Exploring A Bayesian View On Compositional and Counterfactual Generalization

Patrik Reizinger[*1,2] **and Rahul G. Krishnan**[2,3]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Vector Institute, Toronto, Canada
[3]Department of Computer Science, University of Toronto, Toronto, Canada

Abstract

Large models trained on vast data sets can achieve both minimal training and test loss, and, thus, generalize statistically. However, their interesting properties such as good transfer performance or extrapolation concern out-of-distribution (OOD) data. One desired OOD property is compositional generalization, when models generalize to unseen feature combinations. While compositional generalization promises good performance on a wide range of OOD scenarios, it does not account for the plausibility of such combinations. A ubiquitous example is hallucinations in large models. Building on recent advances in Bayesian causal inference, we propose a unified perspective of counterfactual and compositional generalization. We use a causal world model to reason about the plausibility of unseen combinations. By introducing a Bayesian prior, we show that counterfactual generalization is a special case of compositionality, restricted to realistic combinations. This perspective allows us to formally characterize hallucinations, and opens up new research directions to equip generative AI models with a formally motivated "switch" between realistic and non-realistic/creative modes.

1 Introduction

The rise of realistic generative AI models made it simple to fall prey to these models' *realistic* hallucinations, in the form of false statements, image or video "evidence" of events that did not happen [Nie et al., 2023, Anwar et al., 2024]. However, generating fictional data enables creative professionals to tap into a creative well to compose previously unseen scenes or text. Our work does not pass moral judgment on this phenomenon. As realistic hallucinations can have severe implications, we hope to inspire a discourse about the topic by exploring a theoretical description thereof.

Hallucinations often combine plausible pieces of information into an unrealistic combination, e.g., by putting the image of a spaceship into a Medieval codex. The compositionality of large models is studied in recent works [Reizinger et al., 2024, Han and Padó, 2024, Okawa et al., 2023]. The compositional generalization develops theoretical guarantees to ensure generalization to unseen, out-of-distribution (OOD), feature combinations. Importantly, compositional generalization implicitly assumes that all combinations are valid, realistic, and possible. This is rarely the case, nor is it a universal problem—otherwise, fiction could not exist.

Causality [Pearl, 2009] models the world via (a composition of) causal mechanisms, which reflect our beliefs of how the world works. Thus, causality can be thought of as prior knowledge, distilled from and consistent with, the laws of nature.

[*]Work done during an internship at the Vector Institute. Correspondence to patrik.reizinger@tuebingen.mpg.de

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

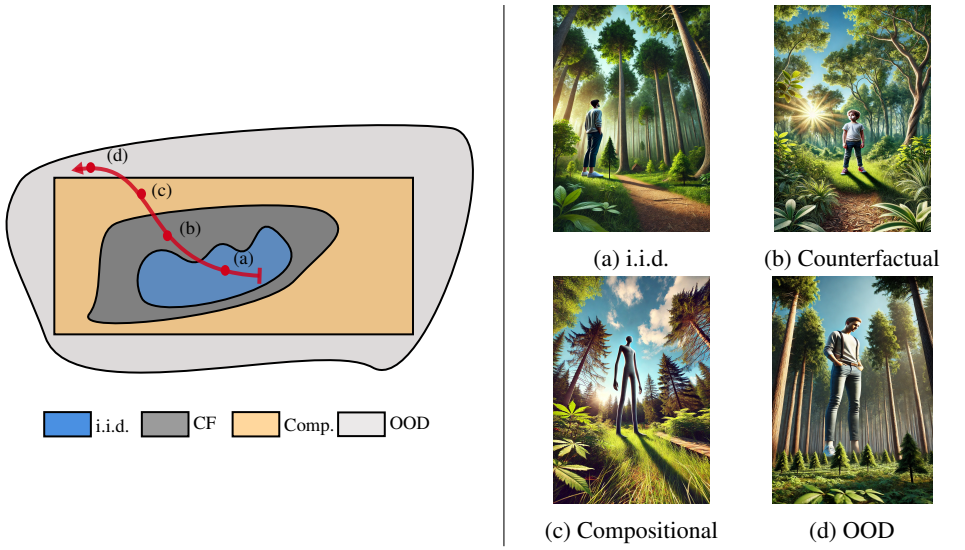

Figure 1: **The relationship of i.i.d. data, counterfactuals, compositions, and OOD data of human weight–height pairs: Left:** a Venn diagram showing the relationship of these concepts with a red arrow indicating a trajectory of changing features, depicted in the four figures on the **Right:** assuming that the data set includes measurements taken on adults, *(a)* an adult, an i.i.d. sample of the data set; *(b)* a child, with a causally plausible combination of weight and height not included in the data set, thus, an OOD sample; *(c)* a "stick man," with a weight–height pair lying in the product space $[W_{\min}; W_{\max}] \times [H_{\min}; H_{\max}]$ of the i.i.d. samples, but a physiologically infeasible one (i.e., the counterfactual assigns a zero probability); and *(d)* a giant with a weight–height combination outside of the support of the product space

We explore how the inherent compositionality of causal mechanisms relates to compositionality, focusing on OOD generalization from a Bayesian perspective. We compare counterfactual and compositional generalization, formulate a Bayesian framework to relate the two, and investigate implications for improving the desired behavior in large AI models. Our **contributions** are:

- We make the first steps towards a Bayesian perspective to unify compositional and counterfactual generalization;
- We discuss the implications for generative models and formally define hallucinations.

2 Background

Compositional generalization. Compositional generalization studies under what assumptions models can generalize to unseen combinations of features such as color and shape. Most works consider computer vision and assume that the independent features form an underlying product space, and each feature's all values (but not all combinations) are observed [Brady et al., 2023, Wiedemer et al., 2023b,a, Lachapelle et al., 2023]. Recent works also started exploring compositional generalization for natural language [Ahuja and Mansouri, 2024, Han and Padó, 2024, Ramesh et al., 2024, Lake and Baroni, 2023, Nogueira et al., 2021, Dziri et al., 2023, Saparov et al., 2023].

Causality and counterfactuals. Causality [Pearl, 2009, Peters et al., 2018] models cause–effect relationships and draws conclusions beyond statistical associations. Structural Equation Models (SEMs) model independent exogenous noise variables (causes) N_i and dependent endogenous causal Z_i variables (effects) via functional mechanisms f_i, i.e., $Z_i := f_i(Pa(Z_i), N_i)$, where $Pa(Z_i)$ denotes the parents of Z_i ($Pa(Z_i) \subset \mathbf{Z}$). Counterfactual queries answer *"What if?"* questions about events that have not occurred. Technically, counterfactual questions ask about the hypothetical value of a specific Z_i given that, all else being equal, the corresponding exogenous variable N_i takes a different value from the observed N_i^0—i.e., it evaluates $N_i^0 \neq \hat{N}_i : f_i(Pa(Z_i), N_i = \hat{N}_i)$

Trustworthiness and robustness in large models. The emergence of large (generative) AI models spurred research to understand why these models work and how can they be made more reliable.

This includes investigations whether these models adhere to causal principles [Jin et al., 2023, Montagna et al., 2024, Willig et al., 2022, Li et al., 2024, Liu et al., 2024, Kasetty et al., 2024], exhibit compositionality [Reizinger et al., 2024, Okawa et al., 2023, Han and Padó, 2024], or are interpretable and trustworthy [Li et al., 2024, Wu et al., 2023, Anwar et al., 2024, Nie et al., 2023].

3 Towards the Bayesian unification of counterfactual and compositional generalization

3.1 Intution

The Independent Causal Mechanisms (ICM) principle [Peters et al., 2018] postulates that causal mechanisms neither inform nor influence each other, which can be interpreted as a notion of *compositionality*. This can be seen from the Markov factorization of the joint distribution over causal variables $p(\mathbf{Z}) = \prod_i p(Z_i|Pa(Z_i), N_i)$, where the probabilistic causal model is composed of causal mechanisms $p(Z_i|Pa(Z_i), N_i)$. The modularity of the causal mechanisms imply that in OOD scenarios, e.g., under distribution shifts, only the affected mechanism needs to be adjusted.

A bivariate case study. To illustrate the relationship between statistical, causal, and compositional generalization, we consider the *simplified* example of weight W and height H (see also Fig. 1). With all else (age, gender, etc.) being equal, we know that for a specific height, there is a corresponding physiologically feasible range of weights. Particularly, a bigger height requires a bigger minimum weight. From the weight, it is not straightforward to draw such conclusions, as a short person can also be extremely overweight. Thus, we model the joint distribution with a causal Markov factorization $p(W, H) = p(W|H)p(H)$ and assume that our data set contains measurements of adults. The modularity implies that if we want to model weight–height relationships in children instead of adults, we only need to adapt $p(H)$. Given a data set of adult weight–height pairs, different generalizations in increasing generality are:

- *Statistical generalization* (Fig. 1a) means generalizing to other (not observed) weight–height pairs of adults, sampled from the same distribution. Statistical generalization cannot capture *realistic* OOD data, e.g., weight–height pairs of children.
- A causal model (Fig. 1b) generalizes to unseen *realistic* OOD combinations, as *counterfactual* queries about the weight and height of children are plausible. However, it cannot capture *unrealistic* OOD combinations.
- *Compositional generalization* (Fig. 1c) generalizes to even the unrealistic OOD combinations, given that they belong to the support of the *observed* product space $[W_{\min}; W_{\max}] \times [H_{\min}; H_{\max}]$, such as a “stick man" character.
- The broadest category is OOD generalization, which includes all weight–height pairs, even outside the support of the product space—e.g., the dwarfs and giants of fairy tales.

The relationship between the counterfactuals and compositionality is *“subset of"* and not *“strict subset of"*—e.g., all possible shape–color combinations of man-made objects are plausible if that is the designer’s intent , such as in dSprites [Watters et al., 2019].

Next, we take a Bayesian perspective in an attempt to formalize these restrictions as a prior expressing the plausible combinations in our causal world. Thus, the difference between compositional and counterfactual generalization becomes a difference of (the support of) priors.

3.2 Towards A Bayesian Framework

Following recent advances in a Bayesian view on causal discovery [Guo et al., 2022] and interventions [Guo et al., 2024], we apply the same lens to counterfactuals, and connect counterfactuals to compositional generalization via a prior expressing our world model rooted in science. In the previous section, we argued that

A key difference between the causal and compositional view on generalization is that counterfactuals assign a non-zero probability only to compositions that adhere to a prior causal world model.

Assume a data distribution $p(x|\theta)$ with a $d-$dimensional parameter θ and corresponding prior $p(\theta)$. In our example in § 3.1, x denotes the weight–height pairs, whereas θ specifies the age, gender, nationality, and, thus, determines the support of $p(x|\theta)$, i.e., the intervals $[W_{\min}; W_{\max}]$, $[H_{\min}; H_{\max}]$. Statistical (i.i.d.) generalization concerns the case when x_{train} and x_{test} have the same θ, i.e., when the prior collapses to a delta distribution $p(\theta) = \delta(\theta - \theta_0)$. In general, OOD generalization concerns $\theta_{\text{train}} \neq \theta_{\text{test}}$, which is only possible for non-delta parameter priors.

Counterfactual generalization requires that the learned model $p(x|\theta)$ correctly models plausible but unseen combinations. That is, $p(x|\theta)$ is non-zero not only for x_{train}, but also for x_{test} sampled from $p(x|\theta_{\text{cf}})$, where θ_{cf} corresponds to our causal prior knowledge of the world, modeled by $p_{\text{cf}}(\theta)$. This prior rules out unplausible combinations, i.e., the $\text{supp}(\theta_{\text{cf}}) \subset \mathbb{R}^d$.

In general, compositional generalization means generalizing over the product of the supports of each latent variable: $\text{supp}(\theta_1) \times \text{supp}(\theta_2) \times \cdots \times \text{supp}(\theta_d)$. It requires each component $\theta_{i,\text{comp}}$ to be in the interval $[\theta_{i,min}; \theta_{i,max}]$, determined by the parameter values in the training set. The product space can include implausible combinations such as the "stick man" in Fig. 1c with a zero counterfactual probability. Thus compositional generalization is more general than counterfactual generalization, i.e., $\text{supp}(\theta_{\text{cf}}) \subset \text{supp}(\theta_{\text{comp}})$. We refer to $p_{\text{comp}}(\theta), p_{\text{cf}}(\theta)$ as *compositional* and *counterfactual* priors and formalize their relationship as:

Proposition 1 (Counterfactual priors are a strict subset of compositional priors)**.** *For real-world scenarios with an underlying causal world model, the counterfactual prior $p_{\text{cf}}(\theta)$ rules out some combinations admitted by the compositional prior $p_{\text{comp}}(\theta)$, i.e.,* $\text{supp}(\theta_{\text{cf}}) \subset \text{supp}(\theta_{\text{comp}})$.

The proof follows indirectly from the above counterexamples, showing that there are parameter configurations which do not adhere to our causal model. However, the two notions can coincide when all combinations are plausible.

Towards a formal definition of hallucinations. Our framework allows us to formally define hallucinations. Namely, if the abstract parameters (e.g., defining the admissible combinations of features) have a zero probability under the causal world model (counterfactual prior), then the resulting sample is a hallucination. Formally:

Definition 1 (Hallucination)**.** *Assume a data distribution $p(x|\theta_{model})$ modeled by a neural network with a corresponding prior distribution $p_{model}(\theta)$. If for any $\theta : p_{model}(\theta) > 0$ the generated sample x_{model} has a zero probability under the counterfactual parameter prior $p_{\text{cf}}(\theta)$, then x_{model} is a hallucination.*

We conclude this section by relating our framework to the performance of state-of-the-art generative models, based on the height–weight example in Fig. 1. We assume access to knowledge about the possible ranges and combinations of the weight and height of humans, represented by $p_{\text{cf}}(\theta)$. We used the DALL-E model available via the free ChatGPT interface (using GPT4o [OpenAI, 2024]) to generate images of a giant. Such a person is physiologically impossible, i.e., $p_{\text{cf}}(\theta_{\text{giant}}) = 0$, thus, by our definition, it is a hallucination. If we render an image of a person who is skinny like a stick man cartoon character, then though $H_{\text{stickman}} \in [H_{\text{min}}; H_{\text{max}}]$ and $W_{\text{stickman}} \in [W_{\text{min}}; W_{\text{max}}]$ but the corresponding parameter has zero density under the counterfactual prior $p_{\text{cf}}(\theta_{\text{stickman}}) = 0$, it is still considered a hallucination, even though this combination is within the domain of compositional generalization.

4 Discussion

Limitations. Our work makes the first step to provide a theoretical connection between compositionality and counterfactuals. Though our contributions can hopefully inspire a discussion about mitigating hallucinations in generative models, our conceptual work has not formalized actionable advice for practitioners.

Consequences for generative models. Compositionality and counterfactuals have merits for different purposes. Enforcing only the realistic combinations is desirable to avoid hallucinations in large models when facts are needed. Whereas the creative use of generative AI necessitates that large models can go into the fictional domain. This can be modeled as a spectrum between facts and fiction, for which, we argue, large models need a "knob" to *intentionally* select the degree of moving beyond reality[2]. Otherwise, not being able to control the consistency between generated samples and observational data can severely limit the practical applicability and trustworthiness of large (generative) models.

Conclusion. We proposed the foundations for a Bayesian framework to unify compositional and counterfactual generalization by introducing a prior that characterizes how realistic a(n unseen)

[2]The closest example we found is the creative/balanced/precise mode selector in Microsoft's Copilot [Microsoft, 2024], though it is unclear whether any of these modes correspond to compositionality or counterfactuals

combination is. In our framework, we can formally define hallucinations in generative models, which, hopefully inspires further research in improving the robustness and trustworthiness of large models by highlighting the key distinction between realistic and fictious samples.

References

Kartik Ahuja and Amin Mansouri. On Provable Length and Compositional Generalization, February 2024. URL http://arxiv.org/abs/2402.04875. arXiv:2402.04875 [cs, stat]. 2

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational Challenges in Assuring Alignment and Safety of Large Language Models, April 2024. URL http://arxiv.org/abs/2404.09932. arXiv:2404.09932 [cs]. 1, 3

Jack Brady, Roland S. Zimmermann, Yash Sharma, Bernhard Schölkopf, Julius von Kügelgen, and Wieland Brendel. Provably Learning Object-Centric Representations, May 2023. URL http://arxiv.org/abs/2305.14229. arXiv:2305.14229 [cs]. 2

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality, 2023. 2

Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. Causal de Finetti: On the Identification of Invariant Causal Structure in Exchangeable Data. *arXiv:2203.15756 [cs, math, stat]*, March 2022. URL http://arxiv.org/abs/2203.15756. arXiv: 2203.15756. 3

Siyuan Guo, Chi Zhang, Karthika Mohan, Ferenc Huszár, and Bernhard Schölkopf. Do Finetti: On Causal Effects for Exchangeable Data, May 2024. URL http://arxiv.org/abs/2405.18836. arXiv:2405.18836 [cs, stat]. 3

Sungjun Han and Sebastian Padó. Towards Understanding the Relationship between In-context Learning and Compositional Generalization, March 2024. URL http://arxiv.org/abs/2403.11834. arXiv:2403.11834 [cs]. 1, 2, 3

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing Causal Reasoning in Language Models. *Advances in Neural Information Processing Systems*, 36:31038–31065, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/631bb9434d718ea309af82566347d607-Abstract-Conference.html. 3

Tejas Kasetty, Divyat Mahajan, Gintare Karolina Dziugaite, Alexandre Drouin, and Dhanya Sridhar. Evaluating Interventional Reasoning Capabilities of Large Language Models, April 2024. URL http://arxiv.org/abs/2404.05545. arXiv:2404.05545 [cs, stat]. 3

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive Decoders for Latent Variables Identification and Cartesian-Product Extrapolation, July 2023. URL http://arxiv.org/abs/2307.02598. arXiv:2307.02598 [cs, stat]. 2

Brenden M. Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, November 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06668-3. URL https://www.nature.com/articles/s41586-023-06668-3. Publisher: Nature Publishing Group. 2

Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. Steering LLMs Towards Unbiased Responses: A Causality-Guided Debiasing Framework, March 2024. URL http://arxiv.org/abs/2403.08743. arXiv:2403.08743 [cs]. 3

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, and Furong Huang. Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey, March 2024. URL http://arxiv.org/abs/2403.09606. arXiv:2403.09606 [cs]. 3

Microsoft. Learn how to use Microsoft Copilot — learn.microsoft.com. https://learn.microsoft.com/en-us/copilot/, 2024. [Accessed 07-09-2024]. 4

Francesco Montagna, Max Cairney-Leeming, Dhanya Sridhar, and Francesco Locatello. Demystifying amortized causal discovery with transformers, May 2024. URL http://arxiv.org/abs/2405.16924. arXiv:2405.16924 [cs, stat]. 3

Allen Nie, Yuhui Zhang, Atharva Amdekar, Chris Piech, Tatsunori Hashimoto, and Tobias Gerstenberg. MoCa: Measuring Human-Language Model Alignment on Causal and Moral Judgment Tasks, 2023. URL https://arxiv.org/abs/2310.19677. 1, 3

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Investigating the limitations of transformers with simple arithmetic tasks, 2021. 2

Maya Okawa, Ekdeep S. Lubana, Robert Dick, and Hidenori Tanaka. Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task. *Advances in Neural Information Processing Systems*, 36:50173–50195, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9d0f188c7947eacb0c07f709576824f6-Abstract-Conference.html. 1, 3

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. [Accessed 07-09-2024]. 4

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2 edition, 2009. ISBN 978-0-511-80316-1. doi: 10.1017/CBO9780511803161. URL http://ebooks.cambridge.org/ref/id/CBO9780511803161. 1, 2

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference: foundations and learning algorithms. *Journal of Statistical Computation and Simulation*, 88(16):3248–3248, November 2018. ISSN 0094-9655, 1563-5163. doi: 10.1080/00949655.2018.1505197. URL https://www.tandfonline.com/doi/full/10.1080/00949655.2018.1505197. 2, 3

Rahul Ramesh, Ekdeep Singh Lubana, Mikail Khona, Robert P. Dick, and Hidenori Tanaka. Compositional Capabilities of Autoregressive Transformers: A Study on Synthetic, Interpretable Tasks, February 2024. URL http://arxiv.org/abs/2311.12997. arXiv:2311.12997 [cs]. 2

Patrik Reizinger, Szilvia Ujváry, Anna Mészáros, Anna Kerekes, Wieland Brendel, and Ferenc Huszár. Position: Understanding LLMs Requires More Than Statistical Generalization. June 2024. URL https://openreview.net/forum?id=pVyOchWUBa. 1, 3

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. Testing the general deductive reasoning capacity of large language models using ood examples, 2023. 2

Nicholas Watters, Loic Matthey, Sebastian Borgeaud, Rishabh Kabra, and Alexander Lerchner. Spriteworld: A flexible, configurable reinforcement learning environment. https://github.com/deepmind/spriteworld/, 2019. URL https://github.com/deepmind/spriteworld/. 3

Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable Compositional Generalization for Object-Centric Learning, October 2023a. URL http://arxiv.org/abs/2310.05327. arXiv:2310.05327 [cs]. 2

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional Generalization from First Principles, July 2023b. URL http://arxiv.org/abs/2307.05596. arXiv:2307.05596 [cs, stat]. 2

Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can Foundation Models Talk Causality?, December 2022. URL http://arxiv.org/abs/2206.10591. arXiv:2206.10591 [cs]. 3

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at Scale: Identifying Causal Mechanisms in Alpaca. *Advances in Neural Information Processing Systems*, 36:78205–78226, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6a8b109d4d4fd64c75e94aaf85d9697-Abstract-Conference.html. 3

A Acronyms

i.i.d. independent and identically distributed

ICM Independent Causal Mechanisms

OOD out-of-distribution

SEM Structural Equation Model