Causal Interventions on Causal Paths: Mapping GPT-2's Reasoning From Syntax to Semantics

Isabelle Lee **Joshua Lum** **Ziyi Liu** **Dani Yogatama**

University of Southern California
lee.isabelle.g@gmail.com

Abstract

While interpretability research has shed light on some internal algorithms utilized by transformer-based LLMs, reasoning in natural language, with its deep contextuality and ambiguity, defies easy categorization. As a result, formulating clear and motivating questions for circuit analysis that rely on well-defined in-domain and out-of-domain examples required for causal interventions is challenging. Although significant work has investigated circuits for specific tasks, such as indirect object identification (IOI), deciphering natural language reasoning through circuits remains difficult due to its inherent complexity. In this work, we take initial steps to characterize causal reasoning in LLMs by analyzing clear-cut cause-and-effect sentences like "I opened an umbrella because it started raining," where causal interventions may be possible through carefully crafted scenarios using GPT-2 small. Our findings indicate that causal syntax is localized within the first 2-3 layers, while certain heads in later layers exhibit heightened sensitivity to nonsensical variations of causal sentences. This suggests that models may infer reasoning by (1) detecting syntactic cues and (2) isolating distinct heads in the final layers that focus on semantic relationships.

1 Introduction

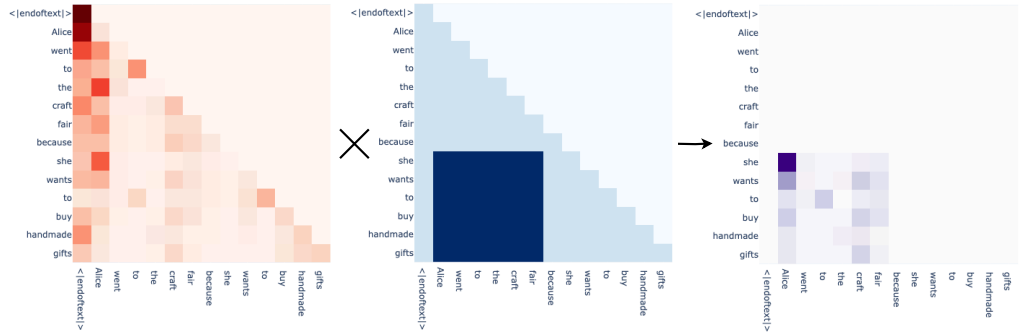
As transformer-based large language models (LLMs) scale up, their performance on diverse downstream tasks has shown remarkable improvement [Wei et al., 2022a, Srivastava et al., 2022]. These models demonstrate remarkable capabilities across various tasks, from reasoning tasks such as math problem solving and commonsense reasoning to question-answering that require knowledge synthesis Kojima et al. [2022], Zellers et al. [2018], Wei et al. [2022b], Brown et al. [2020]. Understanding and benchmarking these capabilities has become a prolific research area, as both technical communities and the general public uncover new ways to harness LLMs. Despite these impressive abilities, however, the mechanisms driving these capabilities remain largely opaque.

As model scales increase, interpreting their associated capabilities becomes increasingly challenging. Nevertheless, notable advancements in interpretability have improved our understanding of these models. Recent work in mechanistic interpretability takes a microscopic approach to analyze models [Olah et al., 2020]. Many studies derive interpretable features and behaviors from attention mechanisms using simplified toy models of transformers, revealing concepts like induction heads and in-context learning [Olsson et al., 2022, Elhage et al., 2021]. Although these insights shed light on interpretable, microscopic mechanisms like feature recognition and copying, they fall short in explaining complex, high-level behaviors in realistic tasks. A major reason for this is that circuits rely on causal interventions, which require clear distinctions between in-domain and out-of-domain examples. However, many natural language tasks are complex and inherently ambiguous; for instance,

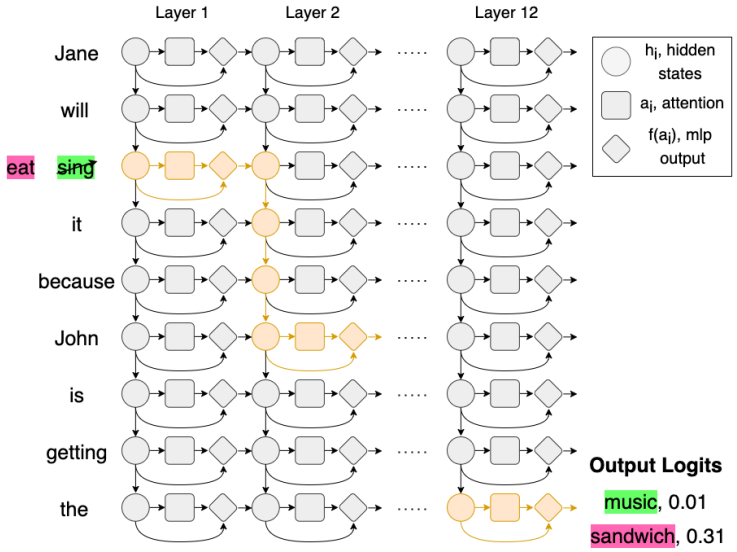
NeurIPS 2024 Workshop on Causality and Large Models (CaLM).

the distinction between reasoning and non-reasoning text is often murky. This ambiguity complicates the scaling of interpretability efforts to such macroscopic behaviors.

To begin to understand macroscopic reasoning behaviors of LLMs and link them to underlying representations, we focus on the simplest cases of reasoning by breaking them down into components like cause-and-effect relations. Specifically, we examine clear-cut causal phrases connected by markers such as "because" and "so." We design scenarios that allow for causal interventions and investigate whether model responses—such as patterns observed in attention maps and logit shifts in the residual stream—can be traced to these semantic perturbations.



(a) Attention Analysis



(b) Activation patching with an example causal trace highlighted in orange

Figure 1: Overview of methods.

In this work, We analyze GPT-2's ability to comprehend causal relationships in sentences with clear, unambiguous causal connections. In these cases, we anticipate that introducing nonsensical perturbations will reveal distinct causal circuits within the model. Our focus is on straightforward instances where action verbs interact causally with specific settings (e.g., locations) or plausible objects. We find that GPT-2 primarily captures syntactic structures within its first 2-3 layers. We then perform causal interventions on the model's semantic activations to identify which attention heads contribute to task performance. Our results reveal that a small set of attention heads consistently activates across subtasks. Future work could explore more complex causal scenarios or sentences with ambiguous causal relationships and compare these findings with larger models to determine if similar patterns emerge across different settings.

2 Overview

We explore how LLMs understand reasoning by examining their responses to sentences with straightforward reasoning structures. We conduct our experiments with GPT-2 small, a 12-layer model with decoder blocks containing self-attention layers with 12 attention heads and multilayer perceptrons (MLPs) [Radford et al., 2019]. We recognize that humans comprehend reasoning in natural language in two steps. First, by identifying syntactic cues associated with reasoning, such as phrases connected by words like "because" and "so", we assess whether a sentence likely contains reasoning relations. Next, we consider the semantic relationships within cause-and-effect phrases. Our experiments are designed to reflect this two-step reasoning process. For syntactic analysis, we use a dataset of diverse sentence structures (see Table 1). For semantic analysis, we modify cause-and-effect phrases in templated sentences (see Table 2) to make the reasoning relations either coherent or nonsensical.

3 Where Is Syntax in a Transformer?

To locate syntactical knowledge in GPT-2, we analyze the model responses to a curated synthetic dataset of causal sentences with varying syntax. We generated the dataset by prompting the language models with multiple templates, as summarized in Table 1. We assess attention patterns based on the causal phrases and delimiters, following an approach similar to the syntactical analysis performed by Vig and Belinkov [2019] on BERT.

Setup and Methods The templates used to generate the syntactical dataset in Table 1 show the syntactical structure of the sentences in the form of $[e_1, \cdots, e_n, d, c_1, \cdots c_m]$ or $[c_1, \cdots c_m, d, e_1, \cdots, e_n]$ where c_i = tokens of a cause phrase, d = causal delimiter token, and e_j = tokens of an effect phrase. Respectively, the first template refers to "because" sentences and the second template refers to "so" sentences. An example of such causal sentence is "Alice went to the craft fair because she wants to buy handmade gifts." Then, we specifically analyze the attention maps by calculating 1) how much attention is paid to the causal delimiters and 2) how much effect token attends to cause tokens. We calculate 1) as

$$P_d = \frac{\sum_{j=1}^{m} \alpha_{d,j}}{\sum_{i=1}^{n+m+1} \sum_{j=1}^{n+m+1} \alpha_{i,j}}, \tag{1}$$

where $\alpha_{i,j} = \left[\text{softmax}\left(QK^T/\sqrt{d_K}\right)V\right]_{i,j}$ with query Q, key K, and value V matrices calculated from the input tokens with attention weights with $1/\sqrt{d_K}$ as a scaling factor calculated from the dimension of the key matrix. We then calculate 2) proportion of cause-to-effect or effect-to-cause attention similarly. As described in Figure 1a, we isolate the cause-to-effect or effect-to-cause attention patterns by masking. The proportion of causal attention pattern can be expressed as

$$P_c = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{i,j}}{\sum_{i=1}^{n+m+1} \sum_{j=1}^{n+m+1} \alpha_{i,j}}. \tag{2}$$

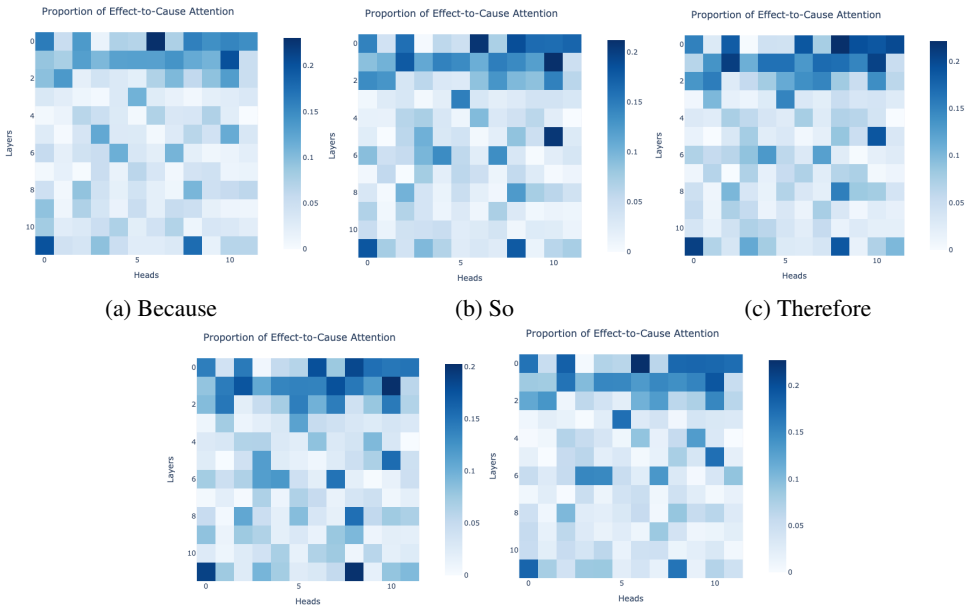
With isolated causal attention map, we perform statistical analyses per head per layers.

3.1 Results

In order to analyze syntactical understanding of GPT-2, we first compute the proportion of attention paid to causal delimiters, P_d, such as "because" and "so." Figure 4 summarizes the results, which shows that the heads that pay attention to delimiters are spread across the layers with some concentrations in the earlier layers of a transformer. On the other hand, Figure 2 shows that the heads that pay particular causal attention, P_c, tend to be most concentrated in the first 2-3 layers.

4 Locating Semantics: Where Does GPT-2 Figure Out a Sandwich Is for Eating, not Singing?

We also consider logit analysis at each layer of the model to analyze model behavior with causal sentences. From the residual stream, we calculate the per token loss at each layer, which we define



(a) Because (b) So (c) Therefore

(d) Resulting (e) Since

Figure 2: Proportion of Effect-to-Cause or Cause-to-Effect Attention.

as per layer loss. We can then hypothesize that the causal relations between two phrases in a sentence would be reflected in the per-layer logit calculation. An illustrated example of per layer loss calculation is shown in Figure 5. In this example, we take a causal sentence “we went shopping because we were bored" and perturb it to make a non-causal sentence. We swap “bored" with “sleepy" to obscure the causal relations between the two phrases of the sentence. In this work, we focus on scenarios where semantic perturbations can occur through straightforward word substitutions. Specifically, we examine sentences that involve an action verb in relation to a specific location or an action verb acting on a particular object. Because these relationships are causally specific and syntactically simple, we can easily distort the sentences to render them nonsensical, such as by replacing a location or object. Our dataset is detailed in Table 2. We see that perturbing a sentence this way is reflected in the per layer loss calculation. With this overall analysis in mind, we can then decompose the residual contribution per attention heads, per neurons, and analyze their implications for finding causal relations.

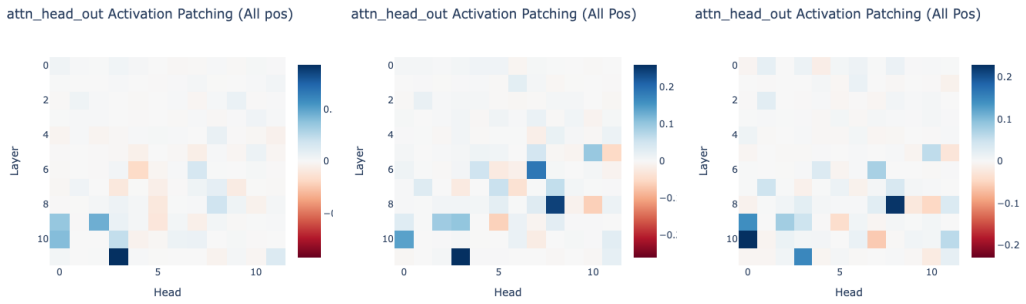
4.1 Activation Patching Results

We apply activation patching to contrastive pairs of causal sentences. As outlined in 1b, we first run our model using an original causal sentence. Next, we introduce a semantic perturbation by replacing the sentence with its contrastive pair and rerun the model. By tracking the activation differences that result in changes to the final logit predictions, we pinpoint specific model components responsible for distinguishing causal semantics from random semantics.

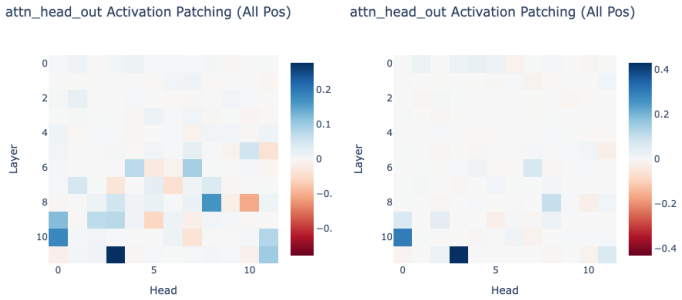
As shown in Figure 3, few distinct attention heads in the middle to last few layers contribute most to the logit difference, especially layer 11 head 2, layer 10 head 0, and layer 8 head 8, light up in most templates. We also note that in the residual stream, the “PERTUBRED” token significantly influences predictions in the earlier layers, as shown in Figures 6, 7, 8, 9, and 10.

5 Conclusion

Our investigation suggests that the model demonstrates a syntactic focus in its initial layers, with attention mechanisms primarily engaging at this stage. As processing deepens, a shift occurs, and the model begins to handle reasoning tasks in a more semantic manner, particularly in the later layers. These findings are evident in cases of clear-cut reasoning, where causal relationships can



(a) ALS template (b) ALB template (c) AOS template



(d) AOB template (e) AOB template

Figure 3: Attention head out activation patching results over all positions. O = Object, L = location, S = So, B = Because

be perturbed with word substitutions. However, ambiguity in reasoning presents a more complex challenge. Future work will aim to explore how the model adapts when faced with ambiguous or less structured reasoning tasks, as understanding these scenarios could significantly enhance the clarity of causal inference and model interpretability.

6 Related Work

Reasoning in LLMs LLMs have demonstrated remarkable “emergent" abilities for which they were not explicitly trained, though mechanisms behind them are not well understood [Wei et al., 2022a, Schaeffer et al., 2023, Lu et al., 2023]. Among them are LLMs’ ability to reason in many domains from informal, commonsense reasoning [Kojima et al., 2022, Bhagavatula et al., 2019, Zellers et al., 2018] to more formal domains such as scientific reasoning [Lu et al., 2022, Birhane et al., 2023] and mathematical reasoning [Cobbe et al., 2021, Yuan et al., 2023]. Behavioral studies have focused significant recent efforts in characterizing and benchmarking model capabilities [Srivastava et al., 2022, Huang et al., 2023], but they are not well connected to the intermediate representations and internal responses of a model. Our work provides first steps in connecting behavioral observations to internal and mechanical model responses with curated tasks.

Attention Analysis and Mechanistic Interpretability Attention maps have been used for interpreting intermediate representations and behaviors of transformers since the transformer architectures took off in language modeling [Jain and Wallace, 2019, Wiegreffe and Pinter, 2019, Clark et al., 2019, Rogers et al., 2020]. Analyzing what the language models pay attention to when making predictions can elucidate relevant features for particular labels. Recently, work in mechanistic interpretability largely approximated transformers with simplified attention matrix multiplications to describe “circuits" [Elhage et al., 2021]. “Circuits" in LLMs can be thought of as information flow through a transformer that make certain decisions or perform a particular task.

Causal Tracing (Activation Patching) and Causal Intervention While many recent behavioral characterizations of LLMs rely on post-hoc benchmarking, some interpretability methods actively engage with model responses. For instance, counterfactual perturbations on input data have been used to study subject-verb agreements in BERT by tracing model responses to particular input representations [Ravfogel et al., 2021, Elazar et al., 2022]. First introduced by Meng et al. [2022], activation patching *causally traces* the effect of perturbed input token on the activations throughout the layers and eventually on the predicted output token. Activation patching has been used to locate factual information in a transformer in the case of Meng et al. [2022], and it is frequently used for identifying circuits in LLMs. Wang et al. [2022] used activation patching to identify a circuit that performs the "indirect object identification task," in which a model predicts the name as object of an action given the previous context.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Yih, and Yejin Choi. Abductive commonsense reasoning. *ArXiv*, abs/1908.05739, 2019. URL https://api.semanticscholar.org/CorpusID:201058651.

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. Science in the age of large language models. *Nature Reviews Physics*, 5:277–280, 2023. URL https://api.semanticscholar.org/CorpusID:258361324.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. URL https://api.semanticscholar.org/CorpusID:218971783.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert's attention. In *BlackboxNLP@ACL*, 2019. URL https://api.semanticscholar.org/CorpusID:184486746.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021. URL https://api.semanticscholar.org/CorpusID:239998651.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schutze, and Yoav Goldberg. Measuring causal effects of data statistics on language model's 'factual' predictions. *ArXiv*, abs/2207.14251, 2022. URL https://api.semanticscholar.org/CorpusID:251134985.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani Gilpin. Can large language models explain themselves? a study of llm-generated self-explanations. *ArXiv*, abs/2310.11207, 2023. URL https://api.semanticscholar.org/CorpusID:264172366.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:67855860.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916, 2022. URL https://api.semanticscholar.org/CorpusID:249017743.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv*, abs/2209.09513, 2022. URL https://api.semanticscholar.org/CorpusID:252383606.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? *ArXiv*, abs/2309.01809, 2023. URL https://api.semanticscholar.org/CorpusID:261531236.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Conference on Computational Natural Language Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:234681155.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. URL https://api.semanticscholar.org/CorpusID:211532403.

Rylan Schaeffer, Brando Miranda, and Oluwasanmi Koyejo. Are emergent abilities of large language models a mirage? *ArXiv*, abs/2304.15004, 2023. URL https://api.semanticscholar.org/CorpusID:258418299.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, ..., and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615, 2022. URL https://api.semanticscholar.org/CorpusID:263625818.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *BlackboxNLP@ACL*, 2019. URL https://api.semanticscholar.org/CorpusID:184486755.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *ArXiv*, abs/2206.07682, 2022a. URL https://api.semanticscholar.org/CorpusID:249674500.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022b. URL https://api.semanticscholar.org/CorpusID:246411621.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Conference on Empirical Methods in Natural Language Processing*, 2019. URL https://api.semanticscholar.org/CorpusID:199552244.

Zheng Yuan, Hongyi Yuan, Cheng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. Scaling relationship on learning mathematical reasoning with large language models. *ArXiv*, abs/2308.01825, 2023. URL https://api.semanticscholar.org/CorpusID:260438790.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724, 2018. URL https://api.semanticscholar.org/CorpusID:53734356.

A Dataset

The datasets for syntactical and semantic analysis are generated using templates which are detailed in Table 1 and in Table 2 respectively.

Template id	Template	Type
1	Alice went to <location> because she wanted to <verb> <object> Alice went to <location> and she <verb> <object> Alice went to <random> because she <verb> <object>	B->A non-causal random
2	Alice went to <location> because <location> is a good place for <object> Alice went to <location> and <location> is <adjective> Alice went to <location> because <location> is a good place for <random>	B->A non-causal random
3	Alice play <object> because she enjoys <verb> <object> Alice play <object> and <pronoun> is <adjective> Alice play <object> because she enjoys <verb> <random>	B->A non-causal random
4	Bob and Chris made <object> so <pronoun> are <adjective1> and <adjective2> Bob and Chris made <object> while <pronoun> are <adjective1> and <adjective2> Bob and Chris made <object> so <pronoun> are <random> and <adjective2>	A->B non-causal random
5	Bob and Chris got work to do so they are <adjective> to <verb> Bob and Chris got work to do but they are <adjective> to <verb> Bob and Chris got work to do so they are <random> to <verb>	A->B non-causal random
6	Alice went to <location> because she wanted to <verb> <object> Alice went to <location> and she <verb> <object> Alice went to <random> because she <verb> <object>	B->A non-causal random
7	Alice went to <location> because she wanted to <verb> <object> Alice went to <location> and she <verb> <object> Alice went to <random> because she <verb> <object>	B->A non-causal random
8	Alice went to <location> because she wanted to <verb> <object> Alice went to <location> and she <verb> <object> Alice went to <random> because she <verb> <object>	B->A non-causal random
9	Alice went to <location> because she wanted to <verb> <object> Alice went to <location> and she <verb> <object> Alice went to <random> because she <verb> <object>	B->A non-causal random
10	Alice went to <location> because she wanted to <verb> <object> Alice went to <location> and she <verb> <object> Alice went to <random> because she <verb> <object>	B->A non-causal random

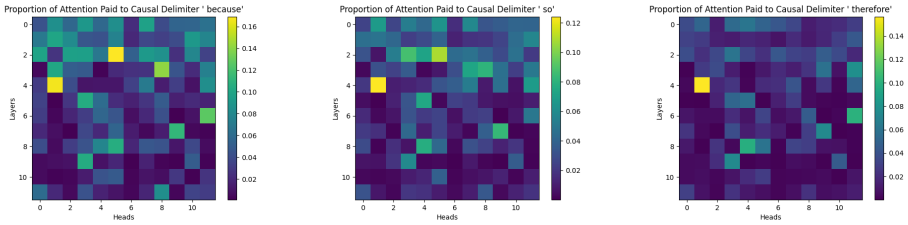
Table 1: Additional Dataset Template for Exploratory Analysis

Id	Template	Task Type	#
ALB	John had to [ACTION] because he is going to the [LOCATION].	Action $\xleftarrow{\text{because}}$ Location	6225
AOB	Jane will [ACTION] it because John is getting the [OBJECT].	Action $\xleftarrow{\text{because}}$ Object	7509
ALS	Mary went to the [LOCATION] so she wants to [ACTION].	Location $\xrightarrow{\text{so}}$ Action	4843
ALS-2	Nadia will be at the [LOCATION] so she will [ACTION].	Location $\xrightarrow{\text{so}}$ Action	5600
AOS	Sarah wanted to [ACTION] so Mark decided to get the [OBJECT]	Action $\xrightarrow{\text{so}}$ Object	6755

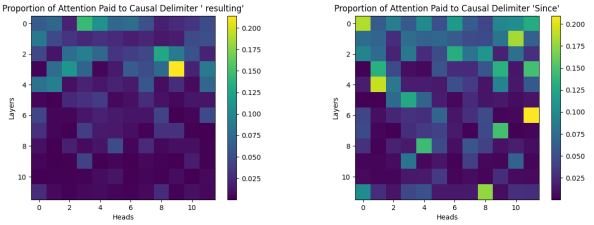
Table 2: Dataset Templates for Causal Relation Prediction

B Proportion of Attention Paid to Delimiters

Heatmap of the proportion of attention paid to causal delimiters such as "because" and "so" in GPT-2.



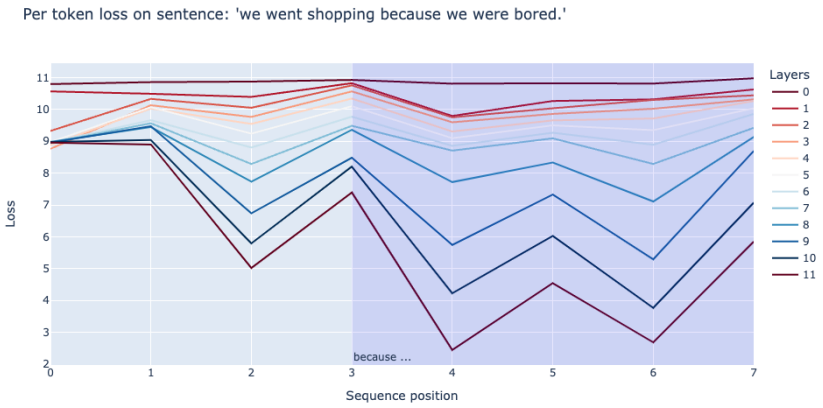
(a) Because (b) So (c) Therefore



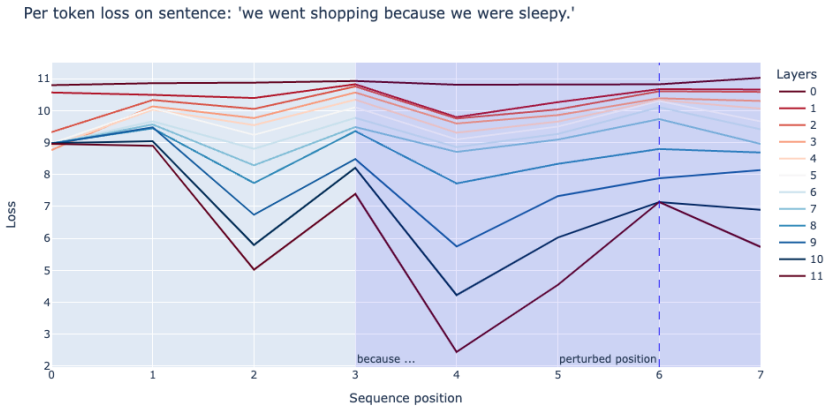
(d) Resulting (e) Since

Figure 4: Proportion of Attention Paid to Causal Delimiters.

C Logit Analysis with Semantic Perturbation



(a) Per layer logit analysis of causal sentence



(b) Per layer logit analysis of causally perturbed sentence

Figure 5: Logit analysis with per layer loss.

D Activation Patching By Model Components

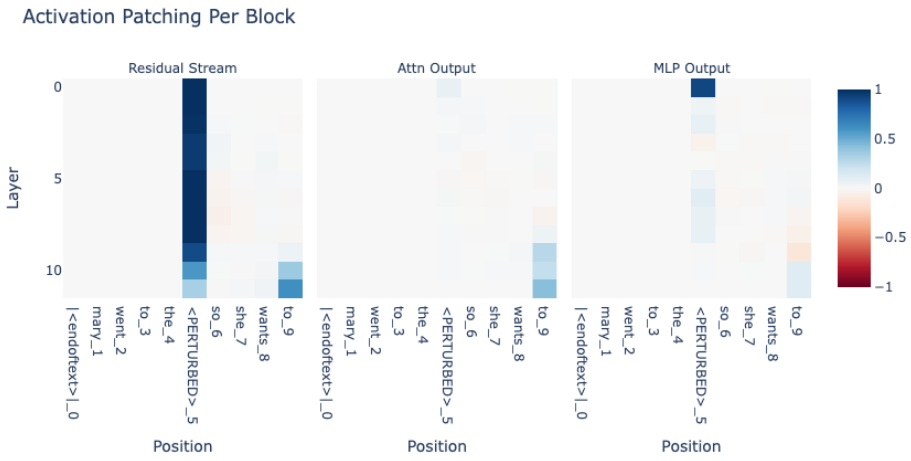


Figure 6: ALS template

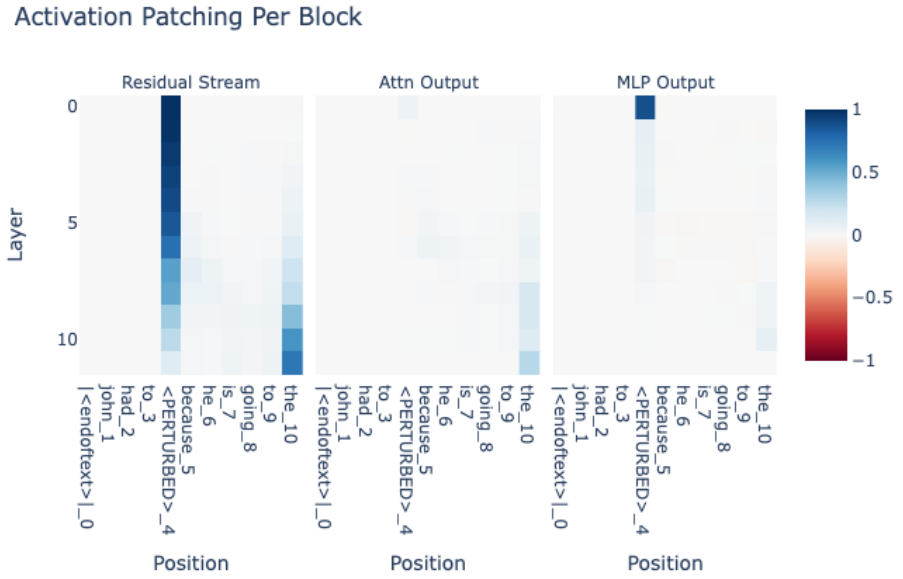


Figure 7: ALB template

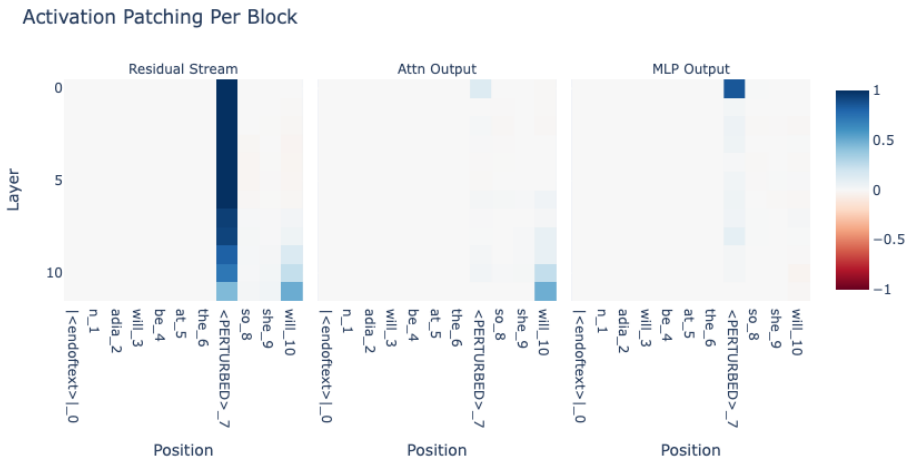


Figure 8: ALS-with template

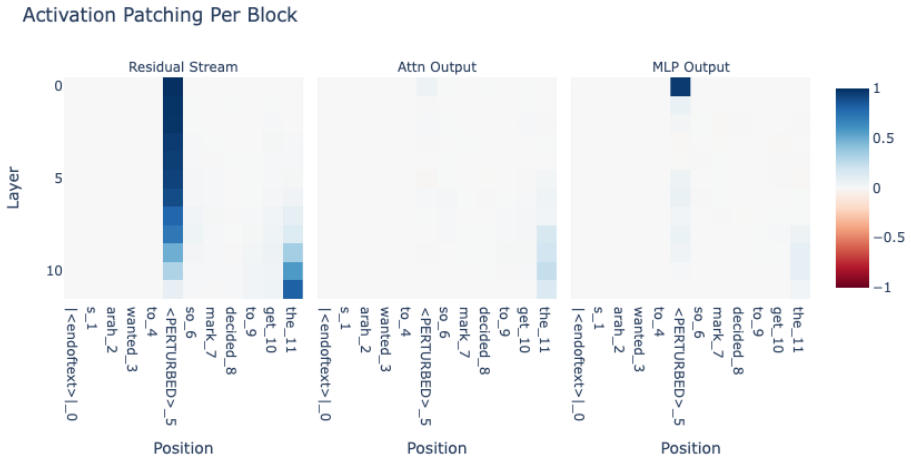


Figure 9: AOS template

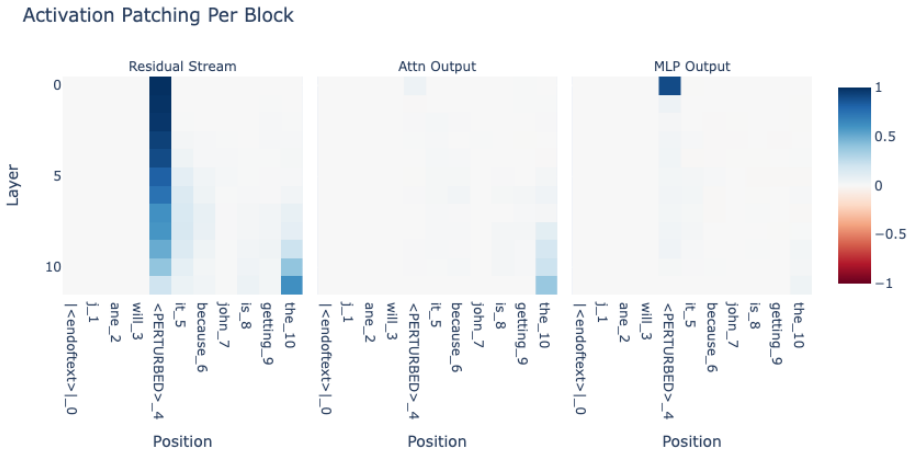


Figure 10: AOB template