BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

On the Eligibility of LLMs for Counterfactual Reasoning: A Decompositional Study

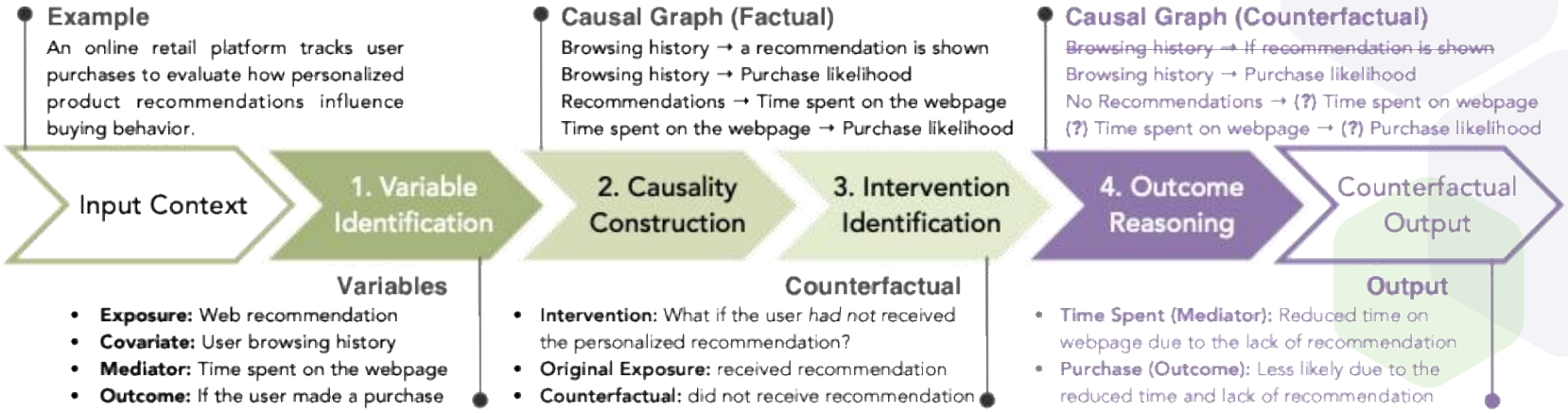
ICLR 2026

Shuai Yang[1] Qi Yang[2] Luoxi Tang[1] Yuqiao Meng[1] Nancy Guo[1]
Jeremy Blackburn[1] Zhaohan Xi[1]
[1]Binghamton University [2]Shanghai University
zxi1@binghamton.edu

Motivation

LLMs often struggle with counterfactual reasoning (i.e., the ability to adjust responses when presented with modified premises), but they lack **a causality-guided standardized** framework for systematically analyzing failure modes

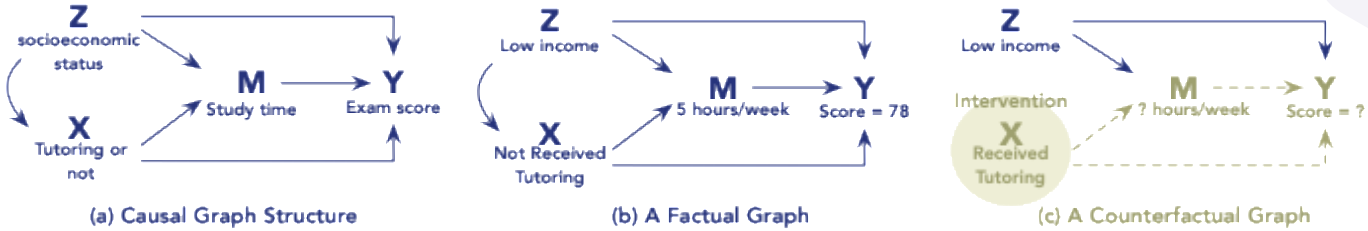
Key Idea



Decompose counterfactual reasoning with **four stages:**

- Identifying causal variables
- Constructing causal graph structures
- Identifying counterfactual intervention
- Counterfactual outcome reasoning

Causal graph to counterfactual



Counterfactual reasoning aims to answer the "**what-if**" question:

Given an observed instance $(X = x, Z = z, M = m, Y = y)$, what would the outcome Y be if the exposure X were set to a different value x', while keeping the covariate Z fixed?

Benchmark

Data	Use Case	Causal Variable				Counterfactual Condition	Modality	Num
		X	Z	M	Y			
CRASS (Frohberg & Binder, 2021)	Question answering	●	◐	◐	●	"What if ..." condition	Text	274
CLOMO (Huang et al., 2024)	Text logic parsing	●	●	●	●	New premise for textual statement	Text	1,100
RNN-Typology (Ravfogel et al., 2019)	Text syntax parsing	●	●	●	●	New syntactic structure of sentence	Text	584
CVQA-Bool (Zhang et al., 2024b)	Question answering	●	◐	◐	●	Hypothetical behavioral pattern	Text,Image	1,130
CVQA-Count (Zhang et al., 2024b)	Numerical reasoning	●	◐	◐	●	Hypothetical numerical pattern	Text,Image	2,011
COCO (Le et al., 2023)	Text-image matching	●	●	◐	●	"What if ..." condition	Text,Image	17,410
Arithmetic (Wu et al., 2024)	Mathematical reasoning	●	●	●	●	Change number base	Symbol	6,000
MalAlgoQA (Sonkar et al., 2024)	Question Answering	●	◐	◐	●	"What if ..." condition	Text,Symbol	807
HumanEval-Exe (Chen et al., 2021)	Code Execution simulation	●	◐	●	●	Hypothetical coding criterion	Text,Code	981
Open-Critic (Vezora, 2024)	Code generation	●	◐	●	●	Hypothetical descriptive functions	Text,Code	8,910
Code-Preference (Vezora, 2024)	Code summarization	●	◐	◐	●	Hypothetical code structures	Text,Code	9,389

●: present, ◐: partially present

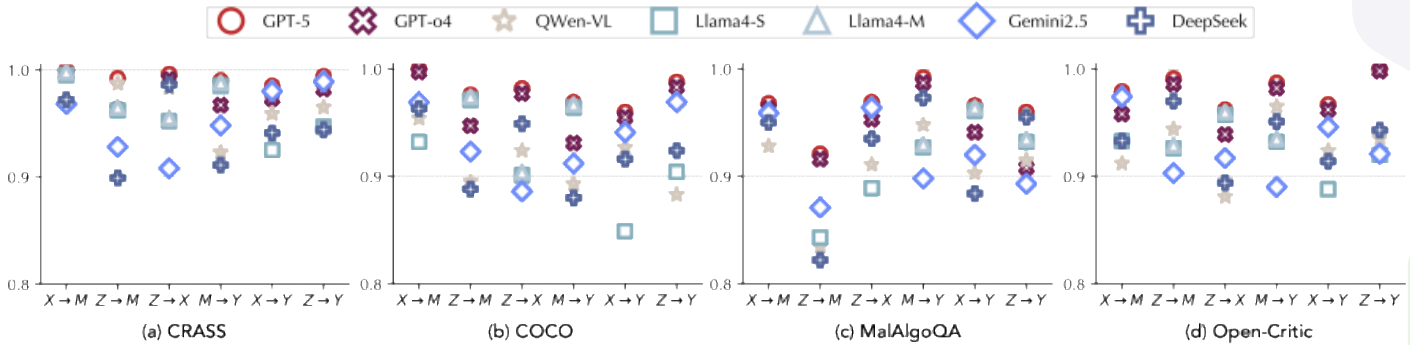
Curated 11 counterfactual datasets of ≈48k instances covering multiple modalities

Experimental setting

- **LLMs:** GPT-5, GPT-o4-mini-high, Qwen3-VL-235B-A22B-Thinking, Llama-4-Scout-17B, Llama-4-Maverick-17B-128E, Gemini2.5-Pro, DeepSeek-VL
- **Metrics:** F1 and accuracy

- ✓ RQ1: How well do LLMs perform when their counterfactual reasoning is decomposed into four distinct reasoning tasks?
- ✓ RQ2: What auxiliary techniques can improve LLMs' counterfactual reasoning?

Insights from counterfactual reasoning:



Performance of causal structure construction

- The primary challenge lies in identifying causal variables and performing causal reasoning
- In particular, the complex input modality and the implicit nature of mediation hinder effective reasoning through causal pathways

Insights from auxiliary techniques:

- (i) In general, improving implicit variable reasoning yields more substantial gains in end-to-end performance in contrast with the improvements in explicit variable identification.

- (ii) The combined strategy achieves the highest overall improvement, although the gains are not strictly additive.

Takeaways

- **Decompositional Framework:**
  We propose a decompositional strategy that spans from causal modeling to counterfactual reasoning

- **Benchmark Construction:**
  We construct a comprehensive evaluation benchmark by curating causal structures and counterfactual instances across multiple domains

- **Evaluation and Improvement Strategy:**
  We identify LLMs' capabilities in specific decompositional stage and propose actionable strategies to improve LLMs' counterfactual adaptability.

Thank You

For questions, please feel free to contact:

zxi1@binghamton.edu