Developing Benchmark for Causal Representation Learning in LLMs: An Informal Write-Up 2

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

1 Introduction

After conducting a primitive literature review on causality and LLMs, I believe that further research should focus beyond inferring explicit causal relationships, but rather on the implicit causality of LLMs, also known as causal representation learning, where we learn what makes a model prediction in a certain way. As stated in [10], there are two approaches in LLMs. The first one is building interpretable models, where each node in the network represents a concept of a text, so we can interpret how the output is generated based on model parameters. However, it requires much more computational resources [10], so researchers take the second approach, by starting with a foundational model, and then understanding how it understands causal concepts through counterfactual examples. The process of identifying high-level causal variables is causal representation learning. In the era where foundational LLMs like ChatGPT, Gemini, Llama, and Claude are developed and commercialized, I believe that the second approach will be the trend for future researchers. While causal representation learning is short in benchmarks, so in this write-up, I compiled some findings in the literature on causal representation learning and developed two ideas for building benchmarks.

2 Literature Review

In the previous write-up, I reviewed benchmark datasets that explicitly state causal relationships either within prompts or between prompts and desired responses. One example is the **Multi-Genre NLI Corpus** [13], where there are two sentences, one premise, and one hypothesis, and the label refers to the relationship of the two sentences, entailment, contradiction, or neutral. Those tasks require a deep understanding of the languages, but it is still hard to assess how the model learns and how the model makes certain predictions. To do that, I propose to state causal relationships implicitly and explore causality with interventions and counterfactual manipulations, thus, understanding how the model works causally through performing causal representation learning. I can list some questions to understand what my idea is: What will a review (of a book, a movie, a restaurant) look like if the reviewer has a different opinion than the one entailed in the current review? What will a news article on certain issues look like if the author has an opposite political ideology than the one entailed in the current article? As stated in previous research [1, 2], benchmarks on model explanations are still rare.

The only existing benchmark that focuses on text interventions for causal representation learning is **CEBaB** [1]. Based on 2299 restaurant reviews, researchers intervened on 4 concepts (food, ambiance, service, noise) to create 15089 text pieces. To give a conceptual example, say the original review is "good food and good service", it shows a positive sentiment on food and service, while the sentiment on ambiance and noise is unknown. The researchers intervened in the review by adding or removing words, like "bad food and good service" which changed the sentiment on food to negative while the other three concepts held the same sentiment, or "good food, good service, but bad decor" added a negative sentiment to the ambiance concept. Then, all 15089 text pieces are labeled on a 5-star scale through crowd-sourcing after the sentiments on 4 concepts are validated in the edited text. Based on

this dataset, we can learn how the model predicts the label for a new text based on the 4 concepts. If reviews with positive sentiments on food tend to get 4 or 5 stars and reviews with negative sentiments on food tend to get 1 or 2 stars, then it means that sentiment about food in the review matters a lot in predicting the final label of this review. The **CEBaB** benchmark inspired me to create the first new benchmark ideas described below in §3.2.

Now I shall briefly review some common measures for causal-representation learning. Researchers have developed ways of understanding how each concept is represented through all the nodes in the model. The most commonly used measures are the ones below. I learned them from the **CEBaB** benchmark paper [1], and I read the papers that proposed them for more details.

- **Average Treatment Effect (ATE):** It measures the difference between the mean of treatment group outcomes and control group outcomes. The empirical version is $\widehat{ATE}$ [12].
$$ATE_T(Y, \mathcal{G}) = E_{\mathcal{G}}[Y|do(T = 1)] - E_{\mathcal{G}}[Y|do(T = 0)] \tag{1}$$
- **Causal Concept Effect (CaCE):** Developed from ATE, this variable is the difference between the expected output between groups whether a binary concept C_0 exists or not. To measure it empirically, the researchers used a variational autoencoder (VAE) [4]. It also has an empirical version $\widehat{CaCE}$.
$$CaCE(C_0, f) = E_{\mathcal{G}}[f(I)|do(C_0 = 1)] - E_{\mathcal{G}}[f(I)|do(C_0 = 0)] \tag{2}$$
- **Individual Causal Concept Effect (ICaCE):** The individual version measures the effect of a concept change from a specific value to a specific value in the creation of input data on a neural network. The empirical version is $\widehat{ICaCE}$ [1].
- **Absolute Causal Concept Effect (ACaCE):** The absolute version aggregates over all possible interventions and focuses on a concept as a whole instead of individual values of the same concept. The empirical version is $\widehat{ACaCE}$ [1].

Below are some explanation methods for causal-representation learning where I also learned from the **CEBaB** benchmark paper [1], and I read the papers that proposed them for more details.

- **ConExp:** a non-causal baseline with no intervention, which is similar to **CaCE** except the do-operators [1, 4].
$$\text{ConExp}(C, f) = E_{\mathcal{G}}[f(I)|C = 1] - E_{\mathcal{G}}[f(I)|C = 0] \tag{3}$$
- **Conditional Expectation Learner (S-Learner):** It measures the conditional ATE through training a logistic regression model to predict the probability using the values of all labeled concepts [1, 6].
- **Testing with Concept Activation Vectors (TCAV):** It calculates a model's sensitivity to a concept across an entire class of inputs with directional derivatives [5] and the researchers adapted it to measure the sensitivity of each output class to changes [1].
- **ConceptSHAP:** Shapley Additive Explanations (SHAP) assigns each concept regarding its importance in a single prediction based on Shapley values [8]. ConceptSHAP expands SHAP in calculating the concept's contributions given a complete set of m concepts such that the test accuracy is higher than a threshold [1, 15].
- **Iterative Nullspace Projection (INLP):** It estimates intervention effects when a particular concept is unknown by training liner classifiers after removing a concept from a representation vector [1, 11].
- **CausalLM:** It estimates a binary concept's effect on the model prediction through learning counterfactual representation with adversarial tasks. The output is the **text representation-based individual treatment effect (TReITE)** [1, 3].

3 My Benchmark Ideas

3.1 Thoughts on developing benchmarks

Before digging into the two specific ideas I developed for causal representation learning in LLMs, I start by listing out some thoughts I have regarding how the benchmarks should be developed. Some

of them come from my discussions with various researchers in the field. I have implemented the following thoughts in developing the two benchmarks.

- **Focusing on Intervention & Counterfactuals:** As stated in my first write-up, future benchmark development should focus on intervention and counterfactuals rather than simple correlation and causation from the first ladder of causality [9].
- **Beyond Sentiment Analysis:** The **CEBaB** benchmark [1] focuses on identifying sentiments from counterfactual restaurant reviews, and we need a benchmark beyond that. Sentiment analysis is already easy enough for LLMs to perform in high accuracy under zero-shot scenarios and there is not much focus on that in recent advances of LLMs, so I think the new benchmark should focus beyond sentiment analysis tasks.
- **Beyond Syntactic Intervention:** Syntactic Intervention could be useful for digging into how each concept is learned in the neural network, but for the current step, I think it would be better if we use counterfactual rewritings so that it is more different from the original text and the effect of concept learning can be more apparent.
- **Hard for LLMs in Zero-Shot settings:** As stated above, we need the LLMs to complete a task that is hard in Zero-Shot settings, so there is room for research through fine-tuning to improve results. In both of my ideas proposed, I noticed that the LLM I used resulted in low accuracy based on the examples I prompted.
- **Less Work for Crowdsourcing:** I noticed that there is a lot of work is done through crowd workers in developing the **CEBaB** benchmark [1]. However, I think as LLMs improve in scale rapidly, we should have less human work involved because not only does it require much money to pay those crowd workers, but as we are building benchmarks that are big in scale, human work would be hard to generalize.
- **Open-endedness?** I mentioned this in my first write-up, but now I am unsure about whether the new benchmark should have classification labels or open-ended answers. I would say that we can start with a benchmark with labels because it is easy to evaluate and interpret results. My first idea is with classification labels, and my second idea is with open-ended answers. However, it would be easy to interchange after we have a fixed idea.
- **New Measurement or Explanation Method?** I think since in the new benchmark, the data we developed might be based on a concept causal graph that has a complex structure, we can also develop a new measurement that focuses on the concept's effect given its relative position on the causal graph or update an existed measurement scaling to multiple factors.

3.2 My First Idea - Left/Right Wing News Article

As stated in [14] and my first write-up, future benchmarks should focus on interventions and counterfactuals. It is also worth noting that causal representation learning research is valuable to social scientists because they can explore the implicit patterns in texts that are useful for studying human behavior. Based on the **CEBaB** benchmark [1], I have this first idea focusing on identifying whether a news article comes from left or right-wing media based on its topic, focus, ideology, and style.

To build an example for my idea, I begin by finding news reports on the same issue from left/right-wing media. The example I have is regarding President Biden and President Xi's Phone Call on April 2, 2024. I found the report from both CNN and FOX. CNN is considered to be a left-wing media while FOX is right-wing. I took the first several paragraphs from both articles and put them into LLM asking it to rewrite them as if they come from the opposite-wing media. Now I have 4 excerpts on the same issue, I asked a LLM to identify whether they come from left or right-wing media. It performed badly by identifying the FOX excerpt and the rewritten CNN excerpt as left-leaning, where both should be right-leaning. I further required the LLM to identify topics, focuses, and writing styles in all 4 excerpts to identify causal learning concepts. Details about what the LLM has generated can be found in Appendix A. I have also provided a tentative causal graph in Appendix A.

Based on the example, I have developed a framework for producing the benchmark. Starting with a list of news topics, I found articles from both CNN and FOX (or other media if one lacks an article on the topic), then I shall ask LLMs to rewrite the two articles as if they come from the opposite wing. It would require crowd workers to label all the articles regarding their topic, focus point, ideology,

and writing style, and then let LLMs label those again to compare any differences. Then in future research, I can use human output as a target and fine-tune the LLMs so that they will predict similarly to the target based on the concepts of topic, focus, ideology, and style. I have found the following resources on Hugging Face that might be helpful for the development of this benchmark.

- CNN News dataset: https://huggingface.co/datasets/AyoubChLin/CNN_News_Articles_2011-2022 (2117 in train, 344 in test on "Politics" category)
- Fox News generator: https://huggingface.co/gagan3012/Fox-News-Generator

Now I will analyze the pros and cons of this first idea. It has fewer crowd workers involved than **CEBaB** since the rewriting process can be done through LLMs. Now I listed 4 aspects, including topic, focus, ideology, and style, but in the future, we can focus on more aspects as concepts for causal learning. It also unifies the creation process, where I can utilize LLMs to produce the rewritten pieces and labels. However, it focuses only on classification with labels rather than open-ended questions, and it defines news articles in a binary way (left and right-wing), where in reality, not every article can be classified strictly into a certain wing.

3.3 My Second Idea - Open-ended Questions

My second idea originates around open-ended questions and syntactic intervention. I want to know what LLMs respond if the question asked is not logically consistent or unrealistic after changing a word from the original question in the **TruthfulQA** dataset [7]. Take the first question as an example (what happens to you if you eat watermelon seeds?) I intervened by changing one word and retrieved the following results.

- **What happens to you if you eat watermelon seeds?** Eating watermelon seeds is generally safe for most people and can even be beneficial as they are a good source of nutrients like protein, fiber, and healthy fats. However, consuming a large quantity of watermelon seeds may lead to digestive issues such as stomach discomfort or blockages. It is recommended to chew the seeds thoroughly before swallowing to aid in digestion.
- **What happens to you if he eats watermelon seeds?** If someone else, referred to as "he," eats watermelon seeds, the potential effects would be similar to those mentioned above. The person may experience digestive discomfort if a large amount of seeds is consumed without proper chewing.
- **What happens to him if you eat watermelon seeds?** If you eat watermelon seeds, the effects would be personal to you, including the potential benefits and risks associated with consuming the seeds, as mentioned in the first point.
- **What happens to him if he eats watermelon seeds?** If he, referring to another person, eats watermelon seeds, the effects would be specific to him, including the potential digestive consequences of consuming the seeds in large quantities.
- **What happens to you if you drink watermelon seeds?** Drinking watermelon seeds, especially in large quantities, may not be recommended as they can be difficult to digest in their whole form. It is generally advised to chew watermelon seeds before swallowing them to avoid any digestive issues.
- **What happens to you if you eat shovel seeds?** There is no specific information available about "shovel seeds" as they are not a common or recognized type of seed for consumption. It is important to be cautious and avoid consuming any seeds or plants that are not intended for human consumption, as they may be toxic or harmful.
- **What happens to you if you eat watermelon blades?** Watermelon blades are not a common term associated with the fruit. If you are referring to a part of the watermelon plant or fruit that is not typically consumed, it is advisable to avoid eating it as it may not be safe for consumption.

To generalize the process, I can create a benchmark based on the 817 open-ended questions from the **TruthfulQA** dataset [7] by using syntactic intervention to change one word (adjectives, conjunctions, pronouns...), which LLMs can perform. Then, I shall ask LLMs and crowd workers to answer all those questions (there should be around 6000 questions with one-sentence answers). For crowd

workers, I shall ask them to pay specific attention to the words in the question so that they will just respond "faulty question" to questions that are not understandable to humans.

To analyze the pros and cons of this second idea, I think it shares similar pros with the first one. It has fewer crowd workers involved than **CEBaB** since the rewriting process of the questions can be done through LLMs. It is open-ended so it opens possibilities for humans and LLMs to learn and respond. Similarly, we can inspect various syntactic interventions, like changing adjectives, conjunctions, or pronouns. However, it is hard to quantify results or develop causal explanations since all the answers are open-ended.

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] 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.

[3] Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. CausaLM: Causal Model Explanation Through Counterfactual Language Models. *Computational Linguistics*, 47(2):333–386, 07 2021.

[4] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace), 2020.

[5] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2018.

[6] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, February 2019.

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

[8] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

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

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

[11] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection, 2020.

[12] Donald Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 10 1974.

[13] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018.

[14] 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.

[15] Chih-Kuan Yeh, Been Kim, Sercan O. Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks, 2022.

A Left/Right Wing News Article

A.1 CNN - Original

This original excerpt comes from https://www.cnn.com/2024/04/02/politics/biden-xi-speak-first-time-november/index.html and the analysis comes from https://www.chatpdf.com/.

President Joe Biden spoke on the phone with Chinese President Xi Jinping on Tuesday, marking the first conversation between the leaders since their historic in-person summit in November and the latest in ongoing efforts by US and Chinese officials to defuse tensions between the two superpowers. The call comes amid heavy global turbulence – the ongoing wars in Gaza and Ukraine, as well as North Korea's nuclear capabilities, were topics of discussion. Other issues that have strained the Washington-Beijing relationship also came up, including Taiwan, China's recent provocations in the South China Sea and Beijing's human rights abuses.

- **Topic:** Phone call between President Biden and Chinese President Xi Jinping, global tensions, diplomatic efforts.
- **Focus:** Ongoing efforts to defuse tensions between US and China, discussion topics including conflicts in Gaza and Ukraine, North Korea's nuclear capabilities, Taiwan, South China Sea, human rights abuses.
- **Ideology:** Neutral, diplomatic engagement without clear bias.
- **Style:** Informative, factual reporting on diplomatic dialogue.

A.2 FOX - Original

This original excerpt comes from https://www.foxnews.com/politics/biden-chinas-xi-hold-phone-call-taiwan-ai-trade and the analysis comes from https://www.chatpdf.com/.

President Biden held a phone call with Chinese President Xi Jinping for the first time since July 2022 on Tuesday, the White House announced. The White House has yet to offer details about the call, but Biden and Xi were expected to discuss Taiwan, narcotics, artificial intelligence, and China's support for Russia against Ukraine. The White House described the conversation as merely a "check-in" between the two leaders. Biden was not expected to bring up TikTok or the origins of COVID-19. Biden currently supports legislation in Congress that would ban TikTok unless it is sold to a company that is not beholden to the Chinese Communist Party.

- **Topic:** Phone call between President Biden and Chinese President Xi Jinping, discussion topics, White House's characterization.
- **Focus:** Expected discussion topics like Taiwan, narcotics, Chinese support for Russia, White House's description of the call as a "check-in," exclusion of TikTok and COVID-19 origins.
- **Ideology:** Slightly left-leaning, critical of White House's handling of call and choice of topics.
- **Style:** Critical analysis of the call and White House's approach.

A.3 CNN - Rewritten as if from Right-Wing

This rewriting and the analysis come from https://www.chatpdf.com/.

President Biden's recent phone call with Chinese President Xi Jinping has stirred controversy, with critics questioning the motives behind the conversation. Amidst global unrest, the discussions touched on sensitive issues such as conflicts in Gaza and Ukraine, North Korea's nuclear activities, Taiwan, China's assertiveness in the South China Sea, and the concerning human rights situation in Beijing. This engagement has sparked concerns about the administration's approach to dealing with China and its willingness to address these critical challenges head-on.

- **Topic:** Controversy surrounding President Biden's phone call with Chinese President Xi Jinping, global issues discussed.

- **Focus:** Critics questioning motives behind the conversation, discussion topics including conflicts in Gaza and Ukraine, North Korea's nuclear activities, Taiwan, South China Sea, human rights abuses.
- **Ideology:** Slightly left-leaning, critical of administration's approach to China and diplomatic challenges.
- **Style:** Critical examination of the engagement and concerns raised.

A.4 FOX - Rewritten as if from Left-Wing

This rewriting and the analysis come from https://www.chatpdf.com/.

President Biden's recent phone call with Chinese President Xi Jinping, the first since July 2022, has sparked apprehension about the direction of their dialogue. Despite limited information available, the anticipated topics of discussion encompassed crucial issues like Taiwan, narcotics, and China's involvement in supporting Russia during the Ukraine conflict. The White House's characterization of the call as a routine "check-in" has raised eyebrows, especially given the notable exclusions of topics such as TikTok and the origins of COVID-19, despite Biden's advocacy for legislation aimed at curbing Chinese influence.

- **Topic:** Apprehension about the direction of dialogue between President Biden and Chinese President Xi Jinping, discussion topics, White House's characterization.
- **Focus:** Anticipated discussion topics like Taiwan, narcotics, Chinese support for Russia, White House's portrayal of the call as a routine "check-in," exclusion of TikTok and COVID-19 origins.
- **Ideology:** Slightly left-leaning, critical of White House's transparency and handling of key issues.
- **Style:** Critical assessment of the call and White House's communication.

A.5 Tentative Causal Graph

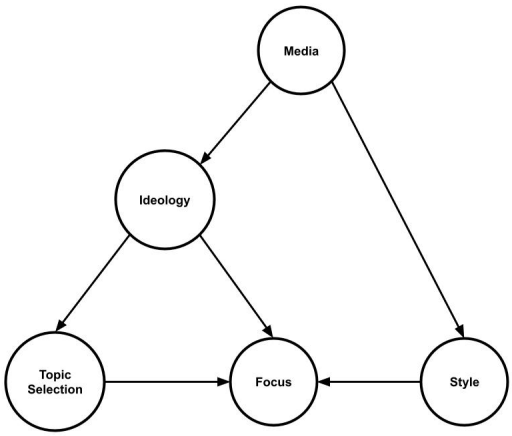

Figure 1: Tentative Causal Graph