My Idea on Developing a New Benchmark for Causal Inference in LLMs: An Informal Write-Up 1

Cheng Guo
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093
c5guo@ucsd.edu

1 A Brief Introduction of Myself and My Idea

My name is Cheng Guo and I am a first-year Master's student studying Computer Science at the University of California, San Diego. I am dedicated to Causality and LLM research and have a strong desire to pursue a PhD degree. I am interested in causal reasoning because it is an essential skill for us. Recently in a deep learning class, my final project was about detecting backdoor attacks, whereby inserting certain words in text data, the classification model trained on those data will always output certain labels when it recognizes an inserted word, which I think is a consequence of LLMs is still short on causal reasoning. Previously, I have done some research towards a human-centered perspective of NLP. In one of the research I did before, I focused on the non-textual characteristics (like capitalization, the usage of exclamation marks, and keyboard smashing) for sentiment analysis, which can be found here (in the "Publications" tab): `https://chengguo2000.github.io/`. Outside of academia, I enjoy theatre arts, and in some dramas, the playwrights use causal inference to make the plotline more surprising, where an example could be black comedy movies.

Based on what I have found, I think that LLM research would be more comprehensive if focused more on causal reasoning. I want to know what is a way of fine-tuning the existing LLMs (or building new ones) that could improve its understanding of causality. My end goal would be something like a Causality-Aware Model Architecture of LLM that uses Causality-Aware Encoding Methods (or incorporates causality into an existing LLM architecture) and can understand causal relationships in natural language. But before that, I need to define what a good understanding of causality for LLM is, so I need to have some datasets, some tasks, and some metrics and test them on the existing LLMs to see where should I improve and see if it is necessary to build a new architecture or start with an existing one. That is a research direction I would like to pursue and here are what I discovered from the literature review and my proposed ideas for building a new benchmark.

2 Literature Review

Below I am presenting some of my findings when I am reading papers about causal inference in NLP and LLMs. Since I am proposing an idea to build a new benchmark, I focused on already existing benchmarks. Causal Inference deserves research attention since it enhances reasoning skills in LLMs. Its relationship with NLP can be seen in two directions. We can use NLP methods to estimate the causal effects of text, and we can also make LLMs predict better with causality [4]. This emerging field is referred to as CausalNLP and various research has been done for incorporating causality in LLMs. The guiding theory in Causality is the ladder of causation proposed by [13], and [18] has proposed a three-level hierarchy that can be seen as the ladder of causality for LLMs.

- The first Rung identifies correlations from data, while the first level identifies causality from domain knowledge. Both can be seen as retrieving relationships from given information.

- The second Rung is about interventions, and the second level is about discovering new knowledge. Both can be seen as reasoning on given scenarios.

- The third Rung is about counterfactuals, and the third level is about estimating the consequences of actions. Both can be seen as reasoning about imaginary scenarios.

The first stage has been studied and there are well-defined benchmarks. However, the second and third stages in both theories have not been studied thoroughly and are short in benchmarks. I think developing a benchmark for the second and third stages is critical because it makes causality research different from other statistical approaches. For the rest of this review, I shall go over some existing datasets and benchmarks.

To begin with, some datasets and benchmarks provided either binary options or labels regarding causality. The **BIG-bench** dataset is a collective work regarding various reasoning capabilities for LLMs, and the data for the "empirical judgment" task and "cause and effect" task belongs to this category [3]. The "empirical judgment" task data contains 99 sentences and each has three options: causal, correlative, or neutral. One of these three options will be labeled "1" and the other two will be "0". For the "cause and effect" task, the "two sentence" subtask is designed for causal reasoning, where two events are listed for each example and the answer is one of the two that is the cause of the other. The first subtask from the **e-CARE** dataset [5] also falls into this category, where given a premise and two hypotheses, the binary label represents which hypothesis is the correct effect of the premise. The **LogiQA** dataset is similar to the **e-CARE** datasets above [11], but this time, there are four choices for each scenario and each scenario is specified into context and question. The **Tuebingen cause-effect pairs** dataset [12] contains 74 pairs of variables and the model needs to decide whether the change in one variable causes the change in the other and the result is binary. This dataset specifically requires the model to retrieve world knowledge to perform causal reasoning. For world knowledge, The counterfactual reasoning dataset evaluates whether the model needs world knowledge [9] with counterfactual and real-world premises.

Some benchmarks dig further into the causal relationships between variables. The **Corr2Cause** benchmark [8] contains the premise and hypothesis. Beyond determining whether the hypothesis is true or false, the causal relationship between the two variables in the hypothesis also falls into the following six classes: Is-Parent, Is-Ancestor, Is-Child, Is-Descendent, Has-Collider, Has-Confounder, and the corresponding class is provided in the dataset. The **Logic** and **LogicClimate** datasets [7] investigate the logical fallacy in a masked statement (or in an article) and each data point is provided with one of the 13 logical fallacy types, examples include faulty generalization and ad hominem. The **CLadder** dataset [6] explores the quantitative aspect of causal reasoning where marginal and conditional probabilities are provided in each prompt for determining causality between two events. Researchers who built the **CLadder** dataset also designed **CausalCoT**, a Causal chain of thought prompting strategy that significantly improves the performance in LLMs.

Among datasets and benchmarks that contain binary options or labels, it is easy to understand and interpret performance results. However, it turns causal reasoning into a classification task, where the model only needs to output a label. Beyond retrieving world knowledge, the model may perform causal parroting or exploiting language cues, and both of them need to be avoided since they cause the model to infer results not with causal reasoning.

Some datasets and benchmarks contain open-ended prompts. The second subtask from the **e-CARE** dataset is about generating a conceptual explanation for a cause-effect pair [5]. The **Intuitive Physics** dataset [17] focuses on the real-world cause-and-effect in physics. The **TimeTravel** dataset [14] is about whether a model can modify the story to make it compatible with a counterfactual event. For these datasets, since it would be hard for the model to output exact results like the one provided, they used automated metrics like **BLEU-4** or **ROGUE-L** score to evaluate performance.

After reading all the above research papers, I found that most of them concluded that the current LLMs performed poorly with the specified task, while some of them proposed improvement strategies like **CausalCoT** [6]. In my opinion, I think one reason could be that most current LLMs are next-word predictors, while they are poor at reasoning and comprehension. Hindrance like causal parroting, exploiting language cues, or retrieving outside knowledge may help with reasoning to some extent, but they did not embody the inherent causal reasoning skills of LLMs.

3 My Benchmark Idea Proposal

3.1 Issues to Address

After reading the above papers, I believe that a new benchmark should be developed to address the following issues:

- **No Causal Parroting:** To not let models recite causal relationships from input data, I think we have to make the causal relationships described in the prompt much different than the expected causal relationship we want the model to learn and conclude.
- **No Exploiting Language Cues:** We can use logic symbols like arrows in the prompt so that it will not exploit language cues when learning causal relationships. We may need to develop new ways of verbalizing a causal relationship expressed with math symbols or causal graphs in natural language to avoid exploitation.
- **Focusing on Interventions & Counterfactuals:** Previous benchmarks focus a lot on identifying causation from correlation [8, 16], while we need a benchmark for interventions and counterfactuals. Instead of asking a model to identify the causal relationship, we can ask for what experiment should we perform or what data we need to reason causality (an example can be found below).
- **Scaling to Multiple factors:** We should include scenarios and prompts with various complexities to assess if a model can understand complex causal relationships. This would require structural understanding discussed below.
- **Open-Endedness:** It is only through open-ended questions can we evaluate whether a model truly understands causal relationships, but we need the model to output its understanding in a certain format so that we can assess its performance.
- **Regarding Retrieval Outside Knowledge:** I think retrieval would make the model not focus on causal reasoning but just augment its answer based on outside sources. I would choose not to encourage the model to retrieve outside knowledge when learning about causality (but I might be wrong about this).

3.2 Proposed Contents of the Benchmark

Based on the above issues to address in the new benchmark, I propose that the new benchmark should contain the following features:

- **Fictional Scenario:** Retrieval instead of understanding and exploiting language cues are prominent issues for models failing to reason about causality [16, 17]. It would be important for us to formulate scenarios and prompts in a way so that they cannot be learned through retrieval or exploiting language cues (masking from [7] is a good idea). Also, the benchmark could be centered around scenarios where we can have multiple questions for one scenario.
- **Hidden Confounder, Collider, or Mediator:** While the model needs to discover hidden factors since we want to avoid them to learn through retrieving [8], we can include some hidden factors in the prompt while not mentioning its causal relationships.
- **Open-ended Questions regarding Interventions and Counterfactuals:** We need open-ended questions on interventions and counterfactuals for the model to generate answers after deep understanding so that it does not become a classification model that relies on language cues. We might need to add some sample answers into the prompt for the model to learn how to answer questions like those (but I am wondering whether it will exacerbate the causal parroting issue).
- **Structural Understanding:** As the complexity of causal reasoning increases, the benchmark needs to understand the causal relationships between various events in a scenario in a structural format. The benchmark should include information on causal relationships like parent and child or ancestor and descendent between events (like a causal graph) as a reference (maybe not provided in the prompt).

Other than the above contents, I think the overall benchmark should include the following aspects for evaluation of performance:

- **Alignment:** Is the causal relationship output from the model correct?
- **Quality:** How comprehensive did the model's output address causal relationships? Did it mention all causal relationships in the given scenario?
- **Robustness:** What would happen to the output if we add invariant perturbations to the given scenario? Can the model still output the same answer?
- **Fairness:** Are there any biases or disparities when facing the various demographic information in the provided scenario?
- **Efficiency:** How long does it take for the model to infer causality in a given scenario concerning complexity?

3.3 After developing the Benchmark

As mentioned above, after developing the benchmark, I would like to research how to make LLMs better perform causal reasoning, which can be achieved either through fine-tuning or through developing new causality-aware encoding methods and model architecture. Here are some ideas I have come across when I read other papers in the field of NLP that I think might contribute to further research.

- **Causal Chain-of-Thought in Prompting:** We would want the LLMs to adapt to causality with multiple factors and higher complexity [8, 16]. The CausalCoT from [6] is an adaptation of CoT in a causality context, and I would like to explore how it can be used in formulating a prompt like a few-shot prompting.
- **Prompt Engineering:** Beyond Chain-of-Thought, it would also be interesting to look at other strategies for prompting, including In-Context Learning and One-Shot Prompting, where we provide a similar causal relationship in the fictional scenario as an example in the prompt for the model to find another one.
- **Active Learning:** Since causality deeply relies on human understanding, it would be important to incorporate some human feedback during training. Active learning would be a good method to add human annotations regarding causality which is proven to improve the efficiency of fine-tuning [2]. If human labeling is not efficient for large data, we could also make LLMs labeling results from each other.
- **Teacher Forcing:** Let's say there is a chain of causal relationships in a fictional scenario (where one caused another, then caused another, and so on...), I am wondering if we can improve model efficiency for convergence if we feed each event in the causal chain back to the model. This is only an analogy from the teacher-forcing strategy used in RNN.
- **Masked Autoencoder:** To make models reasoning about causality, I think we can also take unsupervised approaches. This idea is also an analogy from the masked autoencoder method in computer vision. In causality, for example, on a fictional scenario, we can remove some words that determine a causality relationship and let the model reconstruct the whole scenario by itself, I am wondering if this could improve the understanding of causality.
- **Mixture of Experts:** Although I mentioned that the benchmark should focus on causal reasoning instead of knowledge retrieval, we cannot neglect the fact that most humans use outside knowledge to perform causal reasoning. To develop a causality-aware model, it needs to learn how to retrieve knowledge from various areas, and I think a Mixture of Experts, a prevailing machine-learning technique, can be useful in this situation.
- [1, 15, 10].

References

[1] Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior, 2022.

[2] Gantavya Bhatt, Yifang Chen, Arnav M. Das, Jifan Zhang, Sang T. Truong, Stephen Mussmann, Yinglun Zhu, Jeffrey Bilmes, Simon S. Du, Kevin Jamieson, Jordan T. Ash, and Robert D. Nowak. An experimental design framework for label-efficient supervised finetuning of large language models, 2024.

[3] Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023.

[4] 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.

[5] Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. Is chatgpt a good causal reasoner? a comprehensive evaluation, 2023.

[6] 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.

[7] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. Logical fallacy detection, 2022.

[8] 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.

[9] Jiaxuan Li, Lang Yu, and Allyson Ettinger. Counterfactual reasoning: Do language models need world knowledge for causal understanding?, 2022.

[10] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[11] 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.

[12] Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks, 2015.

[13] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018.

[14] Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation, 2019.

[15] Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning interpretable concepts: Unifying causal representation learning and foundation models, 2024.

[16] 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.

[17] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal, 2023.

[18] 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.