Teaching Transformers Causal Reasoning through Axiomatic Training

Aniket Vashishtha [1] **Abhinav Kumar** [2] **Abbavaram Gowtham Reddy** [3] **Vineeth N Balasubramanian** [3]
Amit Sharma [1]

Abstract

For text-based AI systems to interact in the real world, causal reasoning is an essential skill. Since interventional data is costly to generate, we study to what extent an agent can learn causal reasoning from passive data. We consider an axiomatic training setup where an agent learns from multiple demonstrations of a causal axiom (or rule), rather than incorporating the axiom as an inductive bias or inferring it from data values. A key question is whether transformers could learn to generalize from the axiom demonstrations to larger and more complex scenarios. Our results, based on a novel axiomatic training scheme, indicate that such generalization is possible. We consider the task of inferring whether a variable causes another variable, given a causal graph structure. We find that a 67 million parameter transformer model, when trained on linear causal chains (along with some variations) can generalize well to new kinds of graphs, including longer causal chains, causal chains with reversed order, and graphs with branching; even when it is not explicitly trained for such settings. Our model performs at par (or better) than many larger language models such as GPT-4, Gemini Pro, and Phi-3. Our framework enables learning causal reasoning and arbitrary axioms from passive data.

1. Introduction

Causal reasoning can be defined as a set of reasoning procedures consistent with pre-defined axioms or rules that are specific to causality (Galles & Pearl, 1997). For instance, d-separation and rules of do-calculus can be considered as axioms and specifications of a collider or a backdoor set can be considered as rules that can be derived from axioms. Typically, causal reasoning is done over data corresponding to variables in a system. Axioms or rules are incorporated as inductive biases in a machine learning (ML) model, through regularization, model architecture, or the choice of variables for a particular analysis. Depending on the kind of available data—observational, interventional, or counterfactual—Pearl's ladder of causation (Bareinboim et al., 2022; Pearl & Mackenzie, 2018) defines the kinds of causal reasoning that is possible.

As axioms are the building blocks of causality, we study whether it is possible to directly learn the axioms using ML models. That is, rather than learning from data that is the result of axioms followed by a data-generating process, what if a model can learn an axiom (and thus causal reasoning) directly from symbolic demonstrations of the axiom? Such a model has the advantage that it can be applied for causal reasoning in diverse downstream scenarios, compared to task-specific causal models built using specific data distributions. This question gains relevance as language models make it possible to learn over symbolic data expressed in natural language. In fact, recent studies have evaluated whether large language models (LLMs) can do causal reasoning by creating benchmarks that encode causal reasoning problems in natural language (Kıcıman et al., 2023; Jin et al., 2024a;b; Ban et al., 2023; Long et al., 2023; Willig et al., 2022; Vashishtha et al., 2023).

Specifically, we propose a new way of learning causal reasoning through axiomatic training. We posit that causal axioms can be expressed as the following symbolic tuple, $\langle$*premise*, *hypothesis*, *conclusion*$\rangle$ where *hypothesis* refers to a causal claim and *premise* refers to any relevant information to decide whether the claim is true or not (*conclusion*). The conclusion could simply be "Yes" or "No". For example, the collider axiom from (Jin et al., 2024b) can be expressed as, *premise*: "$A \perp\!\!\!\perp B, B \not\perp\!\!\!\perp C, A \not\perp\!\!\!\perp C$"; *hypothesis*: "Does A cause C?"; and the *conclusion* as "Yes". Based on this template, a large number of synthetic tuples can be generated, e.g., by changing the variable names, changing the number of variables, changing the order, and so on. The key question is: if a model is trained on such data, would it learn how to apply the axiom to new scenarios?

To answer this question, we train a transformer model from scratch on synthetic data generated using symbolic demon-

*Equal contribution [1]Microsoft Research, India [2]MIT, USA [3]IIT, Hyderabad. Correspondence to: Amit Sharma <amshar@microsoft.com>.

Published at the ICML Workshop on Large Language Models and Cognition, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

strations of the causal irrelevance axiom (Galles & Pearl, 1997). To evaluate generalizability, we train on simple chains of the causal irrelevance axiom of size 3-6 nodes and test on multiple different aspects of generalization, including length generalization (chains of size 7-15), name generalization (longer variable names), order generalization (chains with reversed edges or shuffled nodes), structure generalization (graphs with branching). We find that a model trained on simple chains generalizes to applying the axiom multiple times over larger chains, but it is unable to generalize to the more complex scenarios like order or structure generalization. When we train a model on a combination of simple chains and chains with some edges randomly reversed, however, we find that the model generalizes well across all kinds of evaluation scenarios. Extending the findings on length generalization for NLP tasks (Kazemnejad et al., 2023; Bhattamishra et al., 2020; Haviv et al., 2022; Furrer et al., 2021), we find a critical role of different positional embeddings (Radford et al., 2018; Vaswani et al., 2023; Kazemnejad et al., 2023) in ensuring causal generalization across length and other aspects. Our best model has no positional encoding, although we find that learnable and sinusoidal encodings also work well for some scenarios. The axiomatic training approach also generalizes to a harder problem proposed in (Jin et al., 2024b). The task to distinguish correlation from causation given a premise containing statistical independence statements. Solving this task requires knowledge of multiple axioms, including d-separation, Markov property, and others. Using the same method to generate synthetic training data and train the model as above, we find that a transformer trained on task demonstrations over 3-4 variables learns to solve this task for graphs with 5 variables. On this task, our model's accuracy is higher than much larger LLMs such as GPT-4.

Our work provides a new paradigm of teaching models causal reasoning through symbolic demonstrations of axioms, which we call *axiomatic training*. The data generation and training setup is general and can be applied to learn any new axiom, as long as it can be expressed in the symbolic tuple format. More generally, our results contribute to the literature on causal learning from passive data (Lampinen et al., 2023), showing a general way to learn any causal axiom through passive demonstrations.

2. Learning Causal Axioms In Transformers

Instead of performing causal reasoning using observational or interventional data, we study whether it is possible to learn some of the general rules of causality directly from symbolic axioms. More specifically, we incorporate rules for causal reasoning in transformers as inductive biases. We begin by asking the question "are there any minimal sufficient characterization of causal principles that hold true in general?". There has been a fundamental work from Galles & Pearl (1997) where they axiomatize the causal relevance (or equivalently irrelevance). They show that for a given *stable probabilistic* causal model (defined below), there exists a finite set of axioms that completely characterized by axioms of path interception in corresponding directed graphs. We now study how such causal relevance statements can be incorporated into transformer models.

Let $\mathcal{M} = (\boldsymbol{X}, \boldsymbol{U}, \mathcal{F})$ be a causal model defined over a set of endogenous variables $\boldsymbol{X}$, exogenous variables $\boldsymbol{U}$ and the causal relationship between then defined by set of structural equations $\mathcal{F}$ (Galles & Pearl, 1997). Let $\mathcal{G}$ be the causal graph associated with the causal model $\mathcal{M}$ where the nodes $\boldsymbol{V}$ in $\mathcal{G}$ correspond to the variables in $\mathcal{M}$ and an edge $V_i \to V_j$ between any two nodes V_i, V_j denote the causal relationship between them.

Definition 2.1 (**Causal Irrelevance**, Defn. 7 in (Galles & Pearl, 1997))**.** X is probabilistically causally irrelevant to Y given Z, written ($X \nrightarrow Y|Z$) iff: $\mathbb{P}(y|z, do(X) = x) = \mathbb{P}(y|z, do(X) = x')\,, \forall x, x', y, z$ i.e., once we hold Z fixed at z, intervening on X will not change the probability of Y.

Under the stability assumption (see Assm G.1), Galles & Pearl (1997) characterizes six axioms that completely characterize causal irrelevance (Def 2.1) or equivalent causal relevance statements after using the corresponding contrapositive statements. An axiom of causal irrelevance is of the form (given in conjunctive normal form):

$$\bigwedge_s \bigvee_t (\boldsymbol{X}_i^{s,t} \nrightarrow \boldsymbol{X}_j^{s,t} | \boldsymbol{X}_k^{s,t}) \implies \bigwedge_l \bigvee_n (\boldsymbol{X}_i^{l,n} \nrightarrow \boldsymbol{X}_j^{l,n} | \boldsymbol{X}_k^{l,n})$$

where $\wedge$ is "logical and", $\vee$ is "logical or" and for a given (s,t) or (l,n) pair, $\boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{X}_k$ are disjoint subsets of observed variables $\boldsymbol{X}$. In the above causal irrelevance statement, if the antecedent is true, the consequent is also true.

Transitivity Axiom: For the scope of our study, we focus on the transitivity axiom (Axiom 3.6, Fig. 7 in (Galles & Pearl, 1997)) because it is a generic axiom, which can be used to represent complex structures like forks, colliders and chains which are used as building blocks of any causal structure. Below, we restate the transitivity axiom where A, X, Y, Z are endogenous variables of the system. $(\boldsymbol{X} \nrightarrow \boldsymbol{Y}|\boldsymbol{Z}) \implies (\boldsymbol{X} \nrightarrow \boldsymbol{A}|\boldsymbol{Z}) \vee (\boldsymbol{A} \nrightarrow \boldsymbol{Y}|\boldsymbol{Z}) \forall \boldsymbol{A} \notin \{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}\}$ Which could be equivalency converted into a causal relevance statement by taking the contrapositive: $\exists \boldsymbol{A} \notin \boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z}\ s.t. \underbrace{(\boldsymbol{X} \to \boldsymbol{A}|\boldsymbol{Z}) \wedge (\boldsymbol{A} \to \boldsymbol{Y}|\boldsymbol{Z})}_{P:\text{premise}} \implies \underbrace{(\boldsymbol{X} \to \boldsymbol{Y}|\boldsymbol{Z})}_{H:hypothesis}$ Given a premise, we can map the hypothesis based on that to the correct label ('Yes' or 'No'). Thus we can enumerate all possible tuples of $\{(P_i, H_i, L_i)\}_i$ where P_i is the premise, H_i is the hypothesis and L_i is the label *(Yes/No)* for a particular setting of the variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}$. If the premise P_i is true for the given causal graph and the

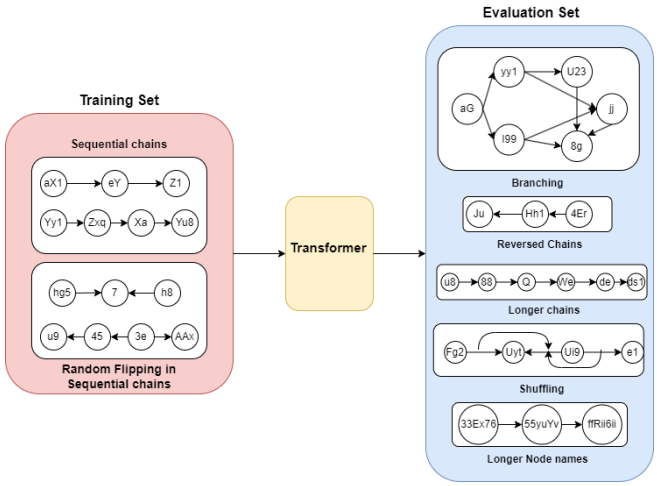

Figure 1. **Evaluating structural generalization of transformers through Axiomatic training**: Our pretraining setup is made of linear sequential chains of small length, no branching, and randomly revrsed edge directions. After training the transformer with our pre-training data D with introduced variability, structural generalization across different dimensions is observed. Specifically across more branched networks with higher average in-degree and out-degree, complete reversals, longer sequences, shuffled natural language statements of sequences and longer node names.

hypothesis can be derived by applying the above axiom (once or more than once inductively), then label L_i has to be *Yes*; otherwise, *No*. For example, let the underlying true causal graph of the system has the topology of a chain, i.e. say $X_1 \to X_2 \to X_3 \to \cdots \to X_n$. Then, a possible premise could be $X_1 \to X_2 \wedge X_2 \to X_3$, and the corresponding hypothesis $X_1 \to X_3$ will have label *Yes* whereas another hypothesis $X_3 \to X_1$ will have label *No*. The above axiom could be inductively applied multiple times to generate more complex premise, hypothesis, and labels.

2.1. Data Format for Training a Transformer Model

To develop a general axiomatic understanding of causality, we propose generating synthetic training data based on causal irrelevance axioms with some variability. This setup tests if transformers can learn and apply these axioms to diverse structures, including those not in the training set. For our training setup, our synthetic dataset D is constructed with N axiomatic instances generated using the transitivity axiom A_t. Each instance in D is structured in the form of a premise P, which is the natural language expression of the causal chain (e.g., "X causes Y, Y causes Z"), followed by the hypothesis in the form of a question H_q (e.g., "Does X cause Y?"), which is then followed by the final label L (e.g., "Yes" or "No"). Each instance in D is structured as $(P_i, H_{q_{ij}}, L_i);\ i \in \{1, \ldots, N\}, q_{ij} \in \{1, \ldots, \binom{n}{2}\}$ where n is the number of nodes in each premise, thus effectively covering all pairs of nodes in each unique chain.

2.2. Axiomatic Learning through Data Variability: A Key to Robust Model Generalization

We introduce variability at multiple levels in the training data to maximize diversity in the distribution of the training set for the transformer model, as explained below.

1. **Length level**: We restrict clength of chains to a range of 3 to 6 nodes, for our training set.
2. **Node Level**: Each node in the chain has a randomly generated alphanumeric name of 1-3 characters to prevent model failure from fixed name lengths.
3. **Edge Level**: We have two main types of causal chains in our models' training set: (a) **Sequential**: All edges are directed forward, creating a typical transitivity chain (e.g., $X \to Y \to Z$). (b) **Random Flipping**: Given a chain of sequential nodes, we randomly reverse some edges, adding complexity by disrupting direct paths between subsequent nodes (e.g., $X \to Y \leftarrow Z$).

Random flipping introduces forks and colliders, which form the building blocks of any causal DAG. This helps incorporate complexity in model training, thus aiding its capability to generalize across multiple structural dimensions.

2.3. Evaluation Strategies

To verify whether models learn axioms rather than surface-level features or trends, we design a true Out-Of-Distribution evaluation set. We evaluate models with various ablations and complex structures to analyze generalization.

1. **Length**: Evaluating if our model accurately infers causal relationships for chains (both sequential and randomly flipped) longer than those in the training set.
2. **Node Name Shift**: Testing the model's performance with longer node names, increasing from 1-3 characters used in the training set to 8-10 characters.
3. **Order of Chains**: (a) **Completely reversed chains**: Inspired by the reversal curse (Berglund et al., 2024),

which reveals LLMs' generalization failures, as they struggle to answer questions in reversed sequences despite knowing the answers in the original order. (b) **Shuffling of Sequences with Random Flipping:** Shuffling of sequences assess transformers' ability to infer accurate relationships regardless of sequence order. Longer chains (>6 nodes) also evaluated.

4. **Branching**: Evaluating causal graphs with branching across all nodes presents a challenging task. Unlike linear chains in the training set, this evaluation involves multiple branches, colliders, forks, and chains, significantly increasing complexity. We evaluate multiple branched networks constructed using the Erdös-Rényi model.

3. Learning Causal Transitivity Axioms

Pretraining Data: Employing a direct transitivity chain-based training setup, we investigate how noise improves the generalization capabilities of our transformer model to manage longer, branched, and reordered complex scenarios. We conduct various ablations using different training sets to grasp potential factors influencing the model's generalization. Our pretraining comprises approximately 175K instances of sequential chains, ranging from 3 to 6 nodes in size. We employ three training data versions. **1) Only causal chains (OCC).** Sequential chains (175K) without any random flip of edges; **2) Training Setup 1 (TS1).** Combines causal chains (101K) and sequences with flipped edges (73K), while ensuring that (reversals removed and model re-trained for evaluating on reversal chains); **3) Training Setup 2 (TS2).** Combines causal chains (132K) and sequences with flipped edges (42K) with a higher fraction of causal chains. We exclude complete reversals in TS1 to check generalization of the model to such a DAG.

Architectural and Model Training Details: We train a GPT (Radford et al., 2018) decoder based 67 million parameter model, trained from scratch on our transitivity based dataset. Our model is trained for 100 epochs (due to optimal loss convergence), with 1e-4 learning rate. Our GPT2 based model using AdamW optimizer has 12 attention layers, 8 attention heads and 512 embedding dimensions. Details of our custom tokenizer are explained in Appx. § D and details about our LLM baselines are explained in Appx. § E.

Loss Function: We optimize loss based on the ground truth label for all settings, represented as $\mathop{\mathbb{E}}_{P,H_q,L\sim P_{\text{train}}} -\log(P(L|P,H_q))$. Our earlier analysis indicated promising results with this approach compared to using next token prediction loss.

Results - Data diversity matters. Models with No PEs generalize well to longer lengths, even though they are only trained on chain length of 3-6. Model trained on only sequential chain (OCC), however, only generalize to longer Sequential chains (Table 4) but not to other DAG structures (Figure 3 for edge flip, Figure 4 for reversal, Table 5 for branching). Models trained on TS1 or TS2 generalize across all scenarios, including edge flip, order, and branching. As sequence length increases without random flipping, TS2 performs best, likely due to less noise in train set from fewer flipped sequences. This suggests that while variability aids structural generalization, excessive variability can hinder it. Reversal and shuffling evaluations are tough because the model hasn't encountered such scenarios during training to learn causal structure regardless of order. Branching is challenging due to increased inter-node connections, since training set only contains linear chains. Our findings underscore the significance of diverse data for generalization.

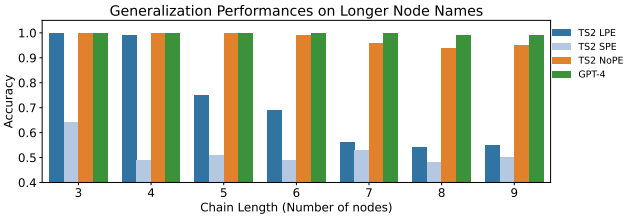

Figure 2. Evaluating generalization on causal sequences with longer node names (than the ones used in sequences in train set), and the impact of different PEs for TS-2 training set, which yields the best performance

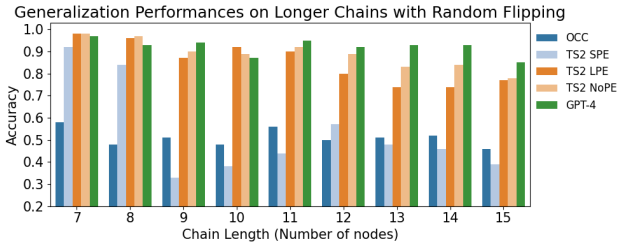

Figure 3. Generalizing to longer unseen causal sequences (>6 nodes) with random flipping using TS2 and OCC (with NoPE) training sets. OCC-trained models struggle due to limited edge-level variability, while TS2 NoPE consistently performs well, with GPT-4 being the best

Axiom-trained transformer generalizes to complex causal scenarios. Considering the model trained with TS2 (best model), it performs well across all setups. Even though our model is not explicitly trained on completely reversed chains, it still performs at par with GPT-4 (Fig. 4). Models trained on TS1 and TS2, trained explicitly without shuffling, show similar trends when evaluated on shuffled sequences with random flipping (Tab. 3). However, transformer trained on OCC setup fails for such settings. Our best models (NoPE trained on TS1 and TS2) outperforms random baselines (50%) and other billion scale models like Gemini Pro and Phi-3 (and GPT-4 in multiple cases).

Role of positional encodings: We also study the effect

of positional encoding. Sinusoidal (SPE) and Learnable PEs (LPE) perform well on longer chains but poorly when node names' length increases, even with small chain lengths (Figure 2). Similarly, SPE does not perform well across different structural dimensions like branching, and order based settings. NoPE performs consistently well across all settings showcasing its efficiency for generalization even beyond length. For a detailed discussion see Appendix H

4. Correlation to Causation w/ Axioms

We use the same model architecture from our transitivity based experiments and train it from scratch for 100 epochs using NoPE, since it performed consistently well across diverse OOD settings in our transitivity based experiments. For creating a train set, we consider the subset of the original dataset with correlational statements for graph consisting of 3 and 4 nodes. As the test set, we evaluate the model's performance directly on 5 node correlational statements.

Model	Precision	Recall	F1 Score	Accuracy
Ours	**0.72**	0.50	**0.59**	**0.64**
Phi-3	0.52	**0.60**	0.56	0.52
Gemini pro	0.52	0.59	0.55	0.52
GPT-4	0.59	0.50	0.54	0.58

Table 1. Correlation to Causation Experiments adapted from (Jin et al., 2024b)

To aid generalization, we take inspiration from our transitivity-based experiments and create different combinations of randomly created alphanumeric node names. We then derive a training set from the original dataset by instantiating the correlational statements with different combinations of alphanumeric node names. We balance the dataset by sampling equally from both classes to avoid bias in our transformer model to get a train set with 113099 instances. Then, we create a test set with 1000 randomly sampled instances of correlational statements for 5-node graph networks. Since the correlational statements are not simplistic unlike the premise from our transitivity experiments, therefore our approach of tokenizing at the character level for nodes, otherwise at the token level for rest is complicated to extend. For a straightforward extension, we tokenize the input text at the token level and use the same node names for evaluation as in the training set to avoid potential out-of-vocabulary issues.

Comparison with Baselines: As reported in (Jin et al., 2024b), due to the complexity of the task, we find that pre-trained LMs such as Gemini Pro and Phi-3 perform similar to a random guess (52% accuracy, see Table 1). While GPT-4 does perform slightly better, it's performance is still low (58% accuracy). Remarkably, our small transformer model performs better than all baselines with 64% accuracy; 6% points higher than GPT-4. With further exploration of different training setups, axiomatically-trained transformer models may be optimised further for such causal reasoning tasks.

5. Discussion and Conclusion

We propose an axiomatic training method to teach causal reasoning to transformers. Our results show that a transformer can learn to apply a causal axiom and generalize to multiple, complex graph structures that were not seen during training. Future work includes extending axiomatic training to learn multiple axioms, operate on naturally-occuring text data, and explore different pre-training losses.

Applicability to Causal Tasks: While our current work focuses on the transitivity axiom for causal relevance, extending the work to other causal axioms from (Galles & Pearl, 1997) is an interesting research direction. In addition, we may consider other axioms that are relevant for downstream tasks such as effect inference. For example, if a transformer model can be trained to validate the d-separation rule—given two variables X and Y, are they independent given a variable set Z?—then repeated applications of the rule can be used to derive a backdoor set. Another interesting direction is to extend the training approach for both deterministic and probabilistic causal models.

Generalization to Logical Reasoning: While we focused on causal reasoning axioms, our axiomatic training approach is general and can be applied to any formal system based on axioms. For instance, the same axiomatic training procedure can be used for teaching LMs logical reasoning tasks such as deductive reasoning. For instance, recent work (Saparov et al., 2023) evaluates the deductive reasoning capabilities of LLMs and measures their generalization abilities along depth, width, and compositional abilities. As the depth increases, performance of LLMs deteriorates. It will be interesting to see whether axiomatic training can be applied to learn deductive reasoning axioms and improve the reasoning abilities of LMs.

References

Ban, T., Chen, L., Wang, X., and Chen, H. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*, 2023.

Bareinboim, E., Correa, J., Ibeling, D., and Icard, T. On pearl's hierarchy and the foundations of causal inference (1st edition). In Geffner, H., Dechter, R., and Halpern, J. (eds.), *Probabilistic and Causal Inference: the Works of Judea Pearl*, pp. 507–556. ACM Books, 2022.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: Llms trained on "a is b" fail to learn "b is a", 2024.

Bhattamishra, S., Ahuja, K., and Goyal, N. On the abil-

ity and limitations of transformers to recognize formal languages, 2020.

Furrer, D., van Zee, M., Scales, N., and Schärli, N. Compositional generalization in semantic parsing: Pre-training vs. specialized architectures, 2021.

Galles, D. and Pearl, J. Axioms of causal relevance. *Artificial Intelligence*, 97(1):9–43, 1997. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(97)00047-7. Relevance.

Haviv, A., Ram, O., Press, O., Izsak, P., and Levy, O. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1382–1390. Association for Computational Linguistics, December 2022. doi: 10.18653/v1/2022.findings-emnlp.99.

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Adauto, F. G., Kleiman-Weiner, M., Sachan, M., and Schölkopf, B. Cladder: Assessing causal reasoning in language models, 2024a.

Jin, Z., Liu, J., Lyu, Z., Poff, S., Sachan, M., Mihalcea, R., Diab, M., and Schölkopf, B. Can large language models infer causation from correlation?, 2024b.

Kazemnejad, A., Padhi, I., Natesan Ramamurthy, K., Das, P., and Reddy, S. The impact of positional encoding on length generalization in transformers. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 24892–24928. Curran Associates, Inc., 2023.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Lampinen, A., Chan, S., Dasgupta, I., Nam, A., and Wang, J. Passive learning of active causal strategies in agents and language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 1283–1297. Curran Associates, Inc., 2023.

Long, S., Piché, A., Zantedeschi, V., Schuster, T., and Drouin, A. Causal discovery with language models as imperfect experts. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

Pearl, J. and Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2018.

Saparov, A., Pang, R. Y., Padmakumar, V., Joshi, N., Kazemi, S. M., Kim, N., and He, H. Testing the general deductive reasoning capacity of large language models using ood examples, 2023. URL https://arxiv.org/abs/2305.15269.

Vashishtha, A., Reddy, A. G., Kumar, A., Bachu, S., Balasubramanian, V. N., and Sharma, A. Causal inference using llm-guided discovery, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

Willig, M., Zečević, M., Dhami, D. S., and Kersting, K. Probing for correlations of causal facts: Large language models and causality. 2022.

Appendix

A. Performance Results for Different Evaluation Setups

Tables 2 and 3 shows the results of generalization to reversal and shuffling; Table 4 shows the results on length generalization; and Table 5 shows the results on branching generalization. Figures 4 and **??** highlight generalization performance on reversal and longer chains.

Model	3	4	5	6
Baselines				
GPT-4	0.97	**0.99**	0.98	0.92
Gemini Pro	0.61	0.59	0.66	0.62
Phi-3	0.80	0.69	0.73	0.69
Axiomatic Training				
TS1 w NoPE	0.98	**0.99**	0.92	0.91
TS1 w SPE	**1.00**	**0.99**	**0.99**	**0.97**
TS2 w NoPE	0.99	**0.99**	0.95	0.94
TS2 w SPE	0.98	0.97	0.93	0.94
TS2 w LPE	0.99	0.98	0.95	**0.97**
OCC w NoPE	0.33	0.18	0.10	0.09

Table 2. Evaluated on completely reversed chains, even when not explicitly trained on reversed chains. Model trained only on sequential chains (OCC), performs the worst, while transfomer trained on both Sequential chains, and sequences with random flipping perform the best (training sets: TS1 and TS2). Accuracy Metric reported. This setup is inspired by the (Berglund et al., 2024) setup

Model Config	3	4	5	6	7	8	9
Baselines							
GPT-4	0.99	**0.97**	**0.89**	**0.85**	**0.95**	**0.90**	**0.90**
Gemini Pro	0.75	0.73	0.72	0.76	0.71	0.68	0.74
Phi-3	0.88	0.86	0.82	0.79	0.76	0.73	0.79
Axiomatic Training							
TS1 NoPE	**1.00**	0.94	0.87	0.84	0.80	0.76	0.73
TS1 LPE	**1.00**	0.95	0.87	0.83	0.78	0.78	0.71
TS1 SPE	**1.00**	0.94	0.86	0.83	0.76	0.73	0.68
TS2 NoPE	**1.00**	0.95	0.87	0.84	0.79	0.76	0.73
TS2 w LPE	**1.00**	0.94	0.87	0.84	0.80	0.76	0.73
TS2 w SPE	0.99	0.94	**0.89**	0.84	0.75	0.74	0.49
OCC w NoPE	0.69	0.62	0.57	0.54	0.57	0.53	0.52

Table 3. Evaluated on shuffled natural language sequence of randomly flipped sequence. Random flipping, length (7-9) and random flipping add complexity to the evaluation setup, since our model is not trained on shuffled set. Accuracy metric is reported

B. Example of Instances from Corr2Causation Benchmark

Following is one of the example instances from the benchmark of Corr2Cause (Jin et al., 2024b), where the model has to infer causal relationships from correlational statements.

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 B. A correlates with C. A correlates with D. B correlates with C. B correlates with D. C correlates with D. However, B and D are independent given A. B and D are independent given A and C. C and D are independent given A. C and D are independent given A and B. Hypothesis: There exists at least one collider (i.e., common effect) of A and B.*

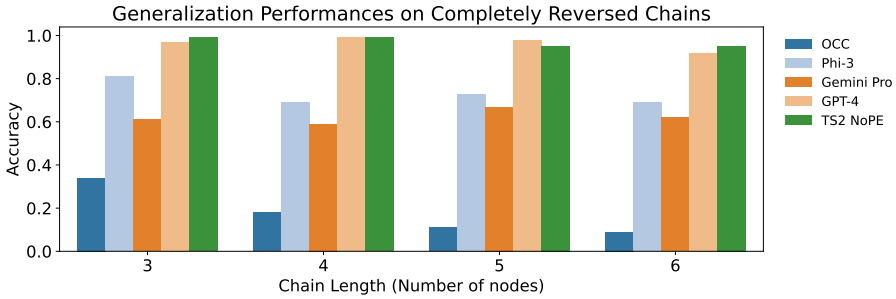

Figure 4. Performance comparison of our best performing transformer model trained on TS2 with NoPE (trained without any completely reversed chains), against larger models like GPT-4, Gemini Pro and Phi-3.

C. Example of Instances from Our Evaluation Sets

Following is one of the instances from the evaluation set for sequences with random flipping, where the model has to infer causal relationships from natural language statements.
Premise:*V causes f. f causes jbj. ag causes jbj. ag causes rBz. rBz causes Tm2. EaT causes Tm2. Hypothesis: Does V cause f?*

Following is one of the instances from the evaluation set for sequences with reversals, where the model has to infer causal relationships from natural language statements.
Premise:*LQw causes e2. p causes LQw. u causes p. a causes u. Hypothesis: Does e2 cause LQw?*

D. Custom Tokenizer details

For tokenization, we develop a custom tokenizer. Alphanumeric node names are tokenized at a character level, while terms like 'causes', 'Does', 'cause', 'Yes', and 'No' are tokenized at the word level. The intuition behind this approach is to avoid out of vocabulary (OOV) tokens in the test time, since the alphanumeric node names of test set are different then the training set and are created randomly, therefore creating a high chance of coming across unseen node names. Following this approach, the vocab size of our transformer model is extremely constrained (69) since it only contains 4-5 word tokens and rest alphanumeric characters along with punctuation marks.

Model	7		8		9		10		11		12		13		14		15	
	FS	RF	FS	RF	FS	RF	FS	RF	FS	RF	FS	RF	FS	RF	FS	RF	FS	RF
Baselines																		
GPT-4	0.95	0.98	0.97	0.93	0.87	**0.94**	0.91	0.87	**0.90**	**0.95**	0.92	**0.92**	0.85	**0.93**	**0.93**	**0.93**	**0.89**	**0.86**
Gem-Pro	0.63	0.73	0.69	0.74	0.64	0.75	0.65	0.81	0.72	0.78	0.60	0.80	0.59	0.68	0.67	0.64	0.61	0.66
Phi-3	0.81	0.85	0.96	0.85	0.85	0.85	0.87	0.89	**0.90**	0.86	0.84	0.85	0.91	0.84	0.90	0.80	0.78	0.85
Axiomatic Training																		
TS1 w NoPE	**1.00**	**0.99**	0.95	0.96	0.88	0.89	0.76	0.88	0.73	0.90	0.77	**0.92**	0.61	0.82	0.67	0.78	0.68	0.81
TS1 w LPE	0.98	0.96	0.92	**0.97**	0.77	0.90	0.59	0.87	0.57	0.86	0.57	0.84	0.55	0.73	0.51	0.76	0.50	0.68
TS1 w SPE	0.99	0.95	0.95	0.94	0.86	0.76	0.80	0.75	0.76	0.79	0.84	0.68	0.79	0.63	0.85	0.65	0.77	0.69
TS2 w NoPE	**1.00**	0.98	**0.99**	**0.97**	**0.92**	0.91	0.88	0.90	0.86	0.92	**0.95**	0.90	**0.96**	0.83	0.81	0.84	0.85	0.78
TS2 w LPE	**1.00**	0.98	0.88	**0.97**	0.80	0.88	0.62	**0.92**	0.66	0.91	0.64	0.81	0.65	0.75	0.62	0.75	0.62	0.77
TS2 w SPE	0.95	0.93	0.81	0.84	0.56	0.34	0.50	0.38	0.50	0.44	0.51	0.57	0.46	0.74	0.52	0.75	0.50	0.77
OCC w NoPE	0.98	0.58	0.79	0.49	0.86	0.51	**0.92**	0.49	0.72	0.57	0.90	0.50	0.81	0.52	0.84	0.52	0.83	0.46

Table 4. Results on longer chains of linear sequential chains with all edges in forward direction (Only causal chains or forward sequence denoted using FS) and sequences with randomly flipped edges (Random flipping so denoted with RF). TS1 and TS2 denote Pretraining data setup 1 and 2 from Section 4. SPE: Sinusoidal PE, LPE: Learnable PE, w/o PE: No PE. Model remains the same across all setups (67 Million parameter based). For longer chains, NoPE performs best on sequential linear setup. Accuracy metric is used

Model	5		8		10		12	
	BF=2	BF=1.4	BF=2	BF=1.4	BF=2	BF=1.4	BF=2	BF=1.4
Baselines								
GPT-4	**0.98**	**0.95**	**0.91**	**0.90**	**0.84**	**0.88**	**0.82**	**0.86**
Gemini Pro	0.77	0.74	0.72	0.76	0.71	0.73	0.73	0.71
Phi-3	0.87	0.83	0.82	0.79	0.77	0.77	0.75	0.80
Axiomatic Training								
OCC w NoPE	0.52	0.51	0.53	0.52	0.52	0.55	0.49	0.47
TS1 w LPE	0.79	0.84	0.71	0.76	0.68	0.69	0.65	0.65
TS1 w SPE	0.72	0.79	0.63	0.64	0.56	0.61	0.52	0.59
TS1 w NoPE	0.77	0.84	0.73	0.76	0.68	0.70	0.62	0.66
TS2 w LPE	0.72	0.80	0.61	0.71	0.62	0.63	0.56	0.63
TS2 w SPE	0.52	0.70	0.49	0.49	0.49	0.49	0.51	0.52
TS2 w NoPE	0.83	0.86	0.74	0.77	0.69	0.74	0.64	0.70

Table 5. Evaluated on branched graphs created using Erdos Renyl, with varying branching factors (calculated by number of edges/number of nodes). TS1 and TS2 denote Pretraining data setup 1 and 2 from Section 3. OCC setup denotes Only sequential Causal Chains with no random flipping. SPE: Sinusoidal PE, LPE: Learnable PE, w/o PE: No PE. Decoder model remains the same across all setups (67 Million parameter), Accuracy metric is used

E. Baselines: How well do LLMs do on these evaluations?

Given recent work on how LLMs can be leveraged for causal reasoning (Kıcıman et al., 2023; Vashishtha et al., 2023; Ban et al., 2023), we include language models such as GPT-4 *(gpt-4-32k)* (**?**), Gemini *(gemini-pro)* (**?**) and Phi-3 *(Phi-3-mini-128k-instruct)* (**?**) as baselines. Note that each of these models is significantly larger than our model and known to perform well on reasoning tasks, with the smallest baseline model Phi-3 having 3.8 billion parameters. We incorporate both commercial (GPT-4 and Gemini Pro) and open-source (Phi-3) models covering a range of size and capabilities. To evaluate the baseline models, we follow a simple zero-shot prompting strategy. For each tuple, we provide the natural language expression of the causal graph *(Premise)* followed by the question *(Hypothesis)* and prompt the LM to answer it in either 'Yes' or 'No' *(Label)*. Here is an example prompt: *"EX causes T. T causes 9. 9 causes W. W causes 7. 7 causes M. M causes a. Does EX cause T? Answer in 'Yes' or 'No' only."*

F. Compute Resources

We run our experiments on 1 A-100 GPU system, for training our models from scratch and evaluating them. We use 1 GPT-4 API for baseline experiments, while Phi-3 and Gemini Pro provide free resources for inferencing.

G. Formal Definitions of Axioms of Causal Irrelevance

Here we restate the stability assumption for a causal model from (Galles & Pearl, 1997) that gives a richer set of finite axiomatization for probabilistic causal irrelevance.

Assumption G.1 (Stability, Definition 9 in (Galles & Pearl, 1997))**.** Let $\mathcal{M}$ be a causal model. Then an irrelevance $(X \nrightarrow Y|Z)$ in $\mathcal{M}$ is stable it is shared by all possible probability distribution over $\mathcal{M}$. The causal model $\mathcal{M}$ is stable if all of the irrelevances in $\mathcal{M}$ are stable.

H. Trend Breakdown of Results

Length Generalization: Table 4 shows accuracy of different models when evaluated on larger causal chains that were not seen during training. Among the baseline pre-trained LMs, GPT-4 obtains the highest accuracy on both standard and randomly flipped chains. It is remarkable that our **TS2 (NoPE)** model obtains competitive performance to the trillion-scale GPT-4 model, even though it had never seen larger sequences during training. In particular, for chains of size 7-13, **TS2 (NoPE)** obtains higher or comparable accuracy than GPT-4 across the standard and randomly flipped chains. Its accuracy decreases for chains of length 14-15 (0.85 for standard chains and 0.78 for randomly flipped chains) but is still significantly higher than that of LMs like Gemini-Pro and Phi-3. Note that a random prediction would yield a 50% accuracy, indicating that the axiomatically-trained **TS2 (NoPE)** model can generalize its reasoning to causal chains much longer than 6 even though it was trained only on chains upto length 6.

Node Name Shift: For models trained on **TS2** dataset, we also evaluate generalization to changes in variable names (Figure 2). We find that **TS2 (NoPE)** is robust to node name changes and retains its high accuracy as new, longer names are introduced. It also retains its generalizability to longer sequences with new node names, performing similarly to GPT-4.

Order of Causal Sequences: We now consider how variations in the causal structure impact generalization of axiomatically-trained models. In Table 3, we consider the complex evaluation setup of shuffling, which includes shuffled order of causal sequences with random flipping for increasing length (even beyond the ones in train set). On this task, **TS2 (NoPE)** obtains higher accuracy than Gemini Pro and Phi-3 on chains of length up to 8. At length 9, **TS2 (NoPE)** obtains 0.73 accuracy which is comparable to Gemini Pro (0.74) and significantly better than random.

We observe a similar pattern for evaluation on completely reversed sequences in Table 2. This is an extreme case of out-of-distribution data since most causal edges are left-to-right in the training data whereas the test data contains all right-to-left edges. On this task, our axiomatically trained model **TS2 (NoPE)** outperforms GPT-4 when restricted to chain lengths of 3-6. In particular, its accuracy (0.94 for chains of length 6) is substantially higher than Gemini Pro and Phi-3 (0.62 and 0.69 respectively).

Branching: Finally, we consider the hardest evaluation task involving non-linear chains where we introduce general Erdos-Renyi graphs as the causal sequences while the training data contains only linear chains. Here the length of sequence corresponds to the number of nodes in the graph and we study the performance differences as the branching factor is varied. While GPT-4 obtains the best accuracy across increasing graph sizes, our **TS2 (NoPE)** model obtains higher accuracy than Gemini Pro for all graph sizes except one. Even when evaluated on graphs with 12 nodes and 1.4 branching factor, the **TS2 (NoPE)** model obtains 70% accuracy, significantly better than random (50%). Note that the training data only included graphs with a branching factor of 1.

Summary: Across all evaluation setups, our axiomatically trained model **TS2 (NoPE)** performs significantly better than random baselines even as chain lengths are increased beyond its training data. In particular, even though our model was not trained on fully reversed chains, it performs at par with the significantly larger GPT-4 model (Fig. 4). For other tasks, it often outperforms or matches the accuracy of billion-scale models like Gemini Pro and Phi-3. These results indicate that a model trained axiomatically can learn to reason about more complex causal structures from demonstrations of simple causal sequences. This suggests the potential of axiomatic training for reasoning over causal graphs.