Human Adults and LLMs as Scientists: Who Benefits from Active Exploration?

Mandana Samiei[1,2*] Eunice Yiu[3*] Anthony GX-Chen[4] Dongyan Lin[5] Jocelyn Shen[6]
Blake A. Richards[1,2,7] Alison Gopnik[3] Doina Precup[1,2]

[1]Mila - Quebec AI Institute [2]McGill University [3]UC Berkeley [4]NYU [5]Meta FAIR [6]MIT Media Lab [7]Montreal Neurological Institute

*Equal contribution mandana.samiei@mail.mcgill.ca ey242@berkeley.edu

Motivation

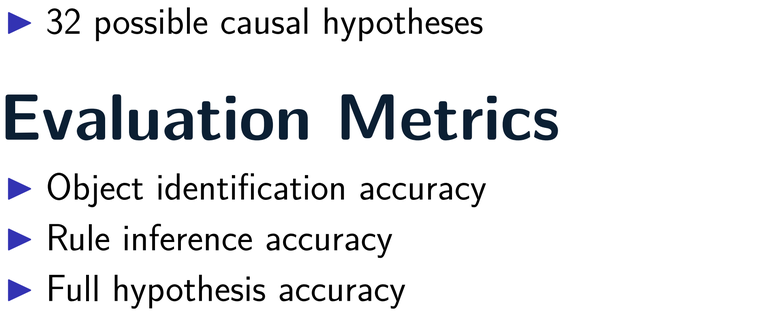
Adults often appear biased toward **disjunctive (OR)** causal rules and struggle with **conjunctive (AND)** rules when evidence is presented passively.

Can adults learn conjunctive causal rules when they are allowed to actively design their own experiments?

Research Questions

1. Does active exploration improve conjunctive causal learning?
2. Why does active exploration help?
3. Can LLMs benefit from active exploration in the same way?

Experimental Paradigm



Participants either actively selected interventions or observed matched intervention sequences before identifying the causal objects and rule.

Participants

- 306 adults (102 Active, 102 Passive Observation, 102 Passive Proposer)
- Conjunctive and disjunctive causal rules
- Six LLMs evaluated in the same interactive environment
- 32 possible causal hypotheses

Evaluation Metrics

- Object identification accuracy
- Rule inference accuracy
- Full hypothesis accuracy
- Search efficiency and information gain

Hypothesis Space

Each hypothesis consists of $h = (S, r)$ where

$S \subseteq \{1, 2, 3, 4\}, \qquad r \in \{\mathrm{AND}, \mathrm{OR}\}.$

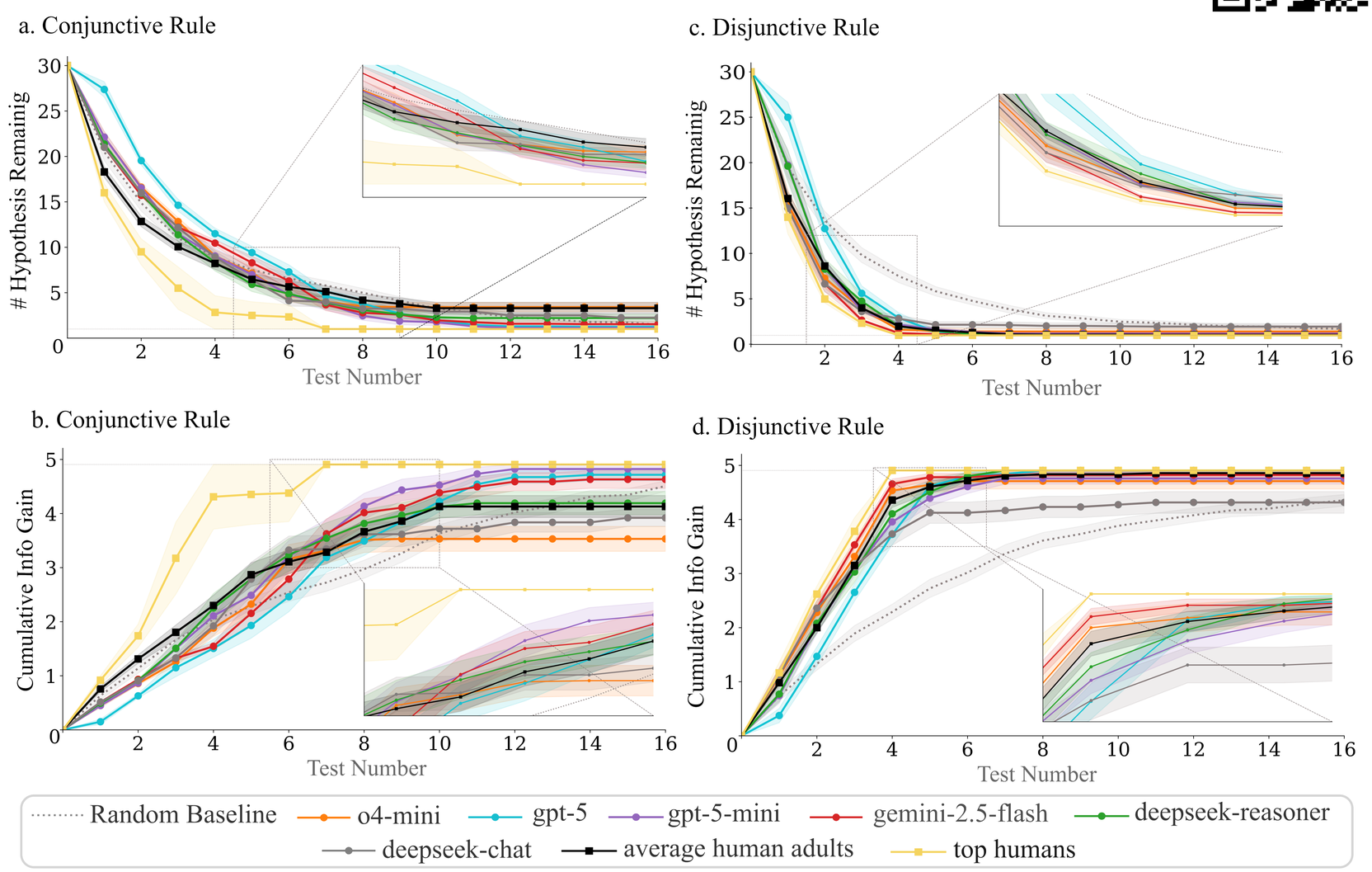
After each intervention,

$$\mathcal{H}_t = \{h \in \mathcal{H}_{t-1} : h(x_t) = y_t\}.$$

Why Are Conjunctive Rules Hard?

Disjunctive rules can often be identified from a single positive intervention. **Conjunctive rules** require eliminating several competing explanations, demanding longer and more informative intervention sequences.

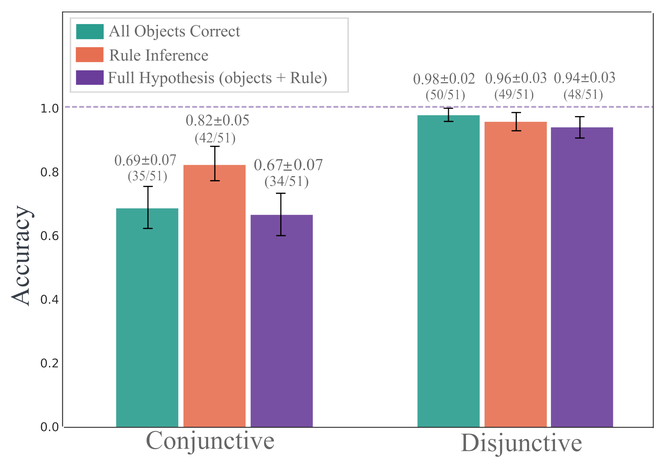
Search Efficiency During Active Exploration



Humans and LLMs progressively prune the hypothesis space. Disjunctive environments permit rapid pruning, whereas conjunctive environments require longer evidence accumulation. "Top humans" are participants achieving perfect full-hypothesis inference accuracy using the least number of tests.

#tests (success. trials)	Conjunctive	Disjunctive
Avg. human	11.24 ± .53	6.48 ± .45
Top human	8.50 ± .99	3.83 ± .17
Top LLM (gpt-5)	10.36 ± .40	7.67 ± .45

Active exploration helps



Passive Proposers generated interventions but did **not** observe outcomes from their own chosen tests. Their conjunctive object accuracy dropped to **0.12**, below Passive Observers (0.45) and Active Explorers (0.69).

We **quantify exploration efficiency** by the reduction in the number of hypotheses still consistent with the observed evidence:

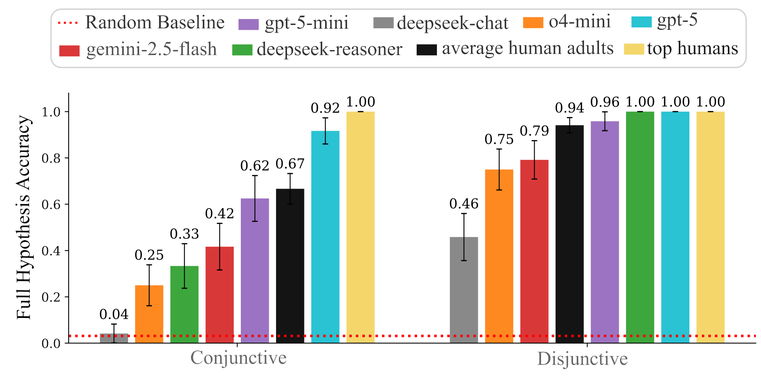
$$\mathrm{InfoGain}_t = \log_2 |\mathcal{H}_{t-1}| - \log_2 |\mathcal{H}_t|.$$

Cumulative information gain is:

$$\mathrm{CumInfoGain}_k = \sum_{t=1}^{k} \mathrm{InfoGain}_t$$

Agency matters as it lets learners collect the evidence their own hypotheses require.

Human vs. LLMs



Some models approach human performance in disjunctive settings, but conjunctive discovery remains harder and more variable.

- `gpt-5` and `deepseek-reasoner` are strong on disjunctive tasks.
- Top-performing humans remain most reliable on conjunctive tasks.
- Active intervention alone is insufficient; efficient hypothesis pruning matters.

Main conclusion

Adult failures in conjunctive causal reasoning are not fixed limits in causal competence. They emerge from **constrained evidence presentation**. When adults can act like scientists - selecting interventions and observing contingent outcomes - they recover flexible causal inference.

Successful reasoning requires: (1) agency (2) adaptive search (3) contingent feedback.

Nexiom Text Adventure: `https://nexiom-text-game.streamlit.app/`