CausalARC

Abstract 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

Abstract. This work introduces CausalARC: an experimental testbed for AI reasoning in low-data and out-of-distribution regimes, modeled after the Abstraction and Reasoning Corpus (ARC) [3]. Each reasoning task is sampled from a fully specified *causal world model*, formally expressed as a structural causal model (SCM). LLM performance on CausalARC varied heavily across tasks:
(1) **abstract reasoning with test-time training**, (2) **counterfactual reasoning**, (3) **program synthesis**, (4) **causal discovery**.

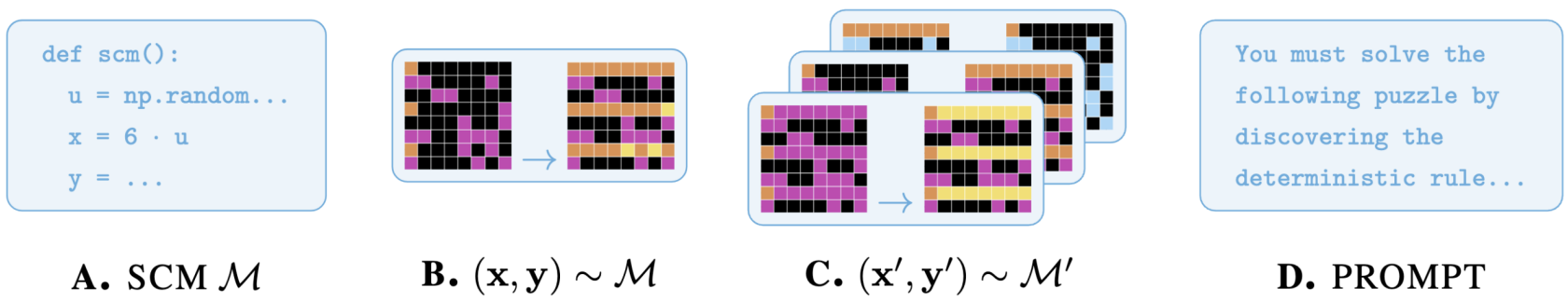

Figure 1: **The CausalARC testbed.** (**A**) First, SCM $\mathcal{M}$ is manually transcribed in Python code. (**B**) Randomly sampled input-output pairs offer observational learning signals about the world model. (**C**) Sampling from submodels $\mathcal{M}'$ of $\mathcal{M}$ yields interventional samples $(\mathbf{x}', \mathbf{y}')$. Performing multiple interventions while holding the exogenous context constant yields a set of counterfactual pairs. (**D**) Automatically generated prompts with in-context demonstrations.

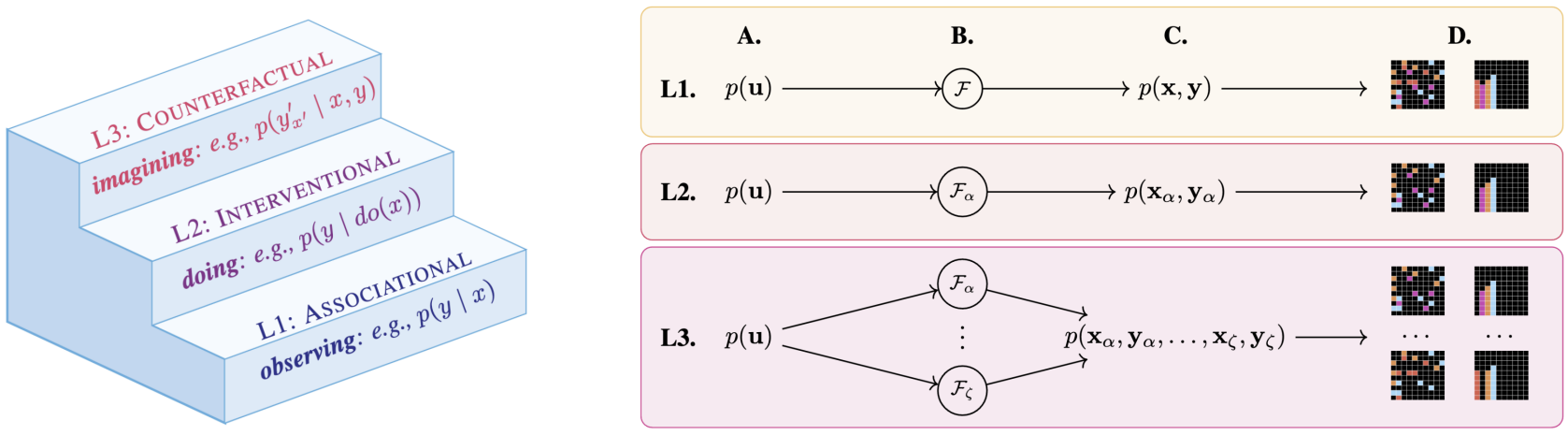

Figure 2: **[Left] The Pearl Causal Hierarchy (PCH)**: observing factual realities (L1), exerting actions to induce interventional realities (L2), and imagining alternate counterfactual realities (L3) [2]. **[Right] Jointly observed counterfactuals in CausalARC.** (**A**) The distribution over the exogenous context (i.e., the external state). (**B**) Transformations applied to the exogenous context (e.g., functions $\mathcal{F}$ in the observational world; updated functions $\mathcal{F}_\alpha$ under intervention α). (**C**) Induced distributions, following from the applied transformation. (**D**) CausalARC samples from each rung of the PCH. Adapted from [2].

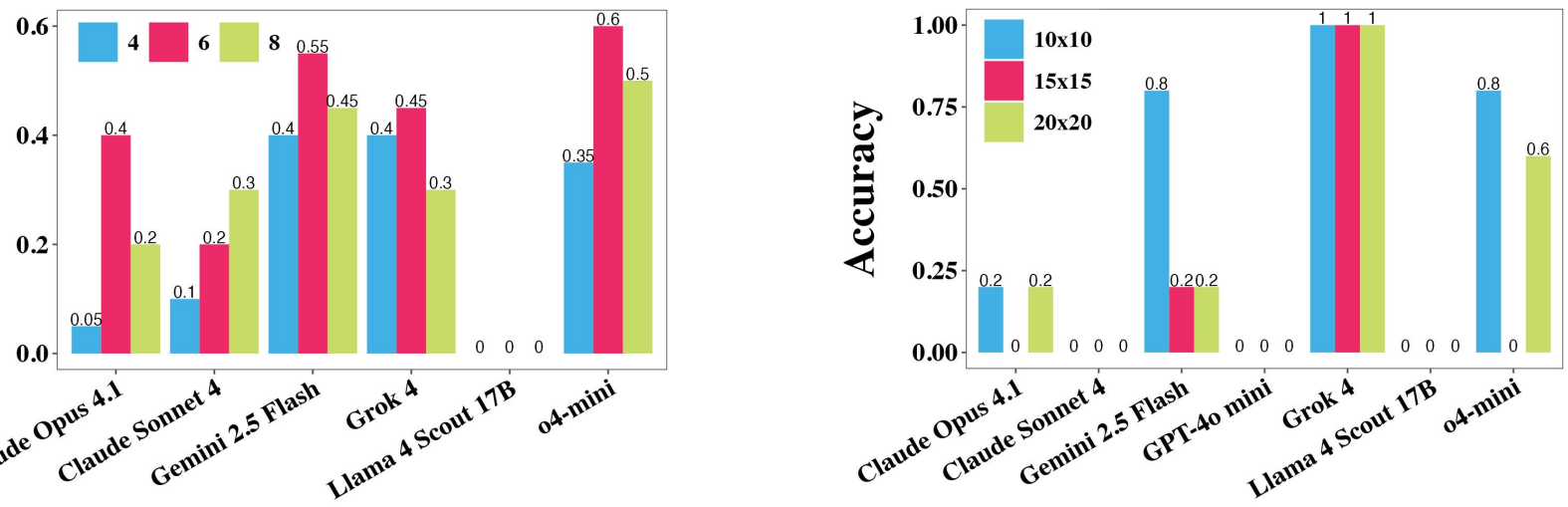

Figure 3: **[Left] Program synthesis** on four counting and extension tasks as total in-context demonstrations increased. **[Right] Causal discovery** with logical reasoning as array size increased. Scores were over five random L1 prompt samples.

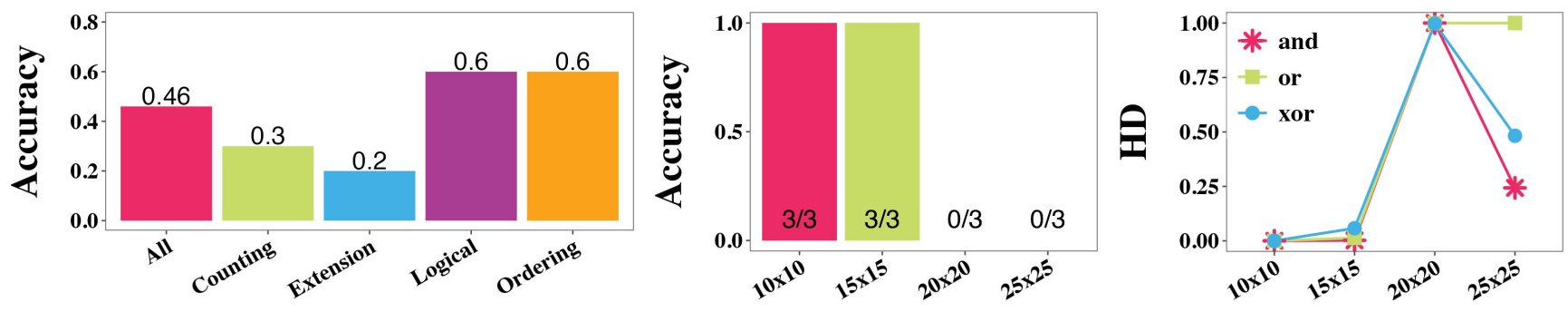

Figure 4: **Abstract reasoning with test-time training** (TTT). **[Left]** Accuracy by CausalARC theme for MARC with TTT (Llama 3 8B base) [1]. **[Center, right]** Performance on *and*, *or*, and *xor* tasks sampled from CausalARC task `SCMdky5` as array size increases.

[1] E. Akyürek, et al. ICML 2025. [2] E. Bareinboim, et al. ACM 2022. [3] F. Chollet, et al. ARC Prize 2024: Technical Report.

NeurIPS 2025 Workshop: Bridging Language, Agent, and World Models for Reasoning and Planning (LAW)