Causal World Representation in the GPT Model

Anonymous Author(s)
Affiliation
Address
email

Abstract

Are generative pre-trained transformer (GPT) models only trained to predict the next token, or do they implicitly learn a world model from which a sequence is generated one token at a time? We examine this question by deriving a causal interpretation of the attention mechanism in GPT, and suggesting a causal world model that arises from this interpretation. Furthermore, we propose that GPT-models, at inference time, can be utilized for zero-shot causal structure learning for in-distribution sequences. Empirical evaluation is conducted in a controlled synthetic environment using the setup and rules of the Othello board game. A GPT, pre-trained on real-world games played with the intention of winning, is tested on synthetic data that only adheres to the game rules. We find that the GPT model is likely to generate moves that adhere to the game rules for sequences for which a causal structure is encoded in the attention mechanism with high confidence. In general, in cases for which the GPT model generates moves that do not adhere to the game rules, it also fails to capture any causal structure.

1 Introduction

In recent years, the generative pre-trained transformer (GPT) model Radford et al. [2018] has demonstrated high-quality generative capabilities, as perceived by humans. Although this model is trained to generate one token at a time, it have been demonstrated to perform a range of tasks beyond next-token prediction, such as visual understanding and even symbolic reasoning [Liu et al., 2024, Team et al., 2023, Chowdhery et al., 2023]. Are these emergent abilities [Li et al., 2023] or merely a 'mirage' resulting from the choice of metric and task Schaeffer et al. [2024]?

In this paper we suggest that there is no restriction in the GPT architecture for learning conditional independence (CI) relations between tokens in a sequence. Moreover, under certain assumptions, a causal structure is directly entailed from these CI relations. One may ask whether this lack of restriction results in implicitly learning a causal model of the world during the pre-training procedure of GPT. Assuming that, both, a causal world model and a model based on surface statistics are sufficient solutions. One possibility is that a causal world model is a more compact solution and thus more probable to be learned during pre-training (Occam's razor). For example, if weights are distributed from a uniform distribution in the surface statistics model, then a causal structure limits the range of their distribution. If so, what are the assumptions underlying this causal world model?

Recently, Rohekar et al. [2024] derived a causal interpretation for unmasked self-attention in BERT models Devlin et al. [2019]. In this paper we follow a similar approach, with several differences, and propose a causal interpretation of the masked attention mechanism in GPT. We additionally define a corresponding causal world model. The ABCD method Rohekar et al. [2024] is adapted and used to learn causal structures of which the induced dependency-relations are encoded in the attention matrices in GPT. We then ask whether errors generated by GPT are correlated with the uncertainty in the causal structure representation by the attention matrices. To this end, we define a metric based on the entropy of p-values of CI tests that are used for inferring the causal structures.

Submitted to NeurIPS 2024 Workshop on Causality and Large Models (CaLM). Do not distribute.

Recent work examined the internal process of large language models and examined whether a world model is implicitly learned using a well-defined and constrained setting, such as in the Chess game setting Toshniwal et al. [2022] and Othello board game setting Li et al. [2023]. For the Othello board game setting, Li et al. [2023] demonstrated that the board state can be inferred from attention matrices in GPT, and Nanda et al. [2023] showed that even a linear classifier suffices to reconstruct the state of the board game from the attention matrices. They claim an emergent world model in GPT. Nevertheless, they do not provide explanation on how the board game is encoded within the attention matrices and why the attention mechanism can represent the board state. In essence, they do not provide an explanation to the apparent emergence of the world model. In addition, the world model they reconstruct (board game state) is specific only to the domain for which the GPT model was trained and lacks recovering the generative mechanism of the token-sequences.

In this paper, we consider the structural causal model as the world model, which describes the generative process and is generally applicable for a range of applications (not domain specific, such as Othello board state, used by Li et al. [2023]). We explore why GPT is able to capture a world model and its apparent emergence.

2 Preliminaries

In this section we provide notations and descriptions of self attention in the GPT architecture, and structural causal models. Matrices are written in bold, vectors in bold-italic, and models in calligraphic font. A summary of the main symbols that are used in this paper is given in Table 1.

2.1 Attention in GPT

Attention is a mechanism that estimates network weights with respect to the context in an input sequence of tokens [Schmidhuber, 1992]. In a GPT model, which is based on the decoder part of the Transformer architecture Vaswani et al. [2017], an attention layer estimates an $n \times n$ lower-triangular (masked) attention matrix $\mathbf{A}$ given an input sequence of n tokens. The input sequence is in the form of an $n \times d$ matrix $\mathbf{Y}$, where the i-th row vector is $\mathbf{Y}(i, \cdot)$, is an embedding (representation) of the i-th token in d dimensions. The attention matrix is estimated by $\mathbf{A} = softmax(\mathbf{Y}\mathbf{W}_{QK}\mathbf{Y}^\top)$, such that $\mathbf{A}$ is lower triangular and the rows sum to 1[1]. In addition to the attention weights, the attention layer calculates a values matrix, $\mathbf{V} = \mathbf{Y}\mathbf{W}_V$, where row $\mathbf{V}(i, \cdot)$ is the value vector of the i-th token. Then, the output embeddings are

$$\mathbf{Z} = \mathbf{A}\mathbf{V}, \tag{1}$$

where the i-th row, $\boldsymbol{Z}_i$, is the embedding of the i-th output token. In a GPT, several attention layers are stacked, and pre-trained such that the i-th output embedding in the last layer predeicts the $(i+1)$-th input token. That is, predicts the next input token.

It is important to note that in the GPT architecture, the embedding of one token is influenced by another token only by the attention matrix, $\mathbf{A}$. In addition, note that an attention matrix $\mathbf{A}$ is estimated *uniquely* for each input sequence of tokens, using weight matrices $\{\mathbf{W}_{QK}, \mathbf{W}_V\}$ learned *commonly* for all in-distribution input sequences.

2.2 Structural Causal Model

A structural causal model (SCM) is a model that can encode causal mechanisms in a domain [Pearl, 2009, Spirtes et al., 2000, Peters et al., 2017] and explain data samples generated from these causal mechanisms Pearl and Mackenzie [2018]. An SCM is a tuple $\{\boldsymbol{U}, \boldsymbol{X}, \mathcal{F}, P(\boldsymbol{U})\}$, where $\boldsymbol{U} = \{U_1, \ldots, U_m\}$ is a set of latent exogenous random variables, $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ is a set of endogenous random variables, $\mathcal{F} = \{f_1, \ldots, f_n\}$ is a set of deterministic functions describing the values $\boldsymbol{X}$ given their direct causes, and $P(\boldsymbol{U})$ is the distribution over $\boldsymbol{U}$. Moreover, each endogenous variable X_i has exactly one unique exogenous cause U_i ($m = n$). The value of an endogenous variable X_i, $\forall i \in [1, \ldots, n]$ is determined by

$$X_i \leftarrow f_i(\boldsymbol{Pa}_i, U_i) \tag{2}$$

[1]The weight matrix is $\mathbf{W}_{QK} = \mathbf{W}_Q\mathbf{W}_K^\top/\sqrt{d_K}$, where generally the weight matrices $\mathbf{W}_Q$ and $\mathbf{W}_K$ are learned explicitly and d_K is the number of columns in these matrices Vaswani et al. [2017].

where $\boldsymbol{Pa}_i$ is the set of direct causes (parents in the causal graph) of X_i, and left-arrow indicates assignment resulting from the cause-effect relation. A graph $\mathcal{G}$ corresponding to an SCM consists of a node per variable, and directed edges for direct cause-and-effect relations that are evident from $\mathcal{F}$.

In this paper we employ a linear-Gaussian SCM having directed and acyclic graphs (DAG). In these models each variable is determined by a linear combination of its direct causes and an independently distributed additive noise determined by a corresponding normally distributed exogenous variable.

For a linear-Gaussian SCM let $\mathbf{G}$ be a weight matrix, where $\mathbf{G}(i,j)$ is the weight of parent (direct cause) node X_j linearly determining the child (direct effect) node X_i. Node X_k is not a parent of X_i if and only if $\mathbf{G}(i,k)=0$. In addition, $\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{\mu_U}, \mathbf{C_U})$, where in this paper we assume $\boldsymbol{\Sigma}$) is a diagonal matrix. The set of functions $\mathcal{F}$ is defined such that $\forall i \in [1,\ldots,n]$,

$$X_i \leftarrow \mathbf{G}(i,\cdot)\boldsymbol{X} + U_i. \tag{3}$$

Assuming a DAG and causally sorted nodes (ancestors precede their descendants), $\mathbf{G}$ is strictly lower triangular (zero diagonal). Given the assignment, we can write in matrix form $\boldsymbol{X} = \mathbf{G}\boldsymbol{X} + \boldsymbol{U}$, and

$$\boldsymbol{X} = (\mathbf{I}-\mathbf{G})^{-1}\boldsymbol{U}. \tag{4}$$

Since $\mathbf{G}$ is a *strictly* lower-triangular weight matrix, $(\mathbf{I}-\mathbf{G})^{-1}$ is a lower *uni-triangular* matrix (ones on the diagonal). Note that this is equal to the sum of a geometric series

$$(\mathbf{I}-\mathbf{G})^{-1} = \sum_{k=0}^{n-1} \mathbf{G}^k. \tag{5}$$

It can be seen that element (i,j) represents the cumulative effect of X_j on X_i via all directed paths having length up to $n-1$. The equivalent weight of a directed path from X_j to X_i is the product of the weights of all edges on that path, and the cumulative effect via all the paths is the sum over equivalent weights of distinct directed paths from X_j to X_i. Note that even if some of the nodes are latent confounders is still $(\mathbf{I}-\mathbf{G})^{-1}$ triangular because, by definition, latent confounders do not have ancestors and are first in a topological ordering. Equation 4 represents a system with input $\boldsymbol{U}$, output $\boldsymbol{X}$ and weights $(\mathbf{I}-\mathbf{G})^{-1}$. The covariance matrix of the output is (see details in Appendix A, Equation 10)

$$\mathbf{C}_{\boldsymbol{X}} = \left[(\mathbf{I}-\mathbf{G})^{-1}\right]\, \mathbf{C}_{\boldsymbol{U}}\, \left[(\mathbf{I}-\mathbf{G})^{-1}\right]^{\top}. \tag{6}$$

In this paper we employ the constraint-based causal discovery approach Spirtes et al. [2000] that use conditional independence (CI) tests to learn the underlying causal graph. This approach generally requires assuming the causal Markov property (Definition 1) and faithfulness (Definition 2).

3 A Causal Interpretation of GPT

We describe the masked attention in GPT as a mechanism that infers correlations between tokens of a given sequence, where these correlations are induced by a causal structure underlying the output sequence tokens. We then describe how to learn a causal graph by estimating independence relations.

3.1 A Relation between GPT and SCM-based World Model

Rohekar et al. [2024] derived a causal interpretation of BERT-based models [Devlin et al., 2019]. We follow a similar approach, with several important modifications and extensions, to derive a causal interpretation to GPT. First, unlike BERT-based models, which are pre-trained to predict the input sequence Devlin et al. [2019], GPT is pre-trained to predict the next tokens in the sequence. That is, given an input sequence of tokens, $\{t_0,\ldots,t_{n-1}\}$, GPT predicts tokens $\{\hat{t}_1,\ldots,\hat{t}_n\}$. Hence, an attention matrix $\mathbf{A}$ and the corresponding values matrix $\mathbf{V}$ have n rows corresponding to tokens $\{t_1,\ldots,t_n\}$ and the output embeddings of of these tokens are the rows of matrix $\mathbf{Z}=\mathbf{A}\mathbf{V}$. Thus, Note that $\mathbf{V}=\mathbf{Y}\mathbf{W}_V$, where $\mathbf{W}_V$ is a weight matrix fixed for all input sequences, and $\mathbf{Y}$ is input embedding of a specific sequence tokens. Each column of $\mathbf{W}_V$ can be viewed as an independent vector onto which the input embeddings are projected. That is $\mathbf{V}(i,j)$ is the projection of token t_i input embedding $\mathbf{Y}(i,\cdot)$ on, common to all in-distribution sequences, vector $\mathbf{W}_V(\cdot,j)$. At inference, each attention matrix of the last attention layer, $\mathbf{A}$, is extracted and a lower uni-triangular matrix is calculated, $\mathbf{D}^{-1}\mathbf{A}$, where $\mathbf{D}\equiv\mathrm{diag}(\mathbf{A})$. Then estimate the covariance matrix

$$\mathbf{C} = \left[\mathbf{D}^{-1}\mathbf{A}\right]\left[\mathbf{D}^{-1}\mathbf{A}\right]^{\top}. \tag{7}$$

Note that unlike Rohekar et al. [2024], which proposed $\mathbf{C} = \mathbf{A}\mathbf{A}^\top$ for *unmasked* self-attention, we utilize the triangular form of the masked attention in GPT to revert the attention normalization performed by the softmax and obtain a uni-triangular form. Thus, this covariance matrix allows us to treat properties calculated from different attention matrices in a similar manner. In this paper (Section 3.2 and Section 3.3), properties we calculate are based on p-values when testing conditional independence relations between tokens. Next, following Rohekar et al. [2024] we relate each token to an endogenous node in an SCM, and $\mathbf{C}_{\boldsymbol{U}} = \mathbf{I}$ from the central limit theorem Rohekar et al. [2024]. Thus, equate covariance $\mathbf{C} = \mathbf{C}_{\boldsymbol{U}}$

$$\left[\mathbf{D}^{-1}\mathbf{A}\right]\left[\mathbf{D}^{-1}\mathbf{A}\right]^\top = \left[(\mathbf{I}-\mathbf{G})^{-1}\right]\left[(\mathbf{I}-\mathbf{G})^{-1}\right]^\top, \tag{8}$$

where both $\mathbf{D}^{-1}\mathbf{A}$ and $(\mathbf{I}-\mathbf{G})^{-1}$ are lower uni-triangular matrices, and the (i,j) elements, $\forall i > j$, of these matrices have the same meaning, influence of token/node j on token/node i. Finally, since GPT is pre-trained to predict tokens $\{t_1,\ldots,t_n\}$ given input tokens $\{t_0,\ldots,t_{n-1}\}$, and since the only cross-token influence on embeddings is in the attention layers, the last attention layer captures the causal structure underlying the output tokens. Earlier attention layers transform embeddings of $\{t_0,\ldots,t_{n-1}\}$ to values, $\mathbf{V}$, which are equivalent to instantiations of exogenous variables, $\boldsymbol{U}$ in SCM. This follows from equating Equation 1 and Equation 4, where $\mathbf{D}^{-1}\mathbf{A} = (\mathbf{I}-\mathbf{G})^{-1}$.

In light of the causal interpretation of GPT, one important question is what is the *causal world model* that is supported by the GPT architecture. Often, a single causal structure is assumed to govern a domain. In contrast, the causal world model that is entailed from the causal interpretation of GPT assumes a distinct structural causal model for each in-distribution sequence. Specifically, in a causal world model supported by a GPT with k-heads in the last attention layer, each in-distribution sequence is assumed to be generated by an ensemble of k distinct SCMs.

In addition, for a given head, the causal structure over a sequence of tokens $\{t_1,\ldots,t_n\}$ is equal to the sub-graph over these tokens for all in-distribution extensions of the sequence. That is, given a sequence of tokens $\{t_1,\ldots,t_n\}$ and a corresponding graph structure $\mathcal{G}_n$, observing any next token, t_{n+1}, such that $\{t_1,\ldots,t_n,t_{n+1}\}$ is in-distribution, should not violate causal relations in $\mathcal{G}_n$ and may only reveal relations between tokens $\{t_1,\ldots,t_n\}$ and token t_{n+1}.

3.2 GPT for Zero-Shot Causal Structure Learning

The causal interpretation presented in this paper leads to a view in which each attention module represents associations (correlations) between input tokens that are induced by the underlying causal structure. Although this allows only rung-1 inference in the ladder of causation Pearl and Mackenzie [2018], under certain assumptions, many of the underlying causal relations can be extracted, even in the presence of latent confounders and selection bias Spirtes et al. [2000]. These relations are generally represented in a type of causal structure called partial ancestral graph (PAG) Richardson and Spirtes [2002]. We follow a procedure called ABCD, proposed by Rohekar et al. [2024] with several modifications. First, since the causal (topological) order is given (restricted by the masked attention in GPT) we can apply causal discovery recursively to efficiently learn the causal structure. To this end we call the ICD, iterative causal discovery, algorithm Rohekar et al. [2021] to reconstruct a causal structure in each recursive iteration (Algorithm 1). The result is a PAG structure. Thus, a causal structure for a specific output sequence can be learned in a zero-shot manner directly from the attention matrix in the last layer (in a k-head setting, a set of k causal structures is learned).

3.3 Causal Structure Confidence

In this section we derive a metric that describe how compatible a sequence is with the causal world model implicitly encoded by GPT. Given an output sequence of tokens, $\boldsymbol{S}$, and a causal structure $\mathcal{G}$ recovered from the last attention layer $\mathbf{A}$, can we score the confidence in this causal structure? Recall that in the proposed world model each sequence has its own causal structure, and each causal structure may have latent variables. It is not clear how to calculate likelihood $P(\boldsymbol{S}|\mathcal{G})$. We therefore propose the following approach approach.

A causal structure-learning algorithm performs multiple statistical tests of conditional independence (CI) using the covariance matrix estimated from the attention matrix. These CI tests calculate p-values and compare them against a predetermined threshold of significance level (α). It is important to note that there is a one-to-one correspondence between the results of these CI test and the entailed

causal structure. That is, a causal structure can be represented uniquely by a set of CI tests and their results. Hence, we propose a scoring function based on the distribution of these p-values to evaluate the confidence in a structure learned from a given attention matrix. A complete undirected graph corresponds to lack of knowledge about causal relations. Generally, causal structure-learning algorithms prune edges from this graph based on statistical CI tests between pairs of variables (tokens in our case). The removal of edges between independent variables then may entail causal relations between other variables Zhang [2008]. Let $\boldsymbol{p} = \{p_1, \ldots, p_\ell\}$ a set of all p-values computed as part of causal structure learning. The null-hypothesis is independence, where p-values greater than the significance threshold, α, correspond to edges removed from the complete graph. We denote $\boldsymbol{p}_{ind} = \{p : p \in \boldsymbol{p} \text{ and } p \geq \alpha\}$, and $\boldsymbol{p}_{dep} = \{p : p \in \boldsymbol{p} \text{ and } p < \alpha\}$. Since p-values are uniformly distributed under the null hypothesis, we expect the entropy of p-values corresponding to independence, redundant relations (spurious correlations), H_{ind} to be higher for matrices that correspond to a structure compared to those that do not. In addition, we expect the distribution of p-values smaller than the significance level to be weighted towards zero. Hence, entropy of p-values corresponding to dependence relations, H_{dep} is expected to be lower for matrices that correspond to a structure compared to ones that do not. We therefore define the following confidence score given an attention matrix $\mathbf{A}$,

$$R(\mathbf{A}) = H_{\text{ind}} - H_{\text{dep}}, \tag{9}$$

where $H_{ind} = -\sum_{p \in \boldsymbol{p}_{ind}} p \log p$ and $H_{\text{dep}} = -\sum_{p \in \boldsymbol{p}_{\text{dep}}} p \log p$, are entropy of p-values corresponding to independence and dependence relations, respectively.

4 Experiments and Results

We use an experiment setup in which the world layout and rules are well defined and known but were not used during training with samples from this world (legal samples). We measure how well attention in the learned GPT model represents a causal model using Equation 9, and whether it is correlated with the ability of the model to generate tokens that adhere to the world rules (legal sequences).

4.1 Setup

We examine a GPT model trained by Li et al. [2023], for predicting the next move given a sequence of consecutive moves in the Othello strategy board game. They trained the model on approximately 132,000 real-world sequences, where it is assumed the players played with the intention of winning. No information about the game board layout or game rules was used in their training process. For example, positional encoding was not used. In our experiments we use a test set that is not in-distribution with respect to the training set, but in-distribution with respect to the game rules. As a test set we use 1000 random sequences of legal moves. That is, each sequence consists of moves that adhere to the Othello game rules but without considering any strategy of winning the game as in the training set. In other words, the support of the test distribution is not a subset of the support of the training distribution, $\text{supp}(P_{\text{train}}) \subset \text{supp}(P_{\text{test}})$. This enables evaluating whether the model implicitly encoded the game rules. In Figure 1 we plot the accuracy of the model in generating a legal next move (vertical axis) given a test input sequence having different sizes (horizontal axis). Although the average accuracy of the model is 95% (dashed red line), it is not uniformly distributed across different sequence lengths. For example, given a sequence of 15 moves, GPT generates a legal 16-th move in 88% of the times (adheres to the game board state and rules). It is evident that the accuracy is significantly lower for input sequence lengths in the range $[10, 30]$ (lower than the average 95%). By definition of the Othello game rules, at the beginning of the game there are only four legal moves, and as the game unfolds, the number of possible legal moves increases before finally decreasing again as the number of vacant spaces on the board decreases. It might be that memorization of surface statistics can take place at the beginning and end of the game. We therefore report experiment results for input sequences with sizes in the range $[10, 30]$ (gray area) where the accuracy is lower than the average.

4.2 Legal Predictions vs. Structural Confidence

Is there a relation between the legality of predicted tokens, with respect to a set of world rules, and compatibility of attention matrices with a causal graph? Recall that the model was not trained explicitly to generate legal Othello game moves but rather to predict the next move played by a

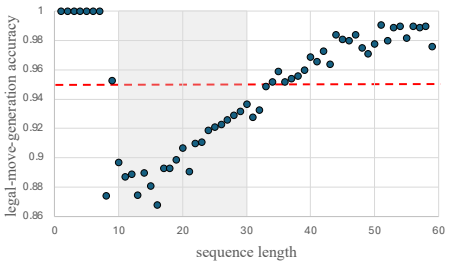

Figure 1: Baseline model accuracy of generating legal Othello game moves. A model trained by Li et al. [2023] on real-world games to predict the next move. Test set consists of randomly generated sequences. Measured accuracy: the percentage of generated moves that are legal according to the Othello game rules. Gray area shows input sequences with sizes in the range $[10, 30]$ where the accuracy is lower than the average of 95% (red dashed line).

human with the intention of winning the game. Moreover, information about the game, such as the existence of a board game and rules, were not provided to the model explicitly Li et al. [2023]. In this experiment we examine whether the cases in which the model generates illegal tokens, are also cases where the causal structure is less distinctive as measured by structural confidence score (Equation 9). In Figure 2 the legal move generation accuracy (vertical axis) monotonically increases with the structural confidence score (horizontal axis) for sequence lengths in $[15, 30]$. For sequence length 10 there is no clear trend. We suspect that for short sequence lengths in the Othello game, memorization of surface statistics enables generating legal tokens with high accuracy.

4.3 Ablation Study

Next we examine if conditional independence tests from which the causal structure is entailed provide an advantage over pair-wise correlations directly represented by elements in the attention matrix. To this end we calculate the confidence score using p-values of a) all pair-wise correlations (from raw attention-matrix elements)—empty conditioning set, b) CI-tests having exactly one node in the conditioning set, c) all CI-tests having empty or exactly one node in the conditioning set, and d) CI-tests used to reconstruct the causal structure without limiting conditioning set sizes. The results are depicted in Figure 3, with corresponding sub-figures. Vertical axis describe the difference between structural confidence averaged over legal and illegal predictions. Error bars indicate 95% confidence intervals. Horizontal axis indicate sequence length. It is evident that relying solely on raw attention values, case a), the difference between legal and illegal generated tokens in not statistically significant, except for sequence length 20. Relying solely on CI-test with exactly one node in the conditioning set, case b), the difference between the structural confidence is positive for all tested sequence lengths but statistically significant only for sequences lengths 17. When employing pair-wise correlations and CI tests with exactly one node in the conditioning tests, the result is statistically significant for both sequence lengths 17 and 20, implying that these two types of tests are complementary. Finally, it is evident that using CI-tests needed to learn the causal graph, without limiting the conditioning set sizes, case d), provide the best results where sequence lengths in $[15, 22]$ are statistically significant and the difference between legal and illegal is positive in all sequence lengths.

5 Conclusions

We presented a causal interpretation of GPT that may explain apparent emergence of world model in recent studies. Following this interpretation, we described a method that utilizes the the triangular form of the attention matrices in GPT to efficiently recover the causal structures for input sequences (zero-shot causal-discovery). Finally, using experiments in the controlled environment of Othello board game we demonstrated that GPT implicitly learns to represent causal structures. Specifically, in cases where the confidence in recovering any structure from the attention matrices is low, GPT generally fails to predict a token that adheres to the Othello board game rules. In future work, these result may provide insights on the sources of hallucination in GPT-based models and may lead to deriving a method to detect them.

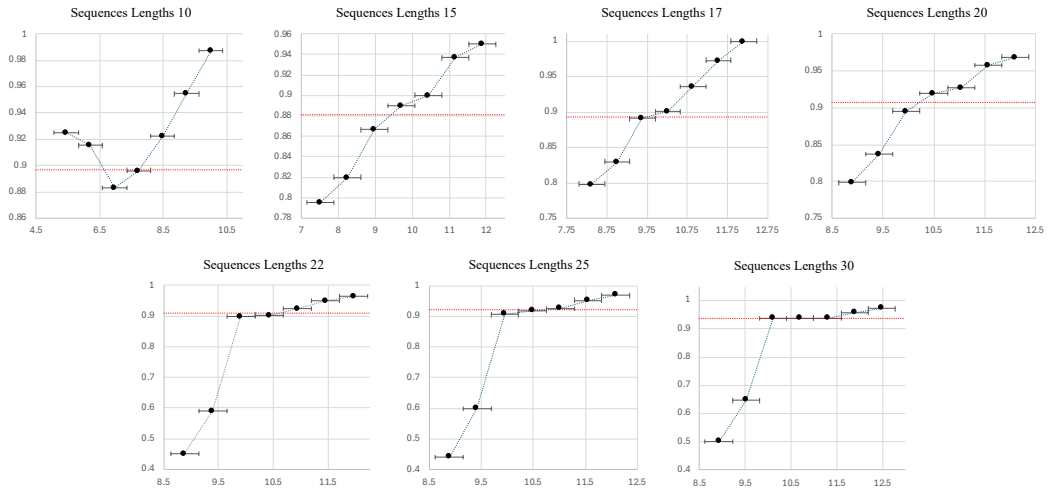

Figure 2: legal move generation accuracy (vertical axis) as a function of structural confidence score R (horizontal axis). Horizontal limits for each point indicates interval of R in which accuracy was averaged. Horizontal dotted red line indicates average accuracy. For sequences of lengths greater than 15 the accuracy increases with the structural confidence score, whereas this trend is not evident for sequences having length 10.

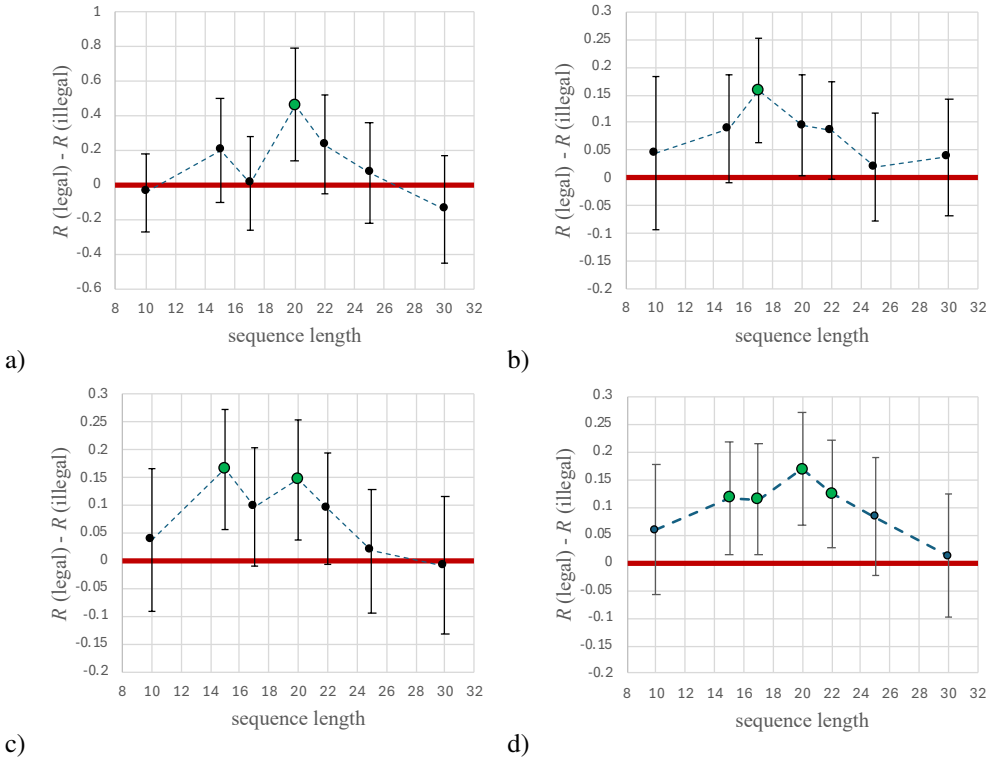

Figure 3: Average difference of structural confidence between legal and illegal move generation (vertical axis) for different input-sequence length (horizontal axis). Error bars are 95% confidence interval. Confidence score are calculated from p-values of: a) all unconditional independence tests calculated directly from raw attention values, b) all CI tests having exactly one conditioning node, c) only tests from cases a) and b), d) CI-tests, without limiting the conditioning set sizes, needed to reconstruct a causal structure.

References

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2024.

Raanan Y Rohekar, Yaniv Gurwicz, and Shami Nisimov. Causal interpretation of self-attention in pre-trained transformers. *Advances in Neural Information Processing Systems*, 36, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. Chess as a testbed for language model state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11385–11393, 2022.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *EMNLP 2023*, page 16, 2023.

Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge university press, second edition, 2009.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction and Search*. MIT Press, 2nd edition, 2000.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

Raanan Y Rohekar, Shami Nisimov, Yaniv Gurwicz, and Gal Novik. Iterative causal discovery in the possible presence of latent confounders and selection bias. *Advances in Neural Information Processing Systems*, 34: 2454–2465, 2021.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

A Additional Preliminaries

In Table 1 we provide common symbols and their meaning used in this paper.

Table 1: Main notations used for the analogy between GPT and attention in SCM. The first set of symbols describes entities in GPT, and the second set of symbols describes entities in SCM.

Symbol	Description
$\boldsymbol{Z}_i$	output embedding of input symbol i, $\boldsymbol{Z}_i \equiv \mathbf{Z}(i, \cdot)$, in attention layer
$\boldsymbol{V}_i$	value vector corresponding to input i, $\boldsymbol{V}_i \equiv \mathbf{V}(i, \cdot)$, in attention layer
$\mathbf{A}$	attention matrix
$\mathcal{T}$	Transformer neural network
$\mathbf{W}_V, \mathbf{W}_{QK}$	learnable weight matrices in GPT
X_i	a random variable representing node i in an SCM
U_i	latent exogenous random variable i in an SCM
$\mathbf{G}$	weighted adjacency matrix of an SCM
$\mathcal{G}$	causal graph (unweighted, directed-graph structure)

A.1 Covariance of Endogenous Nodes in SCM

The covariance matrix of endogenous variables in a linear-Gaussian SCM, as given in Equation 6, is derived as

$$\begin{aligned}\mathbf{C}_{\boldsymbol{X}} &= \mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})^\top] = \\ &= \mathbb{E}[(\mathbf{I} - \mathbf{G})^{-1}\,(\boldsymbol{U} - \boldsymbol{\mu}_{\boldsymbol{U}})(\boldsymbol{U} - \boldsymbol{\mu}_{\boldsymbol{U}})^\top\,((\mathbf{I} - \mathbf{G})^{-1})^\top] = \\ &= [(\mathbf{I} - \mathbf{G})^{-1}]\,\mathbb{E}[(\boldsymbol{U} - \boldsymbol{\mu}_{\boldsymbol{U}})(\boldsymbol{U} - \boldsymbol{\mu}_{\boldsymbol{U}})^\top]\,[(\mathbf{I} - \mathbf{G})^{-1}]^\top = \\ &= \left[(\mathbf{I} - \mathbf{G})^{-1}\right]\,\mathbf{C}_{\boldsymbol{U}}\,\left[(\mathbf{I} - \mathbf{G})^{-1}\right]^\top,\end{aligned} \tag{10}$$

where $\boldsymbol{\mu}_{\boldsymbol{X}} = (\mathbf{I} - \mathbf{G})^{-1}\boldsymbol{\mu}_{\boldsymbol{U}}$.

A.2 Definitions

Definition 1 (Causal Markov) *In a causally Markov graph, a variable is independent of all other variables, except its effects, conditional on all its direct causes.*

Definition 2 (Faithfulness) *A distribution is faithful to a graph if and only if every independence relation true in the distribution is entailed by the graph.*

B Recursive Causal Discovery for GPT

In Algorithm 1 we describe an efficient causal discovery algorithm that utilizes the causal order restricted by GPT by masking attention matrices, forcing them to a lower-triangular form. That is, in a sequence of tokens, $\{t_1, \ldots, t_n\}$, token t_ℓ is not an ancestor of token $t_{\ell-1}$ for all $\ell > 1$. In line 2, the last token is popped from the sequence and placed in t_n resulting in a shorter sequence $\boldsymbol{S}'$. Then, a recursive call is made in in line 3 to learn the structure over tokens in $\boldsymbol{S}'$. Note that since it is ensured that t_n is not an ancestor of any token in $\boldsymbol{S}'$ the skeleton and v-structure relations of $\mathcal{G}'$ is ensured not to be change when adding t_n to the graph Spirtes et al. [2000]. In lines 4–6 token t_n is connected to every node in $\mathcal{G}'$. Finally, using the ICD algorithm Rohekar et al. [2021] edges between t_n and the rest of the graph are learned (removed if conditional independence is found) and the graph is oriented Zhang [2008].

Algorithm 1: Recursive Causal Discovery for GPT

Input: $\boldsymbol{S}$: a sequence of tokens $\{t_1, \ldots, t_n, \}$

Output: $\mathcal{G}$: a partial ancestral graph (PAG)

Function `LearnStructure(`$\boldsymbol{S}$`):`
  $t_n, \boldsymbol{S}' \leftarrow \mathrm{pop}(\boldsymbol{S})$
  $\mathcal{G}' \leftarrow$ `LearnStructure(`$\boldsymbol{S}'$`)`
  $\mathcal{G} \leftarrow \mathcal{G}' + \{t_n\}$
  set $\boldsymbol{E}$ to be the set of edges (circle edge-marks) between t_n and every node in $\mathcal{G}'$
  connect $\boldsymbol{E}$ in $\mathcal{G}$
  test CI for edges in $\boldsymbol{E}$ and orient $\mathcal{G}$ using `ICD` Rohekar et al. [2021]
  return $\mathcal{G}$