that Foundation Models are so effective because they learn an emergent implicit world model containing representations that can be used for cognition and planning (Li et al., 2023; Lindsey et al., 2025; Karvonen, 2024).[1] However, behavioural analysis of Foundation Models has suggested that agents may have much less sophisticated world models than the Interpretability research suggests (Vafa et al., 2024).

In this paper, we analyse how to evaluate world models in Foundation Models, comparing the **Behavioural Evaluations**, **Elicited Evaluations** and **Interpretability Evaluations** of world models. We hence show that since there is a distinction between a model *possessing* and *using* a world model, it is not generally valid to conclude that a model does not have a world model, or that it's world model is fragile or low fidelity or inaccurate, based on Behavioural Evaluations alone. We hence argue that using Elicited Evaluations or Interpretability Evaluations can provide more robust evidence for the presence of a high fidelity world model.

2. Implicit World Models

A world model (Fodor, 1987) is a mental construct that represents the dynamics of an external environment internal to the mind of an agent. Model-based RL systems explicitly use a model of the environment to plan and make decisions and this model is separated from the policy (Hafner et al., 2023; 2020; Sutton et al., 1998; Parr et al., 2022). Recent work has suggested that through autoregressive token prediction Foundation (Language) Models can learn an emergent world model which internally represents important features and dynamics of the AI system's environment (Li et al., 2023; Nanda et al., 2023b; Mikolov et al., 2013; Gurnee & Tegmark, 2023; Engels et al., 2024).[2] The finding of emergent world models in Foundation Models would explain why language models are able to complete tasks that seem beyond the shallow statistics of the training data (Wei et al., 2022a; Kiciman et al., 2023; Delétang et al., 2023). The explanation for these abilities would be that autoregres-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML 2025 Workshop on Assessing World Models. Do not distribute.

[1] See also Park et al. (2024); Nanda et al. (2023b); Olah et al. (2020); Jenner et al. (2024); Taufeeque et al. (2024)

[2] We do not require that the world model be linearly represented within the AI system for our purposes. In practise, (Nanda et al., 2023b; Mikolov et al., 2013; Bricken et al., 2023; Cunningham et al., 2023; Gao et al., 2024) suggest that often world models are linearly represented within AI systems which is computationally and mathematically convenient.

sive AI systems learn implicit world models: high-fidelity, compressed representations of the dynamics of the generating process of their training data distribution which can be leveraged for learning, adapting and acting.

The Causal World Model Theorem from Richens & Everitt (2024) states that "any agent capable of adapting to a sufficiently large set of distributional shifts must have learned a causal [world] model of the data generating process". Intuitively we might expect that agents which have learned a world model are likely to generalise better out of distribution compared to models that are using brittle heuristics.

World Models are internal representations within an agent that correspond to some features of the external environment. When assessing how effective a world model W is with respect to a task T, we are interested in three properties of the world model:

1. **Task Relevance**: A world model is task relevant if it contains the relevant information that might be useful for the task. For example if the task T is exploring New York City by car, then world models about Chicago or about the New York transit system are not relevant.

2. **Accuracy**: A world model is accurate if it accurately represents the true environment. Even though the world model may be incomplete, the information that it does present should be approximately error-free.

3. **Fidelity**: A world model is high-fidelity if it is of high resolution and captures sufficient detail about the environment.

We say that a world model is **effective** if it is *task-relevant*, *accurate* and *high-fidelity* with respect to the task T.

3. Evaluating World Models

There are two main existing approaches to evaluating world models in the literature: *Behavioural Evaluations* (Vafa et al., 2024) and *Interpretability Evaluations* (Nanda et al., 2023b). We also discuss a proposed third approach a protocol for *Elicited Evaluations* which also include a Capability Elicitation stage and an argument for why we should expect that the elicitation was sufficient to conclude that a model doesn't have a certain capability with reasonable likelihood.

3.1. Interpretability Evaluations

A natural approach to assessing the implicit world models taken by Li et al. (2023); Nanda et al. (2023b); Gurnee & Tegmark (2023) is to use supervised (linear) probes to assess whether the model's representations can recover some state from the external environment (Hewitt & Liang, 2019; Wu et al., 2025; Feng et al., 2024; Abdou et al., 2021; Voita & Titov, 2020).[3] Harding & Sharadin (forthcoming) describe a detailed procedure for how to construct probes for this purpose.

Using (Mechanistic) Interpretability Evaluations, it is clear how we can argue for the presence of a world model: if the probes can recover the state of the environment then we can conclude that the AI system's representations contain a world model of the environment.[4]

3.2. Behavioural Evaluations

An alternative approach to evaluating world models for language models is to look at the input-output behaviour of the AI system. Although world models are internal to the AI system, we might think that if an AI system performs tasks that seem to require it to have a world model, then perhaps we can conclude that it indeed does have a world model. For example, consider an existing approach in the game of chess: given some initial sequence of moves in chess notation, Toshniwal et al. (2022); Li et al. (2023) compare the next moves suggested by a generative model to the valid moves from the current state - if the list of moves generally agree then this provides evidence that the model has developed a world model of the chess board and dynamics.

Vafa et al. (2024) appeal to the Myhill-Nerode theorem (Myhill, 1957; Nerode, 1958) to argue that this next-move test is insufficient and that a behavioural evaluation should instead analyse "minimal distinguishing sequences" which may have length greater than a single move. In Vafa et al. (2024)'s setting, the generative model is evaluated on its ability to realise that two sequences of moves lead to the same state and hence that the valid continuations from the shared state are identical and separately the generative model is evaluated on its ability to realise that two sequences of moves lead to different states and hence should have different valid continuations.

The Behavioural Evaluation approach is tempting as it has the following advantages:

- Behavioural Evaluations are easy to conduct and can be applied with only limited API access to the model and no access to model weights or finetuning access.

- Behavioural Evaluations clearly map onto the real-

[3] A more automated and unsupervised approach to linear probing is to use an interpreter model like a sparse autoencoder (Bricken et al., 2023; Cunningham et al., 2023; Gao et al., 2024; McGrath et al., 2024)

[4] Voita & Titov (2020) suggest that the minimum description length of the probe labels given representations naturally characterise not only the quality of the probe but also the strength of the regularity in representations with respect to the labels, which can provide a guide to the world model's fidelity and task relevance.

world performance of the model if the task is well-constructed to be similar to the analogous real-world task.

- Behavioural Evaluations can be applied to any generative model regardless of implementation details. In particular, this means that the same evaluations can be used for hand-written programs or humans to get baselines.

However, despite these advantages, we note that there is a key problem with Behavioural Evaluations - behavioural evaluations bundle together possessing and using a world model and can only provide evidence for the existence of world models that are used for the AI systems behaviour given the exact task and prompt that is used for the evaluation. We further characterise this problem in Section 4.

4. Possessing And Using World Models

Vafa et al. (2024) use the running example of learning a internal geographic map of the streets of New York City which we will adopt and examine.

4.1. New York Map Example

Example 4.1 *Suppose that Alice has never been to New York before and so we provide her with a complete map (i.e. world model) of the city streets before she visits. She says that she would like to travel to Central Park and she assures us that she will bring and use the map. We also ask Alice to share her GPS location with us so that we can track her movements in case she gets lost. Suppose that when looking at the GPS information we notice that Alice has not taken the direct route to Central Park that was implied by the map.* Can we conclude that Alice is no longer in possession of the map?

Though we *can* conclude that Alice did not successfully use the map to guide her actions, it appears that we *cannot* conclude that Alice was not in possession of the map on her journey. We also cannot conclude that the map was not an effective map: the map could be highly task relevant with high accuracy and fidelity.

It is possible that Alice had a map but decided not to use its directions because she was distracted by the sights of the city, or because wanted to go via a route that her friend suggested or because she was unable to read the map or for any other number of reasons. Looking only at Alice's behaviour is not sufficient to make negative claims about her possessing a world model, to make this claim we would need to either (1) look to see if she's holding the map (analogous to an Interpretability Evaluation) or (2) ensure that before we ask her to go to Central Park that she has the desire, capacity and ability to leverage the map to find the shortest route (analogous to an Elicited Evaluation).

4.2. Possession without Use

We may easily construct an model organism for a generative model which clearly contains a world model and yet does not utilise it. Consider a system W which clearly contains a world model, for example given a Deterministic Finite Automata (DFA) with N states, take an RNN which encodes the transition matrix as its state transition function. Now suppose that we have a generative model V defined by the following algorithm:

- Given some sequence of tokens $x_1, x_2, \ldots, x_n$, first apply W in order to find the correct state that the DFA would be in after processing the sequence of tokens.
- Then, return the <eos> token.

The algorithm represented by V can be realised by adding a single layer to the end of the RNN W. It is clear that V does contain an effective world model, after all V contains W which contains an effective world model. However, V does not use the world model to produce a policy: the policy implied by V is to always return the <eos> token regardless of the input sequence. Upon behaviourally evaluating this system, we are likely to conclude that it does not, in fact, possess a world model which would be incorrect.

We provide more realistic examples of agents that possessed yet do not use their world model in Appendix A.

Note that although if an agent possesses a world model that it can never access or use then this is not a practical problem: the societal implications of such an agent are similar to an agent which doesn't have a world model at all. However, if an agent has a world model that it can use, but doesn't use in the scope of our evaluation task, then we are likely to draw false inferences from our evaluation about the model's capabilities. This is in particular relevant if other actors may be able to elicit capabilities from the agent that were not apparent from the evaluation. Here the elicitation could be via specialised prompting, agentic scaffolding, supervised finetuning or reinforcement learning finetuning which allow the model to use its world model and quickly appear to be much more behaviourally capable.

4.3. Towards Elicited Evaluations

As we have argued, Behavioural Evaluations fail to provide evidence against an agent possessing but not using a world model. To make this claim we would like to confirm that with significant effort we were not able to elicit behaviour that corresponds to using the world model from the agent. We suggest that researchers use the following protocol for

Elicitation Evaluations to make stronger claims about agents not possessing an effective world model:

- Find or construct a Behavioural Evaluation which appropriately tests for the use of an effective world model for your task.
- Enumerate a list of elicitation strategies that you will use to help the model leverage its world model if it has one.
- For each elicitation strategy, run the Behavioural Evaluation.
  - If the model still cannot complete the task then the modeller can have increased confidence in the claim that the model does not possess an effective world model.
  - If the model now can complete the task, then we would like to show that the Elicitation technique was uncovering, rather than adding to, the agent's world model. For example, for finetuning elicitations, we might show that the same amount of compute and training for a similar model did not lead to a significant change in the agent's output and this amount of compute is generally not enough to add a new capability to the model.

For example, if the purpose of our evaluations is to show that malicious actors cannot elicit harmful capabilities from the agent for cyber-misuse, which would derive from an agent having an effective world model, then if the elicitation strategy uses more skilled developer time than the the threat actors are expected to then we can be confident that the agent cannot be easily manipulated into exhibiting harmful behavior by the threat actors.

5. Discussion

Researchers often try to make two types of claims from evaluations: *capability claims* (i.e. the agent has capability X or possesses an effective world model for task T) and *inability claims* (i.e. the agent does not have capability X or possess effective world model for task T). We suggest that Behavioural Evaluations alone are insufficient to make inability claims. We can conclude that either (1) that the AI system does not possess an effective world model or (2) that AI system does not use its effective world model to produce an effective policy for the task. Failure on the behavioural evaluation could be a failure of possessing a world model or a failure of using the world model that it possesses.

Frontier AI agents often do not demonstrate their full capabilities without targeted elicitation. In particular, we may be interested in to what extent an agent has a world model but without targeted elicitation, frontier model developers, policymakers and safety researchers may underestimate frontier AI agents. We hence believe that Interpretability Evaluations and Elicitation Evaluations are important for assessing the current state of model capabilities. Understanding the capabilities and limitations of frontier AI systems is critical for Responsible Scaling Policies (OpenAI, 2025), effective AI governance (Bengio et al., 2025; Emanuilov, 2024), safety cases (Buhl et al., 2024; Clymer et al., 2024; Hilton et al., 2025; Buhl et al., 2025), effective feedback loops for AI researchers and for technical predictions about the impact of AI systems on society (Phuong et al., 2024).

The possession-use distinction observed in AI systems parallels well-documented phenomena in biological cognition. Humans often demonstrate implicit knowledge through behavior while failing explicit tests of the same knowledge—such as successfully navigating complex environments yet being unable exhibit the same behaviour when the conditions are subtly changed (Kahneman, 2011). Similarly, while rodent place and grid cells in the hippocampus clearly encode detailed spatial world models (Hafting et al., 2005; O'Keefe & Dostrovsky, 1971), behavioural studies show these animals don't always utilize this spatial knowledge optimally, sometimes persisting with suboptimal routes despite having neural representations of better paths (Tolman, 1948). These biological precedents suggest that the gap between possessing and using world models reflects fundamental architectural constraints in how cognitive systems access internal knowledge.

5.1. Future Work

Though many elicitation methods have been used in previous work there is not currently a consensus on the best approach to elicitation which can be used for Elicitation Evaluations. We would be interested in seeing further work on approaches to elicitation. We would be excited about researchers using evaluating how effective different elicitation method are using model organisms methods similar to (Greenblatt et al., 2024b; Hofstätter et al., 2025).

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., and Søgaard, A. Can language models encode perceptual structure without grounding? a case study in color. *arXiv preprint arXiv:2109.06129*, 2021.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi:10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., Newman, J., Ng, K. Y., Okolo, C. T., Raji, D., Sastry, G., Seger, E., Skeadas, T., South, T., Strubell, E., Tram'er, F., Velasco, L., Wheeler, N., Acemoglu, D., Adekanmbi, O., Dalrymple, D., Dietterich, T. G., Felten, E. W., Fung, P., Gourinchas, P.-O., Heintz, F., Hinton, G., Jennings, N., Krause, A., Leavy, S., Liang, P., Ludermir, T., Marda, V., Margetts, H., McDermid, J., Munga, J., Narayanan, A., Nelson, A., Neppel, C., Oh, A., Ramchurn, G., Russell, S., Schaake, M., Sch"olkopf, B., Song, D., Soto, A., Tiedrich, L., Varoquaux, G., Yao, A., Zhang, Y.-Q., Ajala, O., Albalawi, F., Alserkal, M., Avrin, G., Busch, C., de Carvalho, A. C. P. d. L. F., Fox, B., Gill, A. S., Hatip, A. H., Heikkil"a, J., Johnson, C., Jolly, G., Katzir, Z., Khan, S. M., Kitano, H., Kr"uger, A., Lee, K. M., Ligot, D. V., L'opez Portillo, J. R., Molchanovskyi, O., Monti, A., Mwamanzi, N., Nemer, M., Oliver, N., Pezoa Rivera, R., Ravindran, B., Riza, H., Rugege, C., Seoighe, C., Sheehan, J., Sheikh, H., Wong, D., and Zeng, Y. International ai safety report. Technical Report DSIT 2025/001, 2025. URL https://www.gov.uk/government/publications/international-ai-safety-report-2025.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bubeck, S., Chadrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Buhl, M. D., Sett, G., Koessler, L., Schuett, J., and Anderljung, M. Safety cases for frontier ai. *arXiv preprint arXiv:2410.21572*, 2024.

Buhl, M. D., Pfau, J., Hilton, B., and Irving, G. An alignment safety case sketch based on debate. *arXiv preprint arXiv:2505.03989*, 2025.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., and Wong, E. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

Clymer, J., Gabrieli, N., Krueger, D., and Larsen, T. Safety cases: How to justify the safety of advanced ai systems. *arXiv preprint arXiv:2403.10462*, 2024.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Delétang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Wenliang, L. K., Aitchison, M., Orseau, L., et al. Language modeling is compression. *arXiv preprint arXiv:2309.10668*, 2023.

Emanuilov, I. General purpose ai models. In *CiTiP Webinar-Exploring 5 key topics under the EU AI Act, Date: 2024/02/29-2024/02/29, Location: Online*, 2024.

Engels, J., Liao, I., Michaud, E. J., Gurnee, W., and Tegmark, M. Not all language model features are linear. *arXiv e-prints*, pp. arXiv–2405, 2024.

Feng, J., Russell, S., and Steinhardt, J. Monitoring latent world states in language models with propositional probes. *arXiv preprint arXiv:2406.19501*, 2024.

Fodor, J. A. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press, 1987.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders, June 2024. URL http://arxiv.org/abs/2406.04093. arXiv:2406.04093 [cs], version: 1.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024a.

Greenblatt, R., Roger, F., Krasheninnikov, D., and Krueger, D. Stress-testing capability elicitation with password-locked models. *arXiv preprint arXiv:2405.19550*, 2024b.

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Gurnee, W. and Tegmark, M. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Ha, D. and Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Hafner, D., Lillicrap, T., Norouzi, M., and Ba, J. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., and Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436:801–806, 2005. URL https://api.semanticscholar.org/CorpusID:4405184.

Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., and Hu, Z. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

Harding, J. and Sharadin, N. What is it for a machine learning model to have a capability? *British Journal for the Philosophy of Science*, forthcoming. doi:10.1086/732153.

Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

Hilton, B., Buhl, M. D., Korbak, T., and Irving, G. Safety cases: A scalable approach to frontier ai safety. *arXiv preprint arXiv:2503.04744*, 2025.

Hofstätter, F., van der Weij, T., Teoh, J., Bartsch, H., and Ward, F. R. The elicitation game: Evaluating capability elicitation techniques. *arXiv preprint arXiv:2502.02180*, 2025.

Jenner, E., Kapur, S., Georgiev, V., Allen, C., Emmons, S., and Russell, S. J. Evidence of learned look-ahead in a chess-playing neural network. *Advances in Neural Information Processing Systems*, 37:31410–31437, 2024.

Jung, J. C. et al. Case study: Volkswagen's diesel emissions scandal. *Thunderbird International Business Review*, 59 (1), 2017.

Kahneman, D. *Thinking, fast and slow*. macmillan, 2011.

Karvonen, A. Emergent world models and latent variable estimation in chess-playing language models. *arXiv preprint arXiv:2403.15498*, 2024.

Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., Miller, H., et al. Dspy: Compiling declarative language model calls into self-improving pipelines. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

Kiciman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR*, 2023.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Liu, H., Yan, W., Zaharia, M., and Abbeel, P. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024.

Lubana, E. S., Kawaguchi, K., Dick, R. P., and Tanaka, H. A percolation model of emergence: Analyzing transformers trained on a formal language. *arXiv preprint arXiv:2408.12578*, 2024.

McGrath, T., Balsam, D., Gorton, L., Cubuktepe, M., Deng, M., Nguyen, N., Jain, A., Shihipar, T., and Ho, E. Mapping the latent space of llama 3.3 70b. *Goodfire Research*, 2024. URL https://www.goodfire.ai/papers/mapping-latent-spaces-llama.

Mikolov, T., Yih, W.-t., and Zweig, G. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013.

Myhill, J. Finite automata and the representation of events. *WADD Technical Report*, 57:112–137, 1957.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023a.

Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023b.

Nerode, A. Linear automaton transformations. *Proceedings of the American Mathematical Society*, 9(4):541–544, 1958.

Ngo, R., Chan, L., and Mindermann, S. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

O'Keefe, J. and Dostrovsky, J. The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175, 1971. ISSN 0006-8993. doi:https://doi.org/10.1016/0006-8993(71)90358-1. URL https://www.sciencedirect.com/science/article/pii/0006899371903581.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020. doi:10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

OpenAI. Preparedness Framework (version 2). https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf, April 2025. Last updated 15 April 2025.

Park, C. F., Lee, A., Lubana, E. S., Yang, Y., Okawa, M., Nishi, K., Wattenberg, M., and Tanaka, H. Iclr: In-context learning of representations. *arXiv preprint arXiv:2501.00070*, 2024.

Parr, T., Pezzulo, G., and Friston, K. J. *Active inference: the free energy principle in mind, brain, and behavior*. 2022.

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, 2023.

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., et al. Evaluating frontier models for dangerous capabilities. *arXiv preprint arXiv:2403.13793*, 2024.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Richens, J. and Everitt, T. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024.

Sutton, R. S., Barto, A. G., et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Taufeeque, M., Quirke, P., Li, M., Cundy, C., Tucker, A. D., Gleave, A., and Garriga-Alonso, A. Planning in a recurrent neural network that plays sokoban. *arXiv preprint arXiv:2407.15421*, 2024.

Tolman, E. C. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948. doi:10.1037/h0061626.

Toshniwal, S., Wiseman, S., Livescu, K., and Gimpel, K. Chess as a testbed for language model state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11385–11393, 2022.

Vafa, K., Chen, J., Rambachan, A., Kleinberg, J., and Mullainathan, S. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975, 2024.

van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., and Ward, F. R. Ai sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*, 2024.

Voita, E. and Titov, I. Information-theoretic probing with minimum description length. *arXiv preprint arXiv:2003.12298*, 2020.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.

Welleck, S., Bertsch, A., Finlayson, M., Schoelkopf, H., Xie, A., Neubig, G., Kulikov, I., and Harchaoui, Z. From

decoding to meta-generation: Inference-time algorithms for large language models. *Transactions on Machine Learning Research*, 2024.

Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky, D., Manning, C. D., and Potts, C. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*, 2025.

Zaharia, M., Khattab, O., Chen, L., Davis, J. Q., Miller, H., Potts, C., Zou, J., Carbin, M., Frankle, J., Rao, N., and Ghodsi, A. The shift from models to compound ai systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/, 2024.

A. Examples of Possessing but Not Using World Models in the Wild

In Section 4.2, we presented a simple example of a model that possesses a world model but does not use it. Here we provide some more realistic examples which are likely to have a mismatch between the possession and usage of world models.

Firstly, consider the practise of Capability Elicitation for LLM agents. It has been shown that several elicitation strategies (such as Chain of Thought prompting, prompt engineering, agent scaffolding and finetuning) can boost model capabilities (Wei et al., 2022b; Brown et al., 2020; Zaharia et al., 2024; Khattab et al., 2023; Guo et al., 2024; Welleck et al., 2024). These techniques do not generally teach the model new capabilities but rather help the model to leverage the capabilities that it already has. Given that many times in the past, we have discovered new techniques which make existing agents more capable and reveal world models that were previously hidden, we might expect future elicitation work to also reveal other new capabilities.

Secondly, van der Weij et al. (2024) and Greenblatt et al. (2024a) suggest that future AI models which are situationally aware (Ngo et al., 2022) strategically underperform in evaluations to hide capabilities from auditors. This intentional hiding of capabilities is known as "sandbagging". We may draw an analogy to the Volkswagen emissions scandal (Jung et al., 2017) where the company intentionally designed their cars to pass emissions tests while actually emitting more pollutants than allowed, and the cars were functionally different in the evaluation and deployment settings. Similarly, Perez et al. (2023) found that some LLMs gave worse answers to users introducing themselves as uneducated.

Thirdly, in the process of "grokking" (Power et al., 2022), Nanda et al. (2023a) find that LMs move from utilising a memorising strategy to a generalising strategy over time. In particular, Nanda et al. (2023a) find that the world model corresponding to the generalising strategy is present before the generalising strategy is behaviourally used and this generalising circuit is generally upweighted over time. In this way we see that during the early stages of grokking, the LM directly contains the generalising world model but does not use it.

Fourthly, language model providers often apply finetuning to agents to make agents more helpful and harmless (Bai et al., 2022). Throughout this process, some harmful capabilities that the model has are likely to not be exhibited and hence behavioural evaluations are likely to suggest that the model lacks the relevant effective world model to complete the task. These hidden capabilities can be revealed through elicitation techniques as well as through jailbreaks (Chao et al., 2023).

We would be excited about more future work that explores the extent to which frontier language models possess world models that are not used for behaviour in practically relevant settings.