Stress-Testing Reasoning with Causal World Models

Jacqueline Maasch
Cornell Tech, New York, NY

John Kalantari
YRIKKA, New York, NY

Kia Khezeli
YRIKKA, New York, NY

Abstract

This work probes the failure modes of frontier language models when reasoning over synthetic causal worlds, with an emphasis on counterfactual reasoning, logical reasoning, and abstract reasoning with program synthesis. Performance varied widely between and within tasks, with some models displaying greater aptitude for either transduction or induction tasks.

Introduction Humans are exceptional few- and zero-shot learners that use internal representations of the world to reason and plan in uncertain environments [9, 13, 14]. These internal *world models* can encode beliefs about cause-effect relationships, supporting reasoning at all three levels of the Pearl Causal Hierarchy (PCH): observing factual realities (L1), exerting actions to induce interventional realities (L2), and imagining alternate counterfactual realities (L3) [1]. Though causal reasoning is a hallmark of human cognition [6] and a desideratum for human-like AI [20, 12], state-of-the-art language models (LMs) do not yet display robust reasoning at all three levels of the PCH [11, 8, 27, 24, 10, 15]. Construct validity is a persistent challenge in LM reasoning evaluation [3, 2, 22], and causal reasoning benchmarks often suffer from critical design flaws (e.g., data contamination or weak theoretical justifications) [26]. To address the need for improved evaluation frameworks, we explore the utility of causal world modeling in reasoning evaluation.

A Framework for Stress-Testing Reasoning This work probes the failure modes of LMs when performing abstract, logical, and counterfactual reasoning tasks from the CausalARC testbed (Fig. 1) [16]. Tasks are sampled from ground truth *causal world models* expressed as probabilistic *structural causal models* (SCMs) [20, 21]. Given a fully specified SCM, all three levels of the PCH are well-defined: any L1, L2, or L3 query about the world model can be answered [1]. By framing each

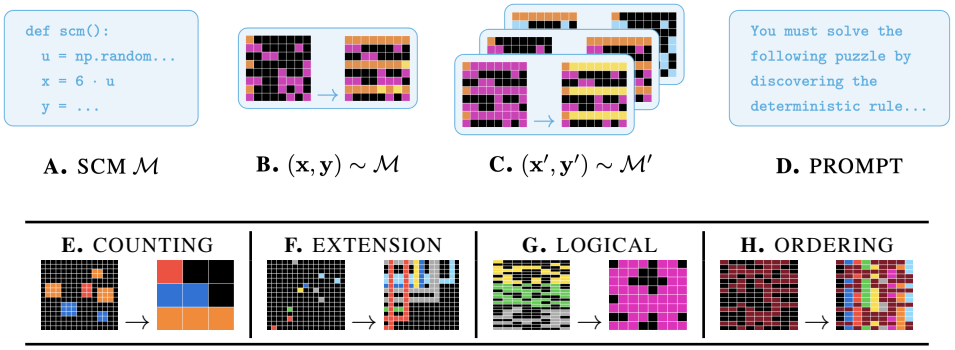

Figure 1: The CausalARC testbed [16], an extension of ARC-AGI [4, 5] (https://arcprize.org/arc-agi). (**A**) First, SCM $\mathcal{M}$ is manually transcribed in Python code. (**B**) Input-output pairs $(\mathbf{x}, \mathbf{y})$ are randomly sampled, providing L1 learning signals about the deterministic transformation that maps inputs to outputs. (**C**) Sampling from interventional submodels $\mathcal{M}'$ of $\mathcal{M}$ yields L2 samples $(\mathbf{x}', \mathbf{y}')$. Given $(\mathbf{x}, \mathbf{y})$, performing multiple interventions while holding the exogenous context constant yields a set of L3 pairs. (**D**) L1, L2, and/or L3 pairs provide in-context demonstrations for natural language prompts. (**E–H**) As in ConceptARC [19], tasks are annotated by theme for detailed error analyses. Code and data can be found at our **project page**: https://jmaasch.github.io/carc/

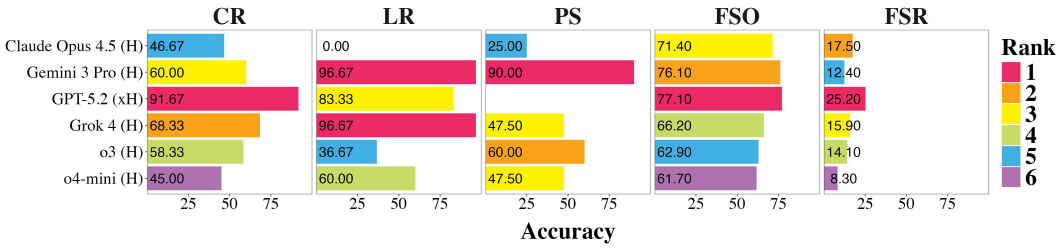

Figure 2: Accuracy with maximum reasoning effort on CR (six counting, extension, and ordering tasks), LR (three logical tasks), and PS (four counting and extension tasks). Each task used five prompts with random L1 examples and five with alternating L1/L3 examples (see [16]). CR prompts used three input-output demonstrations, LR used four, and PS used six. Performance ranged more on CausalARC than FrontierScience-Olympiad (FSO) and Research (FSR; as reported by OpenAI).

task as a generative model, robustness can be benchmarked with respect to a distribution of task instances. Generative benchmarks buffer against static benchmark data leakage, a major challenge in AI evaluation [17, 25, 15, 7]. ARC-like settings address the nontrivial challenge of differentiating true reasoning from factual recall [10, 25, 17, 23], as abstract reasoning relies more on innate cognitive priors than factual knowledge [4]. This formulation makes CausalARC an open-ended playground for testing diverse reasoning hypotheses at all three levels of the PCH.

Experimental Settings We tested three settings with few-shot, in-context learning.

1. *Counterfactual reasoning* (**CR**). Predict the counterfactual array $\mathbf{y}'$ given an intervention on $\mathbf{x}$.
2. *Program synthesis for abstract reasoning* (**PS**). Output Python code that generates $\mathbf{y}$ given $\mathbf{x}$.
3. *Logical reasoning* (**LR**). Identify logical functions governing parent-child relationships in $(\mathbf{x}, \mathbf{y})$.

Setting (1) represents *neural transduction* (i.e., the output is directly predicted), while (2) represents *neural induction* (i.e., the LM generates a program implementing the transformation rule, which is executed to obtain the output array) [18]. We evaluated impacts of prompt formulation by varying the PCH level of in-context examples (all L1 vs alternating L1/L3). An array was deemed correct if all cells were correct. Models were tested at their maximum reasoning effort (H = high, xH = extra high) and/or default.

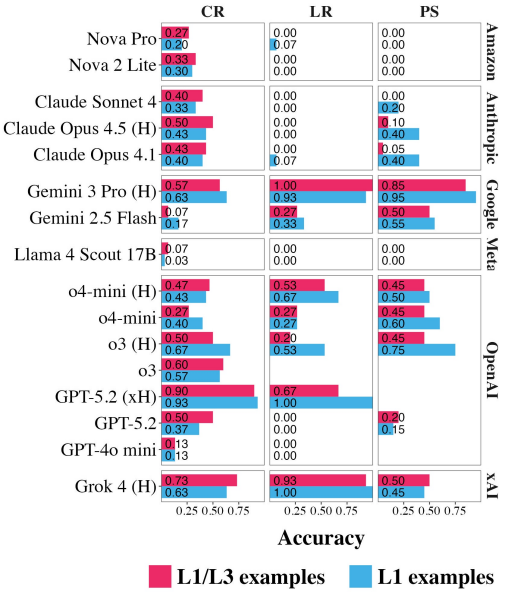

Figure 3: Accuracy on CR, LR, and PS.

General Results LM rankings varied by task, with no across-the-board winner (Figs. 2, 3). While benchmarks like FrontierScience are a tight race, LM performance varied widely on CausalARC. Some models were better at transduction than induction (e.g., Grok 4 (H), Nova) or vice versa (e.g., Gemini 2.5 Flash, o4-mini).

Do L3 Examples Help? L3 in-context demonstrations resulted in better or equal accuracy on CR tasks for 11/17 LMs (64.7%; Fig. 3). L3 demonstrations rarely conferred benefits for LR (2/15 LMs; 13.3%) or PS (2/13 LMs; 15.4%). Future work could explore the utility of fine-tuning on synthetic L3 data, or the use of counterfactual consistency (equal accuracy on L1 vs L3 queries) as a reward signal for reinforcement learning.

Speculations on Training CausalARC performance might provide hints about the closed-source training regimes of proprietary models. Results include some evidence of fine-tuning for program synthesis, reasoning with Boolean logic, and ARC-like abstract reasoning: o4-mini offered "ARC-style input/output" unprompted; Claude Sonnet 4 volunteered to perform program synthesis unprompted (and yet performed poorly on PS); and Grok 4 and Gemini 3 Pro were near-perfect on LR tasks.

References

[1] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*, page 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL https://doi.org/10.1145/3501714.3501743.

[2] A. Bean and et al. Measuring what matters: Construct validity in large language model benchmarks. In *Thirty-ninth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/pdf?id=mdA5lVvNcU.

[3] Z. Cheng, S. Wohnig, R. Gupta, S. Alam, T. Abdullahi, J. A. Ribeiro, C. Nielsen-Garcia, S. Mir, S. Li, J. Orender, et al. Benchmarking is broken–don't let ai be its own judge. *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025.

[4] F. Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[5] F. Chollet, M. Knoop, G. Kamradt, and B. Landers. Arc prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*, 2024.

[6] M. K. Goddu and A. Gopnik. The development of human causal learning and reasoning. *Nature Reviews Psychology*, pages 1–21, 2024.

[7] A. Gong, K. Stankevičiūtė, C. Wan, A. Kabra, R. Thesmar, J. Lee, J. Klenke, C. P. Gomes, and K. Q. Weinberger. Phantomwiki: On-demand datasets for reasoning and retrieval evaluation. In *International Conference on Machine Learning*, 2025.

[8] J. González and A. Nori. Does reasoning emerge? examining the probabilities of causation in large language models. *Advances in Neural Information Processing Systems*, 37:117737–117761, 2024.

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

[10] A. Hüyük, X. Xu, J. Maasch, A. V. Nori, and J. González. Reasoning elicitation in language models via counterfactual feedback. *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.03767.

[11] Z. Jin, Y. Chen, F. Leeb, L. Gresele, O. Kamal, Z. Lyu, K. Blin, F. Gonzalez Adauto, M. Kleiman-Weiner, M. Sachan, et al. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 2023.

[12] B. M. Lake and M. Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.

[13] B. M. Lake, T. Linzen, and M. Baroni. Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*, 2019.

[14] Y. LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[15] J. Maasch, A. Hüyük, X. Xu, A. V. Nori, and J. Gonzalez. Compositional causal reasoning evaluation in language models. *International Conference on Machine Learning*, 2025.

[16] J. R. Maasch, J. Kalantari, and K. Khezeli. Causalarc: Abstract reasoning with causal world models. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*, 2025.

[17] S. I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, 2025.

[18] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2018.

[19] A. Moskvichev, V. V. Odouard, and M. Mitchell. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *Transactions on Machine Learning Research*, 2023.

[20] J. Pearl. Causality: Models, reasoning, and inference. *Cambridge, UK: Cambridge University Press*, 19(2):3, 2000.

[21] J. Pearl. Structural counterfactuals: A brief introduction. *Cognitive Science*, 37(6):977–985, 2013.

[22] I. D. Raji, E. Denton, E. M. Bender, A. Hanna, and A. Paullada. Ai and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[23] P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, and M. Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*, 2025.

[24] R. B. Shrestha, S. Malberg, and G. Groh. From causal parrots to causal prophets? towards sound causal reasoning with large language models. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 319–333, 2025.

[25] X. Xu, R. Lawrence, K. Dubey, A. Pandey, R. Ueno, F. Falck, A. V. Nori, R. Sharma, A. Sharma, and J. Gonzalez. Re-imagine: Symbolic benchmark synthesis for reasoning evaluation. In *International Conference on Machine Learning*, 2025.

[26] L. Yang, V. Shirvaikar, O. Clivio, and F. Falck. A critical review of causal reasoning benchmarks for large language models. In *AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?"*, 2024.

[27] M. Zečević, M. Willig, D. S. Dhami, and K. Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2024.