My Idea on Developing a New Benchmark for Causal Inference in LLMs

Cheng Guo

Overview

- Who am I?
- Literature Review on Existed Benchmarks
  - Corr2Cause, CLadder, CEBaB
- My Idea Proposal

UC San Diego

Who am I?

- Cheng Guo (c5guo@ucsd.edu)
- 1st year MS - dedicated to pursue PhD
- Interested in Causality & NLP
    - Class Project on Backdoor attacks
    - Previous Research
- My Goal:
    - Benchmark -> Test on LLMs -> Fine-Tuning for better performance -> (?)

      Build a Causality-aware model Architecture or Encoding methods

Literature Review on Current Benchmarks

- Ladder of Causation (Pearl & Mackenzie, 2018)
  - Correlation, Intervention, Counterfactuals

  - TimeTravel (Qin et al., 2019)
    - Tuebingen Cause-Effect Pairs (Mooij et al., 2015)
      - IntuitivePhysics (Zečević et al., 2023)
        - BIG-bench (Srivastava et al., 2023)
          - e-CARE (Gao et al., 2023)
            - LogiQA (Liu et al., 2020)
              - LOGIC (Jin et al., 2022)

UC San Diego

Existed Benchmark - Corr2Cause (Jin et al., 2023)

input string · *lengths* 312↔1.21k 36.4%	**label** int64 0 81.5%	**num_variables** int64 4 0.3%	**template** string · *classes* child 16.7%
Premise: Suppose there is a closed system of 4 variables, A, B, C and D. All the statistical relations among these 4 variables are as follows: A correlates with C. A correlates with D. B correlates with C. B correlates with D. C correlates with D. However, A is independent of B. A and D are independent given B and C. A and D are independent given C. B and D are independent given A and C. B and D are independent given C. Hypothesis: C directly causes B.	0	4	child

https://arxiv.org/abs/2306.05836

UC San Diego

Existed Benchmark - CLadder (Jin et al., 2024, proposed CausalCoT)

<table>
<tr><th>id
int64</th><th>prompt
string</th><th>label
string</th><th>reasoning
string</th><th>rung
int64</th><th>query_type
string</th><th>graph_id
string</th><th>story_id
string</th><th>question_property
string</th><th>formal_form
string</th></tr>
<tr><td>4</td><td>Imagine a self-contained, hypothetical world with only the following conditions, and without any unmentioned factors or causal relationships: Husband has a direct effect on wife and alarm clock. Wife has a direct effect on alarm clock. For husbands that don't set the alarm and wives that don't set the alarm, the probability of ringing alarm is 8%. For husbands that don't set the alarm and wives that set the alarm, the probability of ringing alarm is 54%. For husbands that set the alarm and wives that don't set the alarm, the probability of ringing alarm is 41%. For husbands that set the alarm and wives that set the alarm, the probability of ringing alarm is 86%. For husbands that don't set the alarm, the probability of alarm set by wife is 74%. For husbands that set the alarm, the probability of alarm set by wife is 24%.</td><td>yes</td><td>Let X = husband; V2 = wife; Y = alarm clock. X->V2,X->Y,V2->Y E[Y_{X=1, V2=0} - Y_{X=0, V2=0}] \sum_{V2=v} P(V2=v|X=0)*[P(Y=1|X=1,V2=v) - P(Y=1|X=0, V2=v)] P(Y=1 | X=0, V2=0) = 0.08 P(Y=1 | X=0, V2=1) = 0.54 P(Y=1 | X=1, V2=0) = 0.41 P(Y=1 | X=1, V2=1) = 0.86 P(V2=1 | X=0) = 0.74 P(V2=1 | X=1) = 0.24 0.74 * (0.86 - 0.41) + 0.24 * (0.54 - 0.08) = 0.32 0.32 > 0</td><td>3</td><td>nde</td><td>mediation</td><td>alarm</td><td>easy</td><td>E[Y_{X=1, V2=0} - Y_{X=0, V2=0}]</td></tr>
</table>

https://arxiv.org/abs/2312.04350

UC San Diego

Existed Benchmark - CEBaB (Abraham et al., 2022)

			food	ambiance	service	noise	overall
Original text:		Excellent lobster and decor, but rude waiter.	+	+	–	unk	4
Edit Goal							
food:	–	Terrible lobster, excellent decor, but rude waiter.	–	+	–	unk	2
food:	unk	Excellent decor, but rude waiter.	unk	+	–	unk	3
ambiance:	–	Excellent lobster, but lousy decor and rude waiter.	+	–	–	unk	3
ambiance:	unk	Excellent lobster, but rude waiter.	+	unk	–	unk	3
service:	+	Excellent lobster and decor, and friendly waiter.	+	+	+	unk	5
service:	unk	Excellent lobster and decor.	+	+	unk	unk	5
noise:	+	Excellent lobster, decor, and music, but rude waiter.	+	+	–	+	4
noise:	–	Excellent lobster and decor, but rude waiter, and noisy.	+	+	–	–	3

https://arxiv.org/abs/2205.14140

My Idea Proposal - Issues to Address

- No Causal Parroting
  - No Exploiting Language Cues
    - Focusing on Interventions & Counterfactuals
      - Scaling to Multiple Factors
        - Open-Endedness
          - Retrieving World Knowledge?

UC San Diego

My Idea Proposal - Proposed Contents of the Benchmark

- Fictional Scenario
  - Hidden Confounder, Collider, or Mediator
    - Open-ended Questions
      - Regarding Interventions & Counterfactuals
        - Structural Understanding

UC San Diego

My Idea Proposal - Evaluation

- Alignment
- Quality
- Robustness
- Fairness
- Efficiency

Dr. Reid Pryzant (Stanford, Google):

- It would be great if there were a real dataset of paired observational data + RCT for the same problem using text as the independent variable so that researchers can better study the causal effect of text e.g. adding/removing words.

https://arxiv.org/abs/2311.04287

My Idea Proposal - After Developing the Benchmark

- CausalCoT (Jin et al., 2024) in Prompting
  - Other Prompt Engineering Techniques
    - Active Learning
      - Teacher Forcing
        - Masked Autoencoder
          - Mixture of Experts

UC San Diego

Thank you for listening!

References:

- Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023.
- Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M.
- Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond, 2022.
- Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. Is chatgpt a good causal reasoner? A comprehensive evaluation, 2023.
- Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models, 2024.
- Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. Logical fallacy detection, 2022.
- Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation?, 2023.
- Jiaxuan Li, Lang Yu, and Allyson Ettinger. Counterfactual reasoning: Do language models need world knowledge for causal understanding?, 2022.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning, 2020.
- Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks, 2015.
- Judea Pearl and Dana Mackenzie. The Book of Why: The New Science of Cause and Effect. Basic Books, Inc., USA, 1st edition, 2018.
- Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation, 2019.
- Linying Yang, Oscar Clivio, Vik Shirvaikar, and Fabian Falck. A critical review of causal inference benchmarks for large language models. In AAAI 2024 Workshop on ”Are Large Language Models Simply Causal Parrots?”, 2023.
- Matej Zecevic, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal, 2023.
- Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. Understanding causality with large language models: Feasibility and opportunities, 2023.

UC San Diego