CS577 Project Report: Revamping Copyright: Generalized Machine Learning with Open Source NLP

Jasorsi Ghosh
Purdue University
PUID: 34735384
ghosh117@purdue.edu

Preetom Saha Arko
Purdue University
PUID: 34735384
arko@purdue.edu

ABSTRACT

This study investigates the efficacy of generalized machine learning models[29] in navigating copyright[7] constraints within natural language processing tasks. Leveraging both the original Harry Potter corpus[22] and a collection of fan fiction[1], we aim to showcase the models' ability to circumvent copyright limitations. Specifically, we employ the task of summarization of character interactions to demonstrate their performance. Our findings underscore the resilience of these models in transcending legal barriers while maintaining robust functionality. We further establish the provable mathematical framework to support our hypothesis pretext.

1 INTRODUCTION

The Universal Approximation Theorem (Appendix D.) elucidates that machine learning models, particularly neural networks, possess the capability to approximate any continuous function. This theorem underscores the essence of machine learning algorithms as universal function approximates, implying their ability to learn and represent complex relationships between input and output data. In practice, machine learning algorithms iteratively refine their approximated function based on proposals generated by optimizes, aiming to enhance the evaluation score of the approximated function. Ultimately, the process of function approximation is driven by numerical correlations inherent in the data, allowing the model to progressively improve its representation of the underlying relationship between input and output.

In the context of machine learning, causal inference plays a crucial role in understanding and modeling relationships between variables or features in data. This capability is attained through two primary methods: empirical learning, where models discern patterns and correlations from data, and formal causal reasoning, which leverages established principles and methodologies in causal inference. For example, machine learning models can learn from data that certain features are correlated with specific outcomes, but causal reasoning enables us to discern whether these correlations imply causation or are merely coincidental.

Under the context provided in Appendix E, causal systems refer to models or frameworks that capture the causal relationships between variables or features in data. These systems allow us to understand how changes in one variable affect other variables, enabling us to infer causal relationships and make predictions. Causal systems can be used to model structural or hierarchical dependencies by representing the causal relationships between variables in a directed acyclic graph (DAG). In this graph, nodes represent variables, and directed edges represent causal relationships, indicating the direction of influence between variables. By analyzing the structure of the DAG, we can identify direct and indirect dependencies between variables, providing insights into the underlying causal mechanisms governing the system. This formal representation of causal relationships facilitates causal inference and allows us to make predictions and interventions based on causal knowledge.

To broaden the applicability of machine learning models, we employ strategies such as feature deprecation, focusing on essential patterns by removing noise (Appendix A). Markov blankets (Appendix F) identify key variables influencing target outcomes, aiding in understanding complex dependencies. Exploring Markov equivalent (Appendix G) models offers alternative causal structures, enriching the model's ability to approximate various functions efficiently. These approaches collectively enhance model generalization, enabling smoother adaptation to diverse datasets and tasks.

Causal discovery (Appendix H) uncovers the causal structure underlying a system's statistical properties. By analyzing observational data, it identifies causal relationships between variables, revealing direct effects, indirect pathways, and potential confounders. This understanding enables informed decision-making, prediction of intervention outcomes, and optimization of system control strategies. Ultimately, causal discovery bridges the gap between statistical correlations and causal mechanisms, providing deeper insights into complex system behavior.

In Pearl's causal hierarchy (Appendix I), observational data reveals statistical associations between variables, while interventional (Appendix B) data involves actively manipulating variables to uncover causal relationships. Both types of data are essential for causal discovery algorithms to infer the underlying causal structure of a system and construct causal DAGs. Observational data identifies statistical dependencies, while interventional data reveals causal effects, enabling algorithms to discern causality and directionality more effectively.

The ability to bridge the domain gap or distribution shift plays a crucial role in achieving good generalization performance on tasks where publicly available data is used for training on property data. When machine learning models can effectively adapt to changes in data distribution, they demonstrate robust performance when tested on property data. By leveraging causal discovery methods to understand the underlying causal structure, models can adapt more efficiently to diverse datasets, thereby enhancing their ability to generalize well across different domains. This integration of causal inference and distribution shift handling mechanisms empowers models to provide reliable performance, ensuring that insights gained from publicly available data can be effectively applied to property-related tasks leading to law suits.

2 SUMMARY

Here's a point-by-point summary of the project

1. **Causal Mechanism :** We define a causal mechanism M as a function that maps low-level observational data X_{obs} to a set of latent variables Z. This is represented as $M : X_{\text{obs}} \rightarrow Z$, where Z captures the underlying causal relationships inherent in the data. We emphasize the invertibility of this mechanism and explore approximations of linear invertible Structural Causal Models (SCMs) to model complex dependencies in NLP data.
2. **Provable Guarantees Against Interventions:** We aim to establish provable guarantees for interventions on causal systems. Given a causal graph G and a set of observed variables O, we seek to infer the effect of interventions on the outcome variables Y. Formally, this entails finding the causal effect $P(Y \mid \text{do}(X = x))$ under different interventions $\text{do}(X = x)$, where X represents the set of variables being intervened upon.
3. **Our Novel Contribution: Domain Counterfactual Extension:** We extend our framework to handle domain counterfactuals, enabling the modeling of distribution shifts and complex scenarios. This involves defining a counterfactual distribution $P_{\text{counter}}(X)$ that captures the distribution of data under counterfactual interventions, allowing for robust analysis of causal relationships in varying contexts.
4. **Qualitative Analysis of NLP Training Algorithms:** We conduct a qualitative analysis of trending NLP training algorithms, focusing on their ability to capture causal relationships in textual data. This involves evaluating the efficacy of existing methods in modeling causal dependencies and addressing challenges such as sparsity and ambiguity inherent in linguistic data.
5. **Experimental Results:** We present experimental results conducted on the Harry Potter corpus and selected fan fiction pairs, showcasing the effectiveness of our methodologies in real-world scenarios. These experiments demonstrate the practical applicability of our approach in understanding and modeling causal relationships in textual data.

3 METHODOLOGY

3.1 Latent Causal Model

Consider a scenario with linear observations (count p) where $p \geq d$. Let $\hat{G} \in \mathcal{R}^{p \times d}$ be a full-rank matrix, and $X = \hat{G}Z$, where $\hat{G}$ has a generalized pseudo-inverse, and d represents the latent dimension.

Assuming a set of K intervention contexts, denoted as $k \in [K]$, corresponding to interventions on specific variables in the DAG.

Definition 3.1. (Invertible Latent Domain Causal Model): An invertable latent domain causal model $(g, \mathcal{F})$, operates on a shared mixing function $g : Z \longrightarrow X$ on N_d different domains with SCMs $\mathcal{F} \equiv \{f_d : R^m \longrightarrow Z\}_{d=1}^{N_d}$, f_d is invertable autoregressive function, ϵ is the exogenous noise.

Our model adheres to this definition, as the linear relationship and mutually exhaustive exogenous noise lead to a causally sufficient system of SCMs. The invertibility of the f_d function is essential to formulate the domain counterfactual equivariance from injective projection of exhaustive exogenous noise.

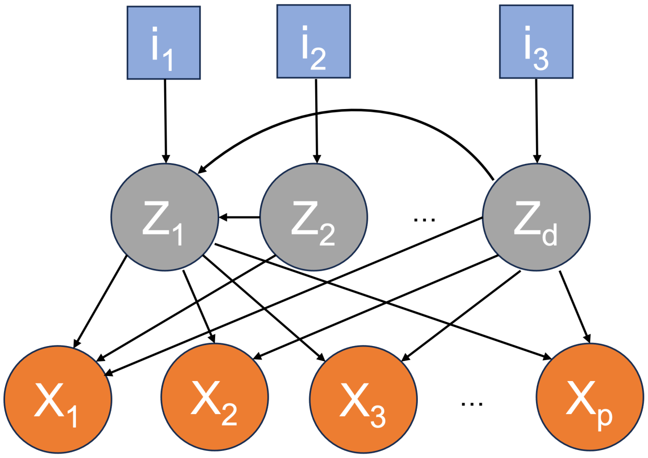


Figure 1: Invertible latent causal model[20], X observed, Z unobserved/latent

Linear invertible causal models (Figure 1) offer a powerful framework for NLP tasks, providing interpretable representations of causal relationships in textual data while maintaining robustness and scalability. By decomposing text into latent variables and modeling causal mechanisms, these models capture complex dependencies and facilitate efficient inference and prediction. Their interpretability and ability to handle uncertainty make them valuable priors for a wide range of NLP applications, enabling insights into language structure and facilitating accurate and scalable text processing.

3.2 Domain Counterfactuals

Definition 3.2. (Distribution Equivalence): Two Invertible Latent Domain Causal Models,$(g, \mathcal{F})$ and $(g‘, \mathcal{F}‘)$ are distributionally equivariant if they induce the same domain distributions.$\forall d, x \in X, J :$ Jacobian of a function

$$P_N\left(f_d^{-1} \circ g^{-1}(x)\right) |J_{f_d^{-1} \circ g^{-1}}(x)| = P_N\left(f‘_d^{-1} \circ g‘^{-1}(x)\right) |J_{f‘_d^{-1} \circ g‘^{-1}}(x)|$$

Causal discovery identifies directed acyclic graphs (DAGs) up to a Markov equivalence class, which entails the same observational distribution. Therefore, it is necessary to formulate a family of models for domain counterfactuals under distribution equivalence.

Definition 3.3. (Domain Counterfactual Equivalenc): Two Invertible Latent Domain Causal Models,$(g, \mathcal{F})$ and $(g‘, \mathcal{F}‘)$ are domain counterfactual equivalent if all domain counterfactuals are equal. $\forall d, d‘$

$$g \circ f_{d‘} \circ f_d^{-1} \circ g^{-1} = g‘ \circ f_{d‘}‘ \circ f_d‘^{-1} \circ g‘^{-1}$$

The current definition for domain counterfactual equivalence only considers the equality of counterfactuals under function composition, overlooking structural aspects like causal ordering and conditional independence. A more comprehensive definition should incorporate criteria ensuring equivalence in both functional transformations and structural dependencies across domains.

Domain counterfactuals represent a formal evaluation framework for testing NLP tasks against a property test dataset, under the condition that the model was solely trained on publicly available data. This evaluation incentivizes achieving robust performance on the property data while adhering to the constraints of training solely on publicly available data. It serves as a rigorous benchmark for assessing the generalization capabilities of NLP models across diverse domains and datasets, emphasizing the importance of real-world applicability and adaptability.

3.3 Domain Counterfactual Extension

However, our formulation currently lacks a direct mechanism for handling domain counterfactuals with provable guarantees. This gap in our approach hinders our ability to achieve true domain-invariant or equivariant distribution shifts for the models. As a result, we have identified macroscopic goals that are essential for addressing this limitation and advancing our research in this direction.

Definition 3.4. (Characterization of Counterfactual Equivalenc): Two ILDs are domain counterfactual equivariant $(g, \mathcal{F})$ and $(g^{‘}, \mathcal{F}^{‘})$

$$(g, \mathcal{F}) \approx (g^{‘}, \mathcal{F}^{‘}) \iff$$

$$\exists h_1, h_2 \in \mathcal{F}_I : g^{‘} = g \circ h_1^{-1} \in \mathcal{F}_I, f_d^{‘} = h_1 \circ f_d \circ h_2 \in \mathcal{F}_A$$

$$\wedge$$

$$|I(\mathcal{F})| = |I(\mathcal{F}^{‘})|$$

where $|I(\mathcal{F})|$ signify the intervension set size, $\mathcal{F}_I$ family of invertable SCMs, $\mathcal{F}_A$ is the family of admissible functions[] restricted to a domain d.

1. h_1 and h_2 must be invertible ($h \in \mathcal{F}_I$): This ensures we can reverse the transformations if needed.
2. The transformed domain-specific models (f_d') must remain admissible ($f_d' \in \mathcal{F}^{‘}_A$): Even after applying h_1 and h_2, the causal relationships within each domain should still be valid.
3. This result implies that to estimate domain counterfactuals, we indeed do not require[1] the recovery of the latent representations or the full causal mode. There could be arbitrarily different latent representations defined $g^{‘} = g \circ h^{-1}$, where h^{-1} can be arbitary invertable function.

3.4 Memorization in NLP models

We reference an intriguing[18] study that aimed to recover text from text embeddings, successfully reconstructing "proper noun" words. This experiment illustrates that standard NLP models tend to memorize training data, relying solely on correlation. This finding emphasizes the importance of shifting focus towards generalized machine learning algorithms that prioritize learning structural relationships.

The paper introduces a two-stage learning objective for recovering text from its embedding. Initially, a text encoder $\phi : V^n \to \mathbb{R}^d$ maps a text sequence of tokens $x \in V^n$ to a fixed-length embedding vector $e \in \mathbb{R}^d$. The goal is to recover the text x given its embedding $e = \phi(x)$, which can be formalized as optimization:

$$\hat{x} = \arg\max_x \cos(\phi(x), e)$$

Subsequently, given a dataset of texts $D = \{x_1, \dots\}$, the objective is to learn to invert the encoder ϕ by estimating a distribution of texts given embeddings, $p(x|e; \theta)$, where θ is learned via maximum likelihood:

$$\theta = \arg\max_{\hat{\theta}} \mathbb{E}_{x \sim D}[p(x|\phi(x); \hat{\theta})]$$

θ are the model parameters to be learned, ϕ is the text encoder function, $\cos(\phi(x), e)$ is the cosine similarity between the embedding $\mathbf{e}$ and $\phi(x)$.

3.5 Experimental Results

3.5.1 Dataset. As proprietary dataset, we use a dataset comprising the original works [22] of J.K. Rowling's "Harry Potter" series for natural language processing (NLP) tasks. In parallel, we contrast this canonical text with a corresponding segment of FanFiction literature [1] that closely mirrors the narrative elements and thematic elements of the original work. This comparison allows for an investigation into the efficacy of NLP techniques when applied to both authentic literary content and its derivative, fan-generated material.

3.5.2 NLP task explained. We build two graphs summarizing the interactions between characters from two different corpora- Harry Potter and Fanfiction. The steps to build such a graph are as follows:

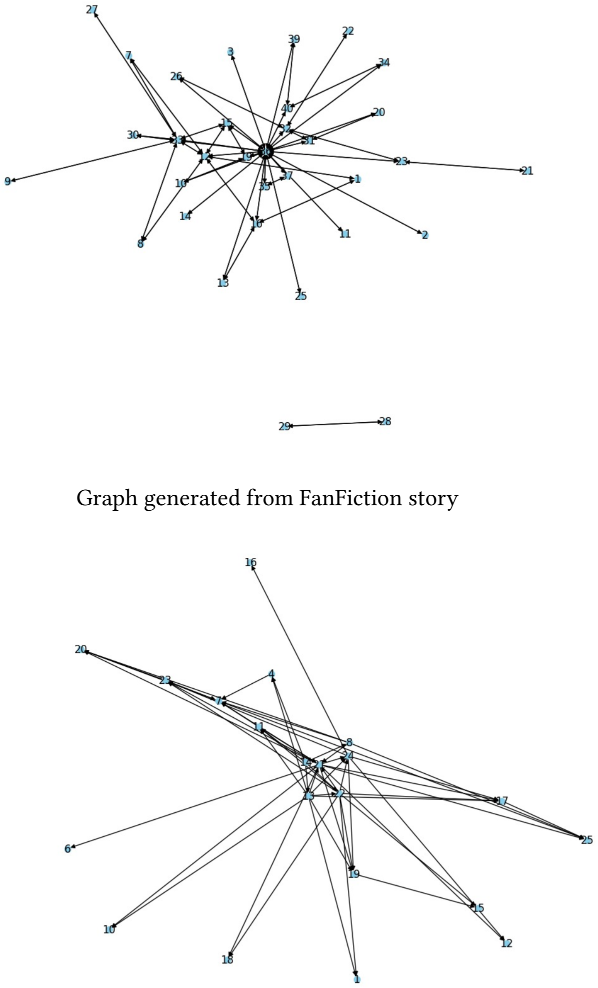
1. Named entity recognition
2. Co-reference resolution for all characters
3. Graph construction where each node corresponds to a character and each edge corresponds to the interaction between a pair of characters. We assume that two characters interact if their positions in the novel are not more than k tokens distant (where k is a hyperparameter). This resembles the idea of using a co-occurrence sliding window. We assign weights to edges according to the number of interactions between each pair of characters.

We choose one story from Fanfiction and another similar story from a chapter of Harry Potter. For named entity recognition, we use Spacy's pretrained en_core_web_lg model. For co-reference resolution, we use the pretrained model of [27] as we have to perform co-reference resolution on a whole story which is usually very long. Even using this approach, we run out of GPU memory if we perform co-reference resolution on the whole story, so we use a cut-down version of the story, both from Harry Potter and Fanfiction. Then we generate one graph summarizing the interactions between characters from each story. Harry Potter stories usually contain more characters than Fanfiction, so we create an induced graph from the generated graph for the Harry Potter story containing only the top n connected nodes, where n is the number of nodes in the graph generated from Fanfiction. This makes the later comparisons fair. Now we have two graphs - one induced graph from Harry Potter and one graph from Fanfiction.

Next we generate graph embedding from each of these two graphs using node2vec [10] algorithm. Then we test for cosine similarity between the two embeddings. If the two embeddings have a high cosine similarity, it can be said that the NLP models have

[1] Theorem 1 from [34]

sufficiently generalized the characters and the interactions between them to such a level that the actual input cannot be identified from the output.



Graph generated from FanFiction story

Induced graph from original Harry Potter story

Figure 2: Character network pair with 82% structural similarity score. Nodes represent characters and edges represent interactions between the characters

3.5.3 Qualitative results. As a proof of concept, we chose a FanFiction story that closely mirrors the first five chapters of the first book of Harry Potter. Due to computation constraints, we consider only the portion that resembles the first three chapters of the first book of Harry Potter for our experiment.

The graph generated from the truncated FanFiction story shows 82% similarity with the graph generated from the first three chapters of the first book of Harry Potter.

This example verifies our claim. Further experiments are needed to validate the claim on more examples.

3.5.4 Practical distribution equivalence. : An approximate formal explanation to the solution can be found in the Appendix K.

4 RELATED WORK

This section can be found at Appendix J.

5 NOVEL IMPACT

(1) Our mathematical formulation explains, the impact of achieving counterfactual performance as assertions of copyright infringement by demonstrating that machine learning models can produce similar results using different datasets. This undermines the notion that proprietary data is indispensable, thereby weakening the basis for making definitive copyright claims.
(2) We introduce the definition of counterfactual equivariance characterization, which is a crucial requirement for achieving true domain counterfactual as future work.
(3) In our project report, we highlight a key observation: strong NLP task performance on property data does not necessarily mean the data was exclusively used for model training. This insight challenges common assumptions about data exclusivity in copyright law.

6 CONTRIBUTION

Jasorsi Ghosh contributed to formulating domain generalization scope of the problem while Preetom Saha Arko integrated with a NLP pipiline. Both Jasorsi Ghosh and Preetom Saha Arko collaborated on literature research, problem definition, coding the model, and analyzed the results.

7 FUTURE WORK

To circumvent the problem of long document coreference resolution, a coreference resolution model built on infini attention [19] can be used. Due to time and computational resource constraints, we could not train any model on the Harry Potter or FanFiction dataset. The models used for named entity recognition and coreference resolution need to be finetuned on our dataset for better results. To get embedding from a graph, a more complex model such as graph neural network (GNN) can be used.

We used only character interactions to compare the similarity between the stories from two corpora. Even to map the interactions between a pair of characters, a simple sliding window approach was used. More complicated approaches can be taken to better represent the interaction between characters and better model the similarity between the stories, probably taking the events into account. Last but not the least, a more comprehensive study on a large number of samples is needed to properly validate the generalization capability of NLP. We did our experiment on a handpicked example as a proof of concept.

REFERENCES

[1] [n. d.]. Works in Harry Potter - J. K. Rowling. Archive of Our Own. https://archiveofourown.org/tags/Harry%20Potter%20-%20J*d*%20K*d*%20Rowling/works

[2] Ryan Abbott and Elizabeth Rothman. 2023. Disrupting Creativity: Copyright Law in the Age of Generative Artificial Intelligence. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4185327. (2023).

[3] Kartik Ahuja, Jason Hartford, and Yoshua Bengio. 2021. Properties from Mechanisms: An Equivariance Perspective on Identifiable Representation Learning. arXiv:2110.15796 [cs.LG]

[4] AI Now Institute. 2023. AI Now Report 2023. https://ainowinstitute.org/wp-content/uploads/2023/04/AI-Now-2023-Landscape-Report-FINAL.pdf.

[5] Blake Brittain. 2023. Judge pares down artists' AI copyright lawsuit against Midjourney, Stability AI. https://www.reuters.com/legal/litigation/judge-pares-down-artists-ai-copyright-lawsuit-against-midjourney-stability-ai-2023-10-30/.

[6] R. Cai, F. Xie, C. Glymour, Z. Hao, and K. Zhang. 2019. Triad constraints for learning causal structure of latent variables. In *Advances in neural information processing systems*, Vol. 32.

[7] CNBC. 2023. In generative AI legal Wild West, the courtroom battles are just getting started. *CNBC* (3 April 2023). https://www.cnbc.com/2023/04/03/in-generative-ai-legal-wild-west-lawsuits-are-just-getting-started.html

[8] Pierre Comon. 1994. Independent component analysis, A new concept? *Signal Processing* 36, 3 (1994), 287–314. https://doi.org/10.1016/0165-1684(94)90029-9 Higher Order Statistics.

[9] F. Eberhardt, C. Glymour, and R. Scheines. 2005. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*. 178–184.

[10] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.

[11] Talika Gupta, Hans Ole Hatzel, and Chris Biemann. 2024. Coreference in Long Documents using Hierarchical Entity Merging. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*. 11–17.

[12] Y. Halpern, S. Horng, and D. Sontag. 2015. Anchored discrete factor analysis. *arXiv preprint arXiv:1511.03299* (2015).

[13] Alexander Hauser and Peter Buhlmann. 2012. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research* 13, 1 (2012), 2409–2464.

[14] A. Hyttinen, F. Eberhardt, and P. O. Hoyer. 2013. Experiment selection for causal discovery. *Journal of Machine Learning Research* 14 (2013), 3041–3071.

[15] Ahmad Jaber, Murat Kocaoglu, Kirthevasan Shanmugam, and Elias Bareinboim. 2020. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *Advances in neural information processing systems*, Vol. 33. 9551–9561.

[16] Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys (CSUR)* 52, 5 (2019), 1–40.

[17] Stephen M. McJohn and Ian McJohn. 2019. Fair Use and Machine Learning. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3406283. *Northeastern University Law Review* (2019).

[18] John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. 2023. Text Embeddings Reveal (Almost) As Much As Text. arXiv:2310.06816 [cs.CL]

[19] Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. 2024. Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention. *arXiv preprint arXiv:2404.07143* (2024).

[20] Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. 2023. Counterfactual Identifiability of Bijective Causal Models. arXiv:2302.02228 [stat.ML]

[21] B. Saeed, A. Belyaeva, Y. Wang, and C. Uhler. 2020. Anchored causal inference in the presence of measurement error. In *Conference on uncertainty in artificial intelligence*. 619–628.

[22] Michael Siebel. 2020. Harry Potter NLP 1. https://siebelm.github.io/Harry_Potter_1/.

[23] R. Silva, R. Scheines, C. Glymour, P. Spirtes, and D. M. Chickering. 2006. Learning the structure of linear latent variable models. *Journal of Machine Learning Research* 7, 2 (2006).

[24] C. Squires and C. Uhler. 2022. Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics* (2022).

[25] Christine Squires, Yuxiang Wang, and Caroline Uhler. 2020. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 1039–1048.

[26] Raghuveer Thirukovalluru, Nicholas Monath, Kumar Shridhar, Manzil Zaheer, Mrinmaya Sachan, and Andrew McCallum. 2021. Scaling within document coreference to long texts. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (2021), 3921–3931.

[27] Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to ignore: Long document coreference with bounded memory neural networks. *arXiv preprint arXiv:2010.02807* (2020).

[28] Thomas Verma and Judea Pearl. 1990. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*. 255–270.

[29] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. 2022. Generalizing to Unseen Domains: A Survey on Domain Generalization. arXiv:2103.03097 [cs.LG]

[30] World Intellectual Property Organization (WIPO). 2023. The State of AI and Intellectual Property. https://www.wipo.int/about-ip/en/frontier_technologies/ai_and_ip.html.

[31] F. Xie, R. Cai, B. Huang, C. Glymour, Z. Hao, and K. Zhang. 2020. Generalized independent noise condition for estimating latent variable causal graphs. In *Advances in Neural Information Processing Systems*, Vol. 33. 14891–14902.

[32] F. Xie, B. Huang, Z. Chen, Y. He, Z. Geng, and K. Zhang. 2022. Identification of linear non-Gaussian latent hierarchical structure. In *International Conference on Machine Learning*. 24370–24387.

[33] Kun Yang, Aaron Katcoff, and Caroline Uhler. 2018. Characterizing and learning equivalence classes of causal DAGs under interventions. In *International Conference on Machine Learning*. PMLR, 5541–5550.

[34] Zeyu Zhou, Ruqi Bai, Sean Kulinski, Murat Kocaoglu, and David I. Inouye. 2024. Towards Characterizing Domain Counterfactuals For Invertible Latent Causal Models. arXiv:2306.11281 [cs.LG]

[35] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. 2021. Contrastive Learning Inverts the Data Generating Process. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 12979–12990. https://proceedings.mlr.press/v139/zimmermann21a.html

A D-SEPRATION IN A CAUSAL DAG

Definition: Let $\mathcal{G}$ be a causal directed acyclic graph (DAG) with nodes $\mathcal{V}$ and directed edges $\mathcal{E}$. Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be disjoint sets of nodes in $\mathcal{G}$.

A set $\mathcal{S}$ of nodes in $\mathcal{G}$ is said to d-separate $\mathcal{X}$ and $\mathcal{Y}$ given $\mathcal{Z}$, denoted as $\mathcal{X} \perp_d \mathcal{Y} \mid \mathcal{Z}$, if for every collider C in $\mathcal{G}$ along any path from a node in $\mathcal{X}$ to a node in $\mathcal{Y}$, either:

(1) $C \in \mathcal{Z}$, or
(2) C is not in $\mathcal{Z}$ and no descendant of C is in $\mathcal{Z}$.

In other words, $\mathcal{S}$ d-separates $\mathcal{X}$ and $\mathcal{Y}$ given $\mathcal{Z}$ if every path from a node in $\mathcal{X}$ to a node in $\mathcal{Y}$ is blocked by $\mathcal{Z}$, except for those paths that pass through a collider C in $\mathcal{G}$ such that neither C nor any of its descendants are in $\mathcal{Z}$.

B HARD SINGLETON INTERVENTION

Definition B.1 (Hard Singleton Intervention). In a causal system $\mathcal{M}$, a *hard singleton intervention* on variable X at value x is denoted by $\mathcal{M}[do(X = x)]$ or $\mathcal{M}[X \leftarrow x]$ and defined as follows:

(1) The value of X is set to x.
(2) All direct causal effects of X are overridden to reflect this new value.
(3) Indirect causal effects and dependencies remain unchanged.

C GRAPH NOTATIONS

(1) Parents of node i are defined by $pa_G(i) := \{j \in G | j \to i \in G\}$, and $\overline{pa_G}(i) := pa_G(i) \cup \{i\}$
(2) Ancestors of $i \ni G$ are defined as $an_G(i) := \{j \in G | k \in \{G, \phi\} | j \to k \to i|\}$ and $\overline{an_G}(i) := an_G(i) \cup \{i\}$
(3) Therefore $\overline{an_G}(I) := \bigcup_{i \in I} \overline{an_G}(i)$
(4) Transative closure of G, denoted by $\overline{G}$, is the DAG with $pa_{\overline{G}}(i) = an_G(i)$. Given a DAG with partial order $\prec_G$ to be $i \prec_G j \iff j \in an_G(i)$.

D UNIVERSAL APPROXIMATION THEOREM

The Universal Approximation Theorem for perceptrons in machine learning formally states that a feedforward neural network with a single hidden layer containing a finite number of neurons (perceptrons) can approximate any continuous function on a compact subset of Euclidean space, given a sufficiently large number of neurons and appropriate choice of activation function.

More formally, let $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ be a non-constant, bounded, and monotonically-increasing activation function (commonly used choices include sigmoid, tanh, and ReLU functions). Then, for any continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and any compact subset $K \subseteq \mathbb{R}^n$, there exists a feedforward neural network with a single hidden layer and activation function Φ that can approximate f arbitrarily well on K, meaning that for any $\epsilon > 0$ there exists a network such that:

$$\sup_{x \in K} |f(x) - F(x)| < \epsilon$$

where $F(x)$ is the output of the neural network with appropriate weights and biases.

In simpler terms, the Universal Approximation Theorem asserts that a neural network with a single hidden layer can approximate any continuous function to an arbitrary degree of accuracy, given a sufficiently large number of neurons and an appropriate choice of activation function.

E CAUSAL INFERANCE

Observed Variables (O): Observed variables, denoted as O, represent the set of variables whose values are directly measured or observed. In a causal system, we can denote the observed variables as $O = \{O_1, O_2, ..., O_n\}$, where n is the total number of observed variables.

Hidden Variables or Confounders (H): Hidden variables, also known as confounders, are unobserved variables that influence both the observed variables and the outcome. We denote the set of hidden variables as $H = \{H_1, H_2, ..., H_m\}$, where m is the total number of hidden variables. Hidden variables are typically not directly measured but have a causal effect on the observed variables and may introduce bias in causal inference.

Structural Causal Model (SCM): A structural causal model is represented by a set of structural equations that describe the causal relationships between variables. Let $V = O \cup H$ be the set of all variables in the system. For each variable V_i in the set V, we have a structural equation of the form:

$$V_i = f_i(Pa(V_i), U_i)$$

where:

- $Pa(V_i)$ represents the parent variables of V_i in the causal graph,
- U_i denotes the set of exogenous variables (i.e., variables not causally influenced by any other variable),
- f_i is a deterministic function representing the causal mechanism governing the relationship between V_i and its parent variables.

Structural Causal Equations: The structural causal equations define the causal relationships between variables in the SCM. Each equation specifies how a variable is determined by its parent variables and exogenous factors. In a structural causal model, the equations take the form:

$$V_i = f_i(Pa(V_i), U_i)$$

where:

- V_i is the variable being determined,
- $Pa(V_i)$ represents the parent variables of V_i in the causal graph,
- U_i denotes the set of exogenous variables,
- f_i is a deterministic function representing the causal mechanism governing the relationship between V_i and its parent variables.

Causal Directed Acyclic Graph (DAG): The causal DAG represents the causal relationships between variables in the system using directed edges without cycles. Formally, a causal DAG is denoted as $G = (V, E)$, where V is the set of variables and E is the set of directed edges representing the causal relationships. Each directed edge $V_i \longrightarrow V_j$ indicates that variable V_i causally influences variable V_j.

In summary, observed variables O are directly measured, hidden variables H are unobserved and influence both observed variables and outcomes, and a structural causal model (SCM) consists of a set of structural equations describing the causal relationships between variables in the system. The causal DAG provides a graphical representation of these causal relationships.

F MARKOV PROPERTY IN A CAUSAL DAG

The Markov property of a node X in a causal Directed Acyclic Graph (DAG) states that the conditional probability distribution of X depends only on its parents in the graph. Mathematically, this can be expressed as:

$$P(X|\text{Non-descendants}(X)) = P(X|\text{Pa}(X))$$

where:

- $P(X|\text{Non-descendants}(X))$ denotes the conditional probability distribution of node X given its non-descendants (nodes not directly or indirectly influenced by X).
- $P(X|\text{Pa}(X))$ denotes the conditional probability distribution of node X given its parents ($\text{Pa}(X)$).

In simpler terms, the Markov property implies that once we know the values of X's parents, the values of other nodes in the graph that are not descendants of X provide no additional information about X. Therefore, X is conditionally independent of its non-descendants given its parents in the causal DAG. This property is fundamental in causal inference as it allows us to infer causal relationships between variables based on their conditional dependencies.

G MARKOV EQUIVALENCE OF CAUSAL GRAPHS

Markov Equivalence of Graphs refers to the concept that two Directed Acyclic Graphs (DAGs) are Markov Equivalent if they encode the same conditional independence relationships among variables.

Formally, let $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ be two DAGs with the same vertex set V. G_1 and G_2 are Markov Equivalent if and only if they have the same skeleton (i.e., the same underlying undirected graph) and the same v-structures (i.e., the same sets of directed triples of nodes).

This equivalence implies that if two DAGs are Markov Equivalent, they represent the same causal relationships among variables, even though their graphical structures may appear different. It highlights the idea that different DAG structures can encode the same causal information, leading to the same conditional independence properties.

H CAUSAL DISCOVERY

Causal discovery, formally represented in mathematical terms, pertains to the identification of causal relationships between variables in a dataset or system. Let $\mathcal{G}$ denote the causal graph representing these relationships, where each node X_i in $\mathcal{G}$ corresponds to a variable and directed edges between nodes indicate causal relationships. The process of causal discovery involves inferring the structure of $\mathcal{G}$ from observational data D. This inference is often formulated as finding the graph $\mathcal{G}$ that best explains the data D under a given criterion or scoring function.

Mathematically, causal discovery aims to find the graph $\mathcal{G}$ that maximizes a scoring function $S(\mathcal{G}, D)$ over all possible graphs, subject to certain constraints or assumptions. The scoring function evaluates the fit of $\mathcal{G}$ to the data D based on statistical criteria or causal principles. Different methods, such as constraint-based, score-based, or hybrid approaches, may be employed to perform causal discovery.

I PEARL'S CAUSAL HIERARCHY

Let $\mathcal{X}$ be the set of all variables in a causal system, and let $\mathcal{G}$ be the set of all possible causal graphs representing causal relationships between variables in $\mathcal{X}$.

1. **Association:** The lowest level of the hierarchy deals with associational relationships between variables, represented by joint probability distributions $P(\mathcal{X})$.
2. **Intervention:** The next level involves interventions, where we can manipulate the values of certain variables to observe the effects on others. This is represented by interventional distributions $P(\mathcal{X} \mid \text{do}(X = x))$, where X is the intervened variable and x is the value at which it is set.
3. **Counterfactuals:** At a higher level, counterfactuals involve reasoning about what would have happened under different conditions. This is represented by counterfactual distributions $P(\mathcal{X} \mid \text{do}(X = x), X' = x')$, where X is the intervened variable, x is the intervention value, and X' is the target variable for counterfactual reasoning.
4. **Structural Causal Models (SCMs):** The highest level of the hierarchy involves structural causal models, which provide a formal representation of causal relationships between variables in terms of structural equations. An SCM consists of a set of equations that describe how each variable is causally influenced by its parents in the causal graph.

In summary, Pearl's causal hierarchy starts with simple associations between variables and progresses to more complex causal reasoning involving interventions, counterfactuals, and structural causal models. This hierarchy provides a systematic framework for reasoning about causality in complex systems.

J RELATED WORK

The related work section is divided into three categories: causal perspective, NLP task perspective, and copyright strikes perspective.

Causal perspective:

1. Identifiable Representation Learning: This area focuses on ensuring learned representations accurately reflect underlying factors. Works like Independent Component Analysis (ICA) [8] establish conditions for identifiability, but often require strong assumptions like independence between latent variables. Recent works by Ahuja et al. [3] and Zimmermann et al. [35] relax these assumptions but don't handle causal relationships between latent variables.
2. Causal Structure Learning: Research in this field investigates methods to learn causal relationships between variables from data. It's established that causal structure can be identified up to a certain point based on available data (interventions) [13, 15, 25, 28, 33]. See Squires & Uhler [24] for a recent review. A key line of work by Eberhardt et al. [9] and Hyttinen et al. [14] characterizes the interventions necessary and sufficient to ensure that the causal structure is fully identifiable. In particular, Eberhardt et al. [9] showed that $d-1$ interventions are in the worst case necessary to fully identify a causal DAG model on d nodes. The current paper extends this line of work to DAG models over latent variables.
3. Learning Latent DAG Models: The task of learning a DAG over latent variables dates back to at least Silva et al. [23]. They introduced the notion of a pure child: an observed variable X_i with only one latent parent, such X_i is also called an anchor [12, 21]. The method of Silva et al. [23] requires that all observed variables are pure children. Recent works relax this assumption by studying the linear non-Gaussian setting, where all latent and observed variables are linear functions of their parents plus independent non-Gaussian noise. For example, Cai et al. [6] propose a method which learns a latent DAG under the assumption that each latent variable has at least two pure children. The pure child assumption can be extended to allow subsets of latent variables with the same observed children, as in Xie et al. [31], which introduces the Generalized Independent Noise condition. This condition was used by Xie et al. [32] to permit latent variables with no observed children; i.e., a hierarchical latent model.

NLP perspective: Character networks [16] have been used on works of fiction for summarization, classification, or role detection. However, usually such documents are very long and require special techniques [27] [11] [26] for coreference resolution using a finite memory. The infini attention mechanism [19] proposed recently has the potential to scale the coreference resolution to arbitrary large texts.

Copyright infringement perspective : Here are some related works you can reference for the section on copyright infringement strikes against machine learning companies using creative property data:

- Legal Cases:
  - Stability AI vs. Midjourney et al. (2023): Briefly discuss this ongoing case where artists sued Stability AI for copyright infringement, alleging the AI art generator used copyrighted elements from their work in its training data. Mention the court's focus on "fair use" as a key factor in determining the outcome [5].
- Academic Articles:
  - "Disrupting Creativity: Copyright Law in the Age of Generative Artificial Intelligence" (2023) [2]: This article explores the legal gray areas surrounding AI's use of copyrighted material and proposes potential solutions for balancing innovation with creator rights.
  - "Fair Use and Machine Learning" (2019) [17]: This paper delves into the concept of fair use in the context of machine learning, analyzing how courts might interpret fair use when copyrighted data is used to train AI models.
- Industry Reports:
  - "The State of AI and Intellectual Property" by The World Intellectual Property Organization (WIPO) (2023) [30]: This report provides a comprehensive overview of the current legal landscape surrounding AI and intellectual property, including copyright concerns.
  - "AI Now Report 2023" by the AI Now Institute [4]: This annual report by a leading AI ethics research group might discuss copyright issues related to AI training data and potential solutions for responsible AI development.

K APPROXIMATE DISTRIBUTIONAL EQUIVALENCE

Notation:

(1) $\mathcal{F}_{\text{equiv}}$: Equivalence class of functions, comprising functions that produce similar outputs for a given input across different data distributions.
(2) f: Function belonging to $\mathcal{F}_{\text{equiv}}$, representing a machine learning model's learned function.
(3) P_i: Data distribution, where i denotes the index of different distributions.
(4) $P_{\text{fanfiction}}$: Joint stationary distribution of linguistic elements present in the collective fan fiction corpus.
(5) P_{Rowling}: Distribution of linguistic elements present in J.K. Rowling's Harry Potter corpus.
(6) w_i: Linguistic element, representing a word or phrase in the corpus.
(7) x: Input to the function f, representing features or data points in a machine learning task.

Consider the scenario,

(1) Equivalence class of functions ($\mathcal{F}_{\text{equiv}}$):
  - $\mathcal{F}_{\text{equiv}}$ comprises functions f such that for any f in $\mathcal{F}_{\text{equiv}}$, the output for a given input remains consistent across different data distributions P_i.
  - Formally, f belongs to $\mathcal{F}_{\text{equiv}}$ if $f(x) \approx f'(x)$ for all x and for all pairs of data distributions P_i and P_j.

(2) Joint stationary distribution of cumulative fan fiction stories ($P_{\text{fanfiction}}$):
  - $P_{\text{fanfiction}}$ represents the joint stationary distribution of all linguistic elements present in the collective fan fiction corpus.
  - Mathematically, $P_{\text{fanfiction}}$ is defined as the probability distribution over the space of all possible linguistic sequences $\{w_1, w_2, \ldots, w_n\}$, where w_i represents a word or phrase in the fan fiction corpus.

(3) J.K. Rowling's Harry Potter as an intervened mechanism:
  - Introducing Rowling's works alters the joint distribution $P_{\text{fanfiction}}$, transforming it into P_{Rowling}, the distribution of linguistic elements present in Rowling's Harry Potter corpus.
  - Mathematically, this intervention can be expressed as $P_{\text{fanfiction}} \rightarrow P_{\text{Rowling}}$, indicating the transition from the distribution of fan fiction to Rowling's corpus.

(4) Generalization of the equivalent function class:
  - Generalization to Rowling's works requires ensuring that $\mathcal{F}_{\text{equiv}}$ encompasses functions capable of producing outputs consistent with Rowling's corpus.
  - This entails establishing distributional invariance between $P_{\text{fanfiction}}$ and P_{Rowling}, i.e., $P_{\text{fanfiction}} \approx P_{\text{Rowling}}$.
  - Mathematically, distributional invariance implies that the probability distributions $P_{\text{fanfiction}}$ and P_{Rowling} exhibit similar statistical properties, enabling models trained on fan fiction data to generalize effectively to Rowling's writings.