Mechanisms of Causal Reasoning in LLMs

Ben Sturgeon, Jacy Reese Anthis, Mark Chimes, Sky Cope

Causal reasoning is a crucial part of how humans safely and robustly think about the world. To what extent do LLMs have causal reasoning? Marius Hobbhahn and Tom Lieberum (2022, Alignment Forum) approached this with probing. For this hackathon, we begin to follow up on that work by exploring a mechanistic interpretability analysis of causal reasoning in the 80 million parameters in GPT-2 Small using Neel Nanda's Easy Transformer package.

Results

We engaged in an exercise of "prompt engineering" to find input text that can tease out causality from text features that the model could be using as proxies for causality. Our primary focus was prompts about colored balls hitting each other, built on a simpler example from Hobbhahn and Lieberum (2022):

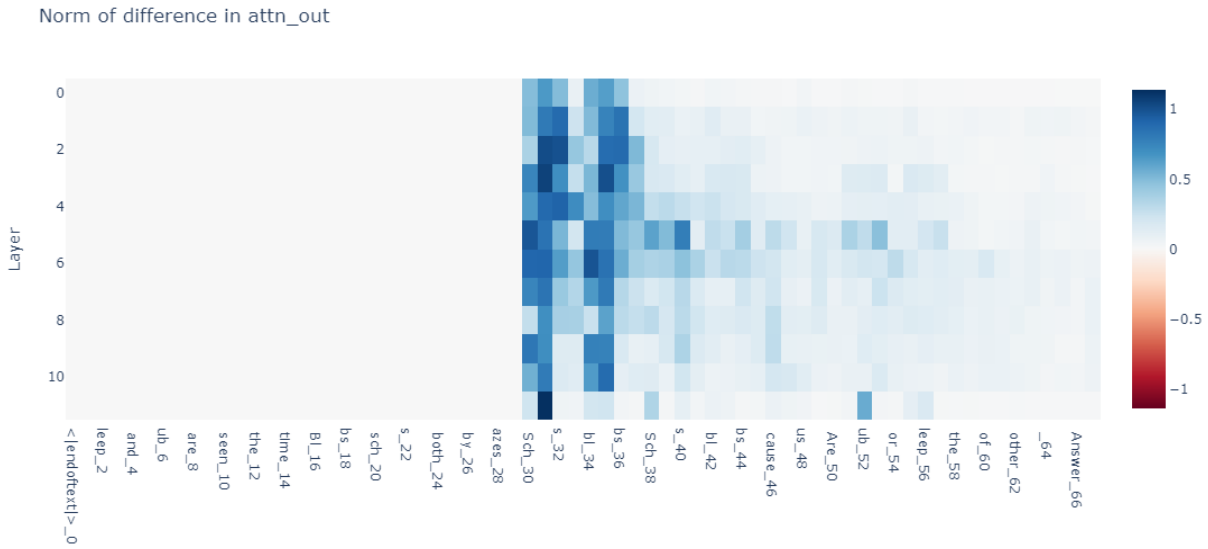
> The red ball hit the yellow ball. The purple ball hit the red ball. The yellow ball fell into the hole. The green ball hit the purple ball. The blue ball hit the green ball. Question: Which ball was third in the chain? Answer in three words:

The correct answer is purple (i.e., "the purple ball"), but there are multiple ways that the model could come to that answer: it could extract the causal chain (blue, green, purple, red, yellow) or it could use first-mentioned word order as a proxy (red, yellow, purple, green, blue). In either case, it would put the highest probability on purple, but we hypothesize that insofar as it properly extracts causal information, it will put higher probability on red than yellow because red is only one step away from purple in the causal chain but yellow is only one step away in the word order. It turns out that GPT-2 Small puts higher probability on red (21.04%) than yellow (4.15%), which could suggest some causal reasoning in GPT-2 Small, though of course that is extremely preliminary, and in this particular case, the model does not get the correct answer. (The model struggles to juggle 5 balls, especially when asked about the second, third, or fourth step in the causal chain. We would like to replicate this with bigger models.)

We also wanted to test nonsense words in a context that would correlational and causal information, again built on a simpler example from Hobbhahn and Lieberum (2022):

> Schleeps and blubbs are frequently seen at the same time. Blubbs and schleeps are both affected by bazes. Schleeps affect blubbs. Schleeps and blubbs both cause fuus. Do blubbs or schleeps cause the other? Answer:

In this case, we are again interested in using the interpretability tools (e.g., visualizing the residual stream) to see what the model is focused on, in particular whether and how it is focused on the key sentence, “Schleeps affect blubbs,” rather than the sentences that contain information on correlation, common cause (“bazes”), and collision (“fuus”). Therefore, we compare this to a prompt in which the key sentence is reversed to, “Blubbs affect schleeps.” We also test different wordings, such as changing the question to, “Are blubbs or schleeps the consequence of the other?” and find that the model only successfully performs this task in cases where we asked about which was the cause rather than which was the consequence.



We found that the model does generally attend to the causal sentence that changed in the reversed example, as well as some attention to the collider sentence, which did not change.

Scaling and automating tests

We began working on a a testing pipeline based on the SERI MATS IOI Demo notebook that can take a spreadsheet of such prompts and evaluate the model, as well as take in a prompt and a reversed prompt to produce these layer-by-layer visualizations.

Pipeline notebook:
colab.research.google.com/drive/1KiMdWMX1yJTefixOOpn6Kyg4H9QmfLFa
Prompt spreadsheet:
https://docs.google.com/spreadsheets/d/1d2eTi-jzkWx_1AfEVB8xnv-uQc6pyAxNL3endiqq8Uc

Theoretical discussion

In *The Book of Why* (2018), Judea Pearl says that deep neural networks (DNNs) cannot grasp causality.[1] Causality is an increasingly common interest in deep learning (e.g., Luo et al. 2020; Whata et al. 2022; Yuan et al. 2020; Zečević et al. 2021), and it seems to be an extremely important part of how humans understand the world. Understanding the extent of causal reasoning in large language models (LLMs) could be useful for several reasons:

a. If we know the causal model embedded in the LLM, we could better tell if it's correct, safe, generalizable, deceptive, etc.
b. We could make models more (or less) causal insofar as causal models are safer (or less safe).
c. We could improve our techniques for interpretability, especially interpreting tasks closely related to causality such as out-of-domain generalization.

What does it mean for an LLM to have causal reasoning? There are many operationalizations of this we can consider. For this interpretability workshop, we highlight two:

1. LLMs can perform well on causal language tasks (e.g., in BIG-Bench).
2. LLMs can perform well on causal language tasks (e.g., in BIG-Bench) under distribution shift.

As we peer into the circuits of LLMs, we can also consider more intrinsic, theoretical, and potentially robust notions of causal reasoning:

3. LLM circuits can be isomorphic to directed acyclic graphs (DAGs) and SCMs.
4. LLM circuits can be isomorphic to potential outcomes (POs) equations.
5. LLM circuits can be isomorphic to Granger causality.
6. LLMs can differentiate spurious and true correlations.
7. LLMs can learn Shapley values, LIME, or other causal attribution measures.
8. LLMs can take in causal inputs.[2]
9. LLMs can model a causal process.
10. LLMs can model manipulation.
11. All of the above may be expressible in LLMs, but are they learnable?
12. Causal understanding in practice is mostly about data, not architecture.

---

[1] Specifically Pearl says deep learning systems are unable to "go beyond rung one of the Ladder of Causation." In his model, "rung one" is mere association or correlation, not Rung Two ("Intervention") or the final Rung Three ("Counterfactuals"). This is formalized somewhat in Bareinboim et al. (2022).

[2] We don't know a formalization of this criterion, but Pearl said "behind every causal conclusion there must lie some causal assumption that is not testable in observational studies" (Pearl 2009). Also see this tweet and Cartwright (1995)'s adage, "No causes in; no causes out." (But where then do causes come from?)

13. LLMs can perform well under distribution shift (i.e., syntactic or semantic variation).
14. LLMs have Independent Causal Mechanisms (ICM) and Sparse Mechanism Shift (SMS).

C3 would build on work mapping LLMs / deep neural networks (DNNs) modules or circuits, or work on causal discovery algorithms. One challenge here is that for a single NN, the direction of causality is set from NN inputs to outputs. To learn the direction of a structural causal model (SCM), we may need to test each set of input and output pairing and compare performance. C5 is similar, perhaps showing that a time series DNN changes output iff older data Granger causes new data. C8 is a vague theory that some ML people bring up, and C9 and C10 are somewhat popular in philosophy, so they might be worth engaging, but any precise operationalization might be controversial. Regardless of the standard causal formalisms, one important view of causality is the functionalist duck test: "If it solves causal problems like a causal model, it is a causal model." So even if conventional DNNs are in some sense stuck in Rung One of Pearl's ladder, if they have SOTA performance on causal problems (e.g., distribution shift), they may be well-classified as causal models.

By understanding the circuits of LLMs and other DNNs, we can begin to understand the extent to which Claims 3–14 are true. While our hackathon project only begins to address these by searching for traces of causal reasoning (e.g., high attention weights on causal language in prompts), we think this may be a promising—albeit very challenging—direction for future work. Thank you for reading, and please feel free to reach out (bwm.sturgeon@gmail.com, jacy@uchicago.edu, markjchimes@gmail.com, skycope@gmail.com).