Trusted AI Symposium

Stress-Testing Reasoning with Causal World Models

Jacqueline Maasch[1], John Kalantari[2], Kia Khezeli[2]
[1]Cornell Tech, New York, NY [2]YRIKKA, New York, NY

Amazon Nova

Abstract

Contribution. This work probes the failure modes of frontier language models (LMs) when reasoning over synthetic causal worlds, with an emphasis on **counterfactual reasoning, logical reasoning, abstract reasoning, and program synthesis**. Models were tested on CausalARC: an experimental testbed for AI reasoning in low-data and out-of-distribution regimes [1], modeled after the Abstraction and Reasoning Corpus (ARC) [2]. Each reasoning task is sampled from a fully specified *causal world model*, formally expressed as a structural causal model (SCM). Principled data augmentations provide observational, interventional, and counterfactual feedback about the world model in the form of few-shot, in-context learning demonstrations. Within- and between-model performance varied heavily across tasks, indicating room for significant improvement in LM reasoning.

Background

ARC-AGI Benchmark for Fluid Intelligence. ARC [2] is a grid world of two-dimensional arrays (1×1 to 30×30) where pixels can take on one of ten colors each. The test-taker must solve each task by discovering the deterministic rule or transformation that maps input arrays **x** to output arrays **y**. Each task provides approximately 2–5 input-output pairs (**x**,**y**) as examples to demonstrate the rule, with no additional clues provided. An average human should ostensibly be able to solve most or all tasks from these demonstrations alone, with no specialized knowledge or training. Instead, problem-solving requires innate cognitive priors (elementary arithmetic, basic geometry, intuitive physics, etc.).

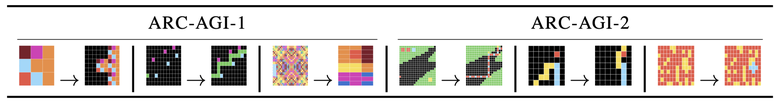

Fig 1. Example input-output pairs from ARC-AGI-1 and ARC-AGI-2.

Structural Causal Model (SCM). An SCM is a tuple $M := \langle \mathbf{U}, p(\mathbf{u}), \mathbf{V}, F \rangle$ where **U** is the set of exogenous variables explained by mechanisms external to M, $p(\mathbf{u})$ is the distribution over **U**, **V** is the set of endogenous variables explained by variables in **U** ∪ **V**, and F is the set of structural functions such that $v_i = f_i(pa_{vi}, u_i)$ for endogenous parent set pa_{vi} and exogenous context u_i [3].

Counterfactual. Let M_x be the submodel of M induced by an intervention on X in **V**. Let Y in **V** be a variable whose value we wish to query. The counterfactual Y_x under model M is then expressed as $Y_x(\mathbf{u}) := Y_{Mx}(\mathbf{u})$. For example, when $Y_x(\mathbf{u}) = y$: "Y would have been y had X been x when **U** = **u**."

The Pearl Causal Hierarchy (PCH). Let M be a fully specified SCM. The PCH is the set of all observational (layer L1), interventional (layer L2), and counterfactual (layer L3) distributions induced by M [3].

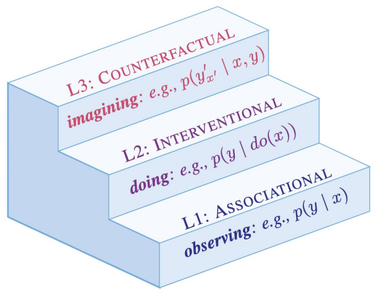

Fig 2. The PCH: observing factual realities (L1), exerting actions to induce interventional realities (L2), and imagining alternate counterfactual realities (L3).

The CausalARC Reasoning Testbed

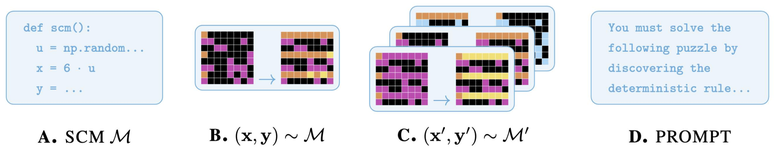

Fig 3. The CausalARC testbed.

Reasoning tasks as generative models. From Fig 3: (**A**) First, SCM M is transcribed in Python code. (**B**) Input-output pairs are randomly sampled, providing observational (L1) learning signals about the world model. (**C**) Sampling from interventional submodels M' of M yields interventional (L2) samples (**x**',**y**'). Given pair (**x**,**y**), performing multiple interventions while holding the exogenous context constant yields a set of counterfactual (L3) pairs. (**D**) Using L1 and L3 pairs as in-context demonstrations, we can generate LM prompts for diverse reasoning tasks.

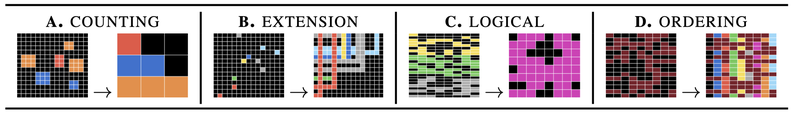

Fig 4. Example demonstration pairs for CausalARC themes.

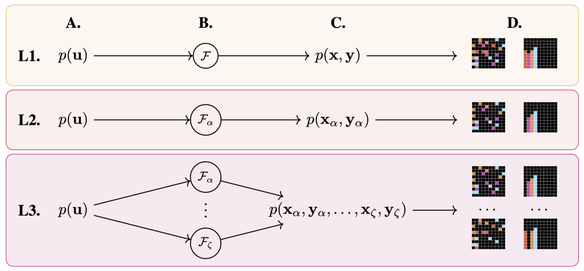

Fig 5. Unlike real-world data settings, counterfactuals can be jointly observed in the synthetic causal worlds of CausalARC. L1, L2, and L3 denote the rungs of the PCH (Fig 2). Figure adapted from [3].

Jointly observed counterfactuals. From Fig 5: (**A**) The distribution over the exogenous context (i.e., the external state). (**B**) Transformations applied to the exogenous context (e.g., functions F in the observational world; updated functions F_α under intervention α). (**C**) Induced distributions, following from the applied transformation. (**D**) CausalARC samples from each rung of the PCH.

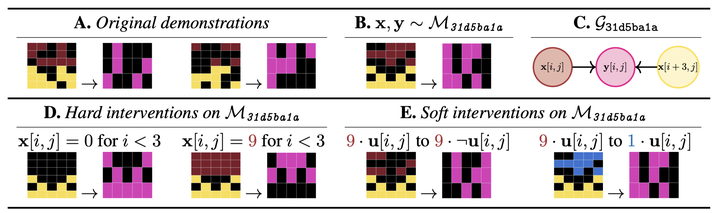

Fig 6. Examples of hard interventions, soft interventions, and the causal graph for a CausalARC task.

Empirical Results

Abstract Reasoning with Test-Time Training (TTT). To gauge the difficulty of CausalARC, we benchmarked MARC with TTT on the full static dataset [4]. MARC was the second-place paper winner for ARC-AGI-1. It takes a neural transduction approach using a Llama 3 8B base model fine-tuned on large ARC-like datasets, with TTT plus in-context learning at inference.

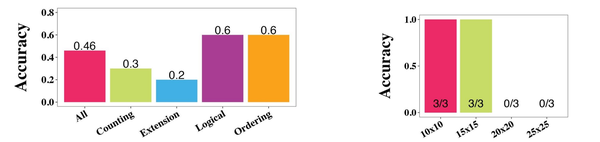

Fig 7. (Left) Accuracy on the full CausalARC static dataset (50 tasks) by theme for MARC with TTT (Llama 3 8B base). (Right) Performance on logical *and*, *or*, and *xor* tasks sampled from SCMdky5 as array size increases.

MARC's overall accuracy of 46% on CausalARC is similar to its pure transduction score of 47.1% accuracy on ARC-AGI-1, **suggesting that CausalARC is of comparable difficulty to ARC-AGI-1**.

Reasoning with Few-Shot, In-Context Learning. We tested three settings:

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

Transduction vs Induction. 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).

Setup. Each task used five prompts with L1 examples and five with alternating L1/L3 examples. We compare to accuracy on FrontierScience-Olympiad (FSO) and Research (FSR; as reported by OpenAI). Models were tested at max reasoning effort (H = high; xH = extra high) and at default.

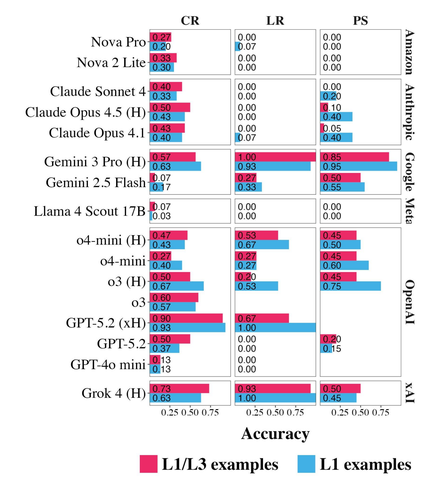

Fig 8. Accuracy on CR, LR, and PS.

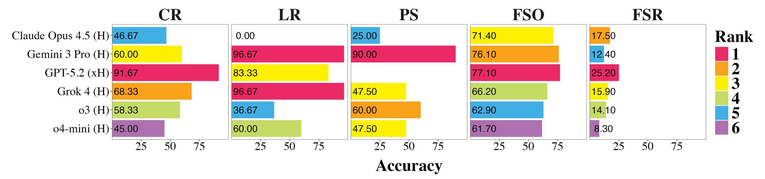

Fig 9. Accuracy with max reasoning effort on CR (six counting, extension, and ordering tasks), LR (three tasks), and PS (four counting and extension tasks).

References. [1] Maasch, Kalantari, and Khezeli. Causalarc: Abstract Reasoning with Causal World Models. In NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning, 2025. [2] Chollet. On the Measure of Intelligence. arXiv preprint arXiv:1911.01547, 2019.
[3] Bareinboim, et al. On Pearl's Hierarchy and the Foundations of Causal Inference. ACM, 2022. [4] Akyürek, et al. The Surprising Effectiveness of Test-Time Training for Few-Shot Learning. International Conference on Machine Learning, 2025.

*Corresponding author: maasch@cs.cornell.edu