Position: Causality can systematically address the monsters under the bench(marks)

Anonymous Authors[1]

Abstract

Effective and reliable evaluation is essential for advancing empirical machine learning. However, the increasing accessibility of generalist models and the progress towards ever more complex, high-level tasks make systematic evaluation more challenging. Benchmarks are plagued by various biases, artifacts, or leakage, while models may behave unreliably due to poorly explored failure modes. Haphazard treatments and inconsistent formulations of such “monsters” can contribute to a duplication of efforts, a lack of trust in results, and unsupported inferences. In this position paper, we argue causality offers an ideal framework to systematically address these challenges. By making causal assumptions in an approach explicit, we can faithfully model phenomena, formulate testable hypotheses with explanatory power, and leverage principled tools for analysis. To make causal model design more accessible, we identify several useful Common Abstract Topologies (CATs) in causal graphs which help gain insight into the reasoning abilities in large language models. Through a series of case studies, we demonstrate how the precise yet pragmatic language of causality clarifies the strengths and limitations of a method and inspires new approaches for systematic progress.

1. Introduction

Machine learning achievements continue to break records and grab headlines, drawing attention from both the public and the research community. However, the rapid proliferation of powerful models and the increasing complexity of tasks continue to amplify existing challenges in reliable evaluation of these models (Mao et al., 2024). Between inflated expectations (Bubeck et al., 2023; Ullman, 2023; Grace et al., 2024), opaque or misleading assessments (Martínez, 2024), and even the occasional mistake (Chowdhuri et al., 2023), the poor communication (Bowman, 2022) and unreliable benchmarks (Raji et al., 2021; Bowman & Dahl, 2021; Alzahrani et al., 2024) can significantly undermine

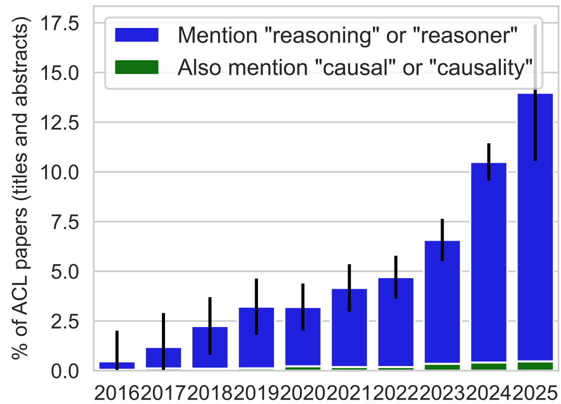


Figure 1. Growth of reasoning papers in ACL Anthology, among which the concept of “causality” is not growing at the same rate, suggesting that NLP is underutilizing causality.

our understanding of the capabilities and limitations of these models (Nezhurina et al., 2024; Yan et al., 2024). This risks a decline of public trust (Bender et al., 2021; Green & Hu, 2018; Hu & Kohler-Hausmann, 2020b) and perhaps even an AI winter. A key issue is that many evaluations focus on performance alone (Liang et al., 2023), failing to account for the reasoning process behind a model’s behavior. For instance, a model may arrive at the right answer for the wrong reasons, making the performance alone an incomplete indication of its capabilities beyond the test set.

To systematically address the challenges in evaluating, in particular, large models, **this position paper argues for a shift toward causality-driven experimental design.** By making causal assumptions explicit, we formulate precise hypotheses and underlying assumptions, diagnose model limitations, and leverage principled tools for analysis.

One subfield that is particularly well-fitted for more causal analyses is the evaluation of reasoning abilities in large language models (LLMs) (Huang & Chang, 2023; Yu et al., 2023). A cursory analysis of the recent NLP papers in the ACL anthology reveals a dramatic rise in the attention in reasoning capabilities of models, as seen in Figure 1. However, curiously, the subset of these papers that mention “causality” or “causal” in the title or abstract is not growing in tandem (yet). In fact, the dendrogram in Figure 2 shows that among

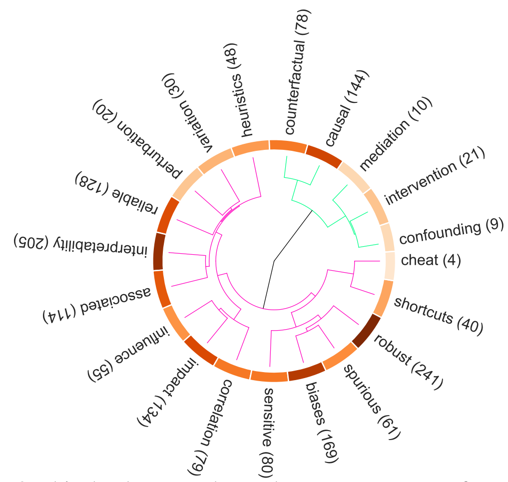


Figure 2. This dendrogram shows the co-occurrences of causal and causality-adjacent terms of papers that contain "reasoning" in the abstracts (total 3181 papers) from the ACL anthology from the past 10 years. The numbers in parentheses indicate the number of papers that mention the term. Note, that the very first split separates all the causality-related terms from the rest of the terms, suggesting relatively poor co-occurrence with other invariably related concepts.

the reasoning papers, causality-related terms tend not to co-occur very much with many non-causal mimics (discussed in Section 2).

Despite many of the issues appearing to be quite disparate based on the distinct terminology that is used, we argue that causality can serve as the framework to systematically study a wide array of issues thus bridging gaps between different subfields and approaches. The expertise required to understand and satisfactorily mitigate these issues is very diverse, and we do not mean to suggest that causality is *all* you need. Rather, we argue that causality already (often implicitly) underlies much of the design, analysis, and interpretation of machine learning experiments, despite the apparent gap in terminology as suggested by Figure 2.

To make an explicit causal framing more accessible and attractive, we develop a guide to common causal structures and then use these abstract topologies to gain insights across case studies about reasoning abilities in LLMs. While we focus on research questions and issues concerning the evaluation of reasoning abilities in LLMs, all four of our main claims (particularly 2-4) largely apply to the whole of empirical machine learning research. Our contributions include:

- We survey a wide variety of common challenges, or *monsters*, that hinder reliable evaluation and discuss how these can faithfully be modeled using causality
- We propose intuitive templates for causal modeling, called Common Abstract Topologies (CATs), and show how these aid in effective model design
- We motivate practical causal models to approaches in several case studies

Prompt: Listen here, I'm starvin' and it's already half past four. Gotta rake in 10 grand in protection dough before I eat. I'm hittin' up every joint in the hood with my bat, collecting an average of 800 bucks a pop. Takes me 10 minutes to move from one mark to the next. Am I gonna get to chow down before 7?
LLM Chatbot: If you're aiming to collect 10 grand by **shaking down joints** for an average of $800 each and it takes you 10 minutes to move from one target to the next, it would take you approximately **3 hours and 20 minutes** (200 minutes) to reach your goal. Starting at half-past four, you should be able to hit your target before 7 PM. So, **yes, you can have a meal before 7** if everything goes as planned. **Enjoy your dinner!**

Example 1. **The final answer is correct, and yet the reasoning is very wrong.** Although this LLM's response is technically correct in a narrow sense, its answer for this GSM8k-like multi-step reasoning question (Cobbe et al., 2021) is unsatisfactory both the purposes of evaluating the model's arithmetic reasoning, as well as commonsense or moral reasoning as the LLM appears not to pick up on the obvious unethical context implied by the scenario.

- We promote causality as a tool to encourage hypothesis-driven research with precise, transparent assumptions and more systematic evaluation

2. A Bestiary of Monsters

As LLMs become an increasingly integral component of decision-making systems, reasoning failures (including seemingly toy problems like in Example 1) may have dire real-world consequences (Nezhurina et al., 2024; Bender et al., 2021). This necessitates more powerful tools to identify, catalog, and address the bestiary of issues that arise in the design as well as evaluation of large models.

Example 1 is indicative of several common reasoning failures in LLMs, and, by implication, our evaluation procedures for addressing such issues. The input prompt is structurally similar to the questions found in GSM8k (Cobbe et al., 2021), a benchmark used for evaluating a model's multi-step arithmetic and commonsense understanding skills. However, we select the subject matter and word choice to evoke a mobster discussing plans to extort money from local businesses. Although the LLM's final answer is factually correct, there are several problems with the rationale: (1) it makes several arithmetic mistakes which happen to cancel out, (2) it fails to pick up on the unethical situation implied by the scenario, and (3) by implicitly condoning the criminal behavior, it does not consider the broader consequences of the response. Crucially, if we only check for correctness, as is standard practice (Huang & Chang, 2023), we would find no fault in the response.

The problem is that to demonstrate good reasoning abilities, a correct answer is insufficient. We need to show that the model answers the question correctly *for the right reasons*. In other words, our evaluation must verify that the model's processing of the input information *leads to* the correct answer consistently and reliably. This criterion makes a *causal* claim about the model's reasoning process, and thus must be supported by a causal analysis.

> **Claim 1:** Evaluating reasoning involves causal inference
>
> A correct answer can be reached through very poor reasoning, but poor reasoning will not generalize beyond the lab bench. To generalize well, the model's reasoning must rely on robustly predictive (i.e. causal) features and relationships rather than spurious ones. Consequently, evaluating the reasoning abilities involves causal inference.

2.1. "Here be dragons" [1]

To get a qualitative sense of the myriad of issues, or *monsters*, that plague our benchmarks and experiments, we will briefly survey recent approaches, including broad overviews into the nature of reasoning tasks (Huang & Chang, 2023; Yu et al., 2023) and the evaluation of LLMs (Mao et al., 2024; Chang et al., 2023; Hajikhani & Cole, 2023). For investigations of more specific issues, we separate efforts into three clusters depending on whether the problem originates with the (1) models, (2) datasets, or (3) evaluation procedures.

Models This line of work focuses on characterizing the reasoning failures and biases of language models, which is nontrivial given their opaque behavior (Binz & Schulz, 2023). These failures range from well-defined formal errors such as logical fallacies (Jin et al., 2022), red herrings (Naeini et al., 2023), or invalid inferences (Saparov & He, 2023) to broader issues including sensitivity to superficial features (Hajikhani & Cole, 2023; Ullman, 2023), overconfidence (Nezhurina et al., 2024), hallucinations (Dziri et al., 2022; Cui et al., 2023), and lack of robustness (Zheng et al., 2024; Wang et al., 2023; Jin et al., 2020). Some studies explore how models exhibit "content effects" (Poesia et al., 2023), absorbing and amplifying human biases (Dasgupta et al., 2022; Zečević et al., 2023) including social and cultural biases (Bender et al., 2021; Messner et al., 2023; Hutchinson et al., 2020; Vig et al., 2020; Cao et al., 2023b; AlKhamissi et al., 2024; Motoki et al., 2024), such as stereotyping (Kotek et al., 2023).

[1]The heir of vagueness and discomfort that researchers frequently use when mentioning potential undesirable biases or systematic limitations in their analysis is not unlike the way medieval cartographers would fill the mysterious edges of their maps with dragons.

Datasets Meanwhile, subtle variations of popular benchmarks, such as premise order in reasoning tasks(Chen et al., 2024) or minor changes in problem parameters (Mirzadeh et al., 2024; Wu et al., 2024), can cause large performance drops (Nezhurina et al., 2024; Yan et al., 2024), raising concerns not just about whether models genuinely reason (Zhou et al., 2024), but also about exploitable issues in the training data and benchmarks (Rogers & Rumshisky, 2020; Bowman & Dahl, 2021). These are can be described as enabling cheating (Zhou et al., 2023), heuristics (McCoy et al., 2019), or shortcuts (Branco et al., 2021; Li et al., 2022; Marconato et al., 2023), possibly due to sampling biases (Razeghi et al., 2022) or in certain cases even leakage between the training and testsets (Zhou et al., 2023) which can result in memorization (Feldman, 2021). Poor dataset construction can lead to annotation artifacts (Gururangan et al., 2018; Fleisig et al., 2024) such as priming effects (Gardner et al., 2021), which degrade the quality and reliability of results (Byrd & Srivastava, 2022) while also unintentionally reinforcing social biases or cultural inequities (Bender et al., 2021; Hu & Kohler-Hausmann, 2020a; Naous et al., 2024).

Evaluation Even with well-constructed datasets, evaluation methodologies can introduce systematic errors (Dominguez-Olmedo et al., 2024) or lead to misleading conclusions (Bowman, 2022). For example, automated scoring systems can obscure obvious failures (Chowdhuri et al., 2023), while static benchmarks can emphasize surface-level accuracy at the cost of other important factors, such as generalization (Liang et al., 2023) or interpretability (Loftus, 2024) or social costs (Raji et al., 2021; Bender et al., 2021). While standardized leaderboards (Beeching et al., 2023) and evaluation procedures (Srivastava & et al.) can enable more direct model comparisons, these benchmarks can gradually become less representative of real-world tasks (Schlangen, 2019; Alzahrani et al., 2024; Shirali et al.; Kiela et al., 2021), introduce biases that favor certain model families (Zhang et al., 2024b), or inadvertently leak information from the test set (Zhou et al., 2023) which can be difficult to detect due to closed-source models and proprietary datasets (Mao et al., 2024).

Despite the diverse, at times redundant, terminology, we observe certain structural similarities in the approaches of these contributions. Terms like "ablation", "perturbations", "edits", "flips", "masking" can often be interpreted as interventional or counterfactual analyses, while "sensitivity"/"robustness", "consistency", "shortcut", "leakage", "bias", etc. refer to how the model's behavior is impacted by, for example, (seen or unseen) confounders.

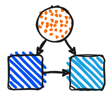
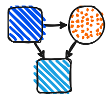
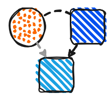
Name	Graph	Example Phenomena
Confounding		• prompt wording, instruction tuning, or prompting strategies • dataset sourcing, annotation artifacts, missing context • overlap or leakage between the benchmark and training data
Mediation		• circuit analysis such as mechanistic interpretability • tool use or integrating an LLM in a larger application • editing individual tokens or ablating model parameters
Spurious Correlations		• social and cultural biases in the data collection process • imbalances in the surface form such as symbol or label bias • variable selection and construction

Table 1. Some simple Common Abstract Topologies (CATs) which can be used to formalize a wide variety of *monsters* both known and unknown that may lurk in a benchmark or experiment analysis and some example issues that they may help represent. For the graphs, is the independent variable, is the dependent/outcome variable, and represents a third variable factor such as a confounder or mediator. Note that the examples are partially overlapping, reflecting that depending on the specific setting, a similar issue may be represented by different CATs or combinations thereof.

Claim 2: The monsters are causal

The disparate and often vague formulations of the issues that lurk within our benchmarks and models such as biases or failure modes can often faithfully be described in terms of causality. Whether the factors are known or unknown, their influences can be captured by an appropriate causal model to guide the experimental design and analysis.

3. Common Abstract Topologies

Coming up with a causal graph that faithfully represents the underlying structure of an experiment or data generating process can be very challenging. Especially, since usually when we design an experiment, we think in terms of more vague concepts like independent, dependent, and controlled variables, and consequently only implicitly make causal assumptions. However, explicit causal graphs:

- precisely communicate the assumptions that go into an approach, experiment, or analysis
- leverage the machinery of causal inference for a more principled analysis
- understand the implications of our design choices including the particular strengths and limitations on both technical and conceptual levels

To help make the process of constructing a causal graph more accessible and systematic, we identify some Common Abstract Topologies (CATs) of causal graphs and discuss associated phenomena (see Table 1) in the context of evaluating reasoning abilities in large models where these structures may be useful.

However, there may be some hesitancy to commit to a specific causal graph that faithfully captures all the factors that may affect the analysis (Bareinboim et al., 2022). Especially since, in practice, the graph is often severely underdetermined by available data, or depends on precise definitions or interpretations of relevant factors. As pointed out by Loftus (2024), researchers may even avoid causal language because it offers more assumptions for reviewers to challenge.

Claim 3: Instrumentalism is all you need

A causal model does not need to be perfect to be useful. Plausible simplifying assumptions and abstractions can yield valuable insights and motivate practical experiments. As research advances, the model can be refined to mark our progress, while providing transparent falsifiable hypotheses at every step of the way.

Here we urge the community to be more pragmatic, much like Loftus (2024); Janzing & Garrido (2022). Due to subtle differences in the model design such as variable construction or selection, the same issue may be represented by various causal models, perhaps even ones that appear incompatible. For example, depending on the level of abstraction (Chalupka et al., 2016; Rubenstein et al., 2017; Beckers et al., 2020), certain causal relationships may be omitted, and the graph may be simplified or augmented with additional variables. Nevertheless, as long as a proposed causal model does not directly conflict with the available data, it may be sufficient to improve performance or produce insights (such as more interpretable or explainable models).

Aside from the additional explanatory power, if a more formal treatment is necessary or desired, there is a whole world of tools and techniques to explore. The field of causal inference (Pearl, 2009; 2020; Imbens & Rubin, 2015; Peters et al., 2017; Bareinboim et al., 2022) has developed a language for formalizing the effects of subtle design choices and their, potentially counterintuitive, consequences for the analysis. For example, Simpson's paradox can be elegantly explained, to "resolve" the apparent paradox based on the appropriate causal assumptions of the problem (for a deep dive into this topic see Pearl (2022) and Chapter 6 of Pearl (2020)).

Claim 4: Towards *explicit* causal assumptions

An experimental design involves a variety of assumptions about what factors matter, how they interact, and how this relates to the proposed approach. Here the language of causality provides a powerful framework for motivating an approach, precisely formulating the hypothesis, and answering questions in a principled way.

Causal inference is valuable not only for formal analysis but also as a conceptual framework for understanding the structural assumption behind an approach or argument. By making the concepts and tools of causal inference more accessible, we aim to develop a practical guide to recognize familiar causal structures in common phenomena, as well as build an intuition for the implications of model design choices on analysis and interpretation. To this end, we present three simple CATs that correspond to the three causal interpretations of a statistical dependence between two variables according to Reichenbach's common cause principle (Reichenbach, 1956).

3.1. Confounding

Confounding occurs when there is a common cause between the independent and dependent variables. For our purposes, we further restrict the "confounding" CAT to the case where the confounder is known and can, in principle, be controlled for. This is in contrast to the "spurious correlation" CAT, where the confounder is unknown or too complex to be modeled explicitly.

Confounding makes evaluation difficult or unreliable because the observed statistical relationship between the stimulus and response is not representative of the underlying causal relationship, thus unbiased causal effect estimation necessitates controlling for the confounder.

3.2. Mediation

Another important type of causal topology is mediation, where there are multiple causal paths between the stimulus and response. For simplicity, we illustrate this general structure with one direct causal link and one that goes through a mediator variable. Mediation analysis is often used to quantify the impact of subcomponents or side-effects on the model's behavior. For example, a common setting may be to study the impact of a specific prompting strategy or representation on the model's response, which can be modeled as mediation as in Figure 3.

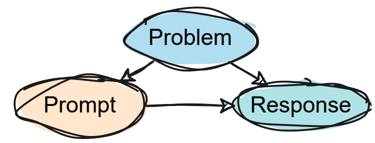


Figure 3. Sketch of a conceptual causal model treating the prompt (i.e. surface form) as a mediator between the underlying problem or task of interest and the model's response.

The impacts of the individual causal paths can be studied by estimating the natural direct effect (NDE), natural indirect effect (NIE), or controlled direct effect (CDE) (Pearl, 2009). However, notably controlling for the mediator is not always appropriate, for example, for estimating the total causal effect (TCE). This underscores one of the key benefits of causal inference: given the specific causal query, the appropriate analysis method is dictated by the graph structure, thereby prescribing specific and principled experiments.

3.3. Spurious Correlations

The final common pattern we discuss here is spurious correlations, which are closely related to confounding but differ in the interpretation and implications for analysis. Spurious correlations (depicted as a dashed curved edge) are statistical associations between variables that are not causally related (neither is an ancestor of the other), but are correlated due to some external factor (a common cause), which is usually unknown.

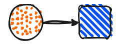
If a model is only trained on observational data (as is almost always the case) as opposed to interventional or counterfactual data, then there is no way to differentiate a spurious correlation from a causal relationship. Consequently, a model can learn to rely on spurious correlations in the data to make predictions, effectively forming an undesirable causal link between the spurious feature and the model's output.

A common cause of spurious correlations, particularly in datasets, is due to selection bias in the data generative process, which may also be described as a collider bias (Pearl, 2009). Generally, it is not feasible to entirely eliminate spurious correlations, as seemingly innocent choices in variable construction and selection are invariably informed by the experimenter's biases (Hu & Kohler-Hausmann, 2020a; Pietsch, 2015). Nevertheless, there is extensive causal inference machinery to address spurious correlations depending on the specific setting (Plecko & Bareinboim, 2023).

4. Case Studies

In this section, we discuss a variety of specific research projects which either make use of one of the Common Abstract Topologies (CATs) or could benefit from a more *explicitly* causal framing.

4.1. Confounding



One project that explicitly uses a causal framing to understand the biases in the text generation of large language models (LLMs) is Xia et al. (2024). To address confounding due to biases in the training data or prompt, they propose using a reward model as an instrumental variable.

Zhang et al. (2024a) formulate a human-LM collaborative writing setting as a causal inference problem where the past human commands and LLM responses are confounders for the current command and the overall interaction outcome. To identify strategies that improve the collaboration, they introduce a new causal estimand, the Incremental Stylistic Effect (ISE), which allows them to abstract away from specific interactions and focus on how actions incrementally contribute to the desired stylistic outcome of the text.

Meanwhile, a good example of an active area of research that largely revolves around the confounding CAT, despite "confounding" rarely being mentioned explicitly, is the study of how the mathematical reasoning abilities of LLMs are affected by various undesirable factors (Zhou et al., 2024; Patel et al., 2021). In particular, a variety of projects have focused on using the dataset GSM8K (Cobbe et al., 2021) to evaluate multi-step arithmetic reasoning as well as common sense understanding (Mirzadeh et al., 2024; Chen et al., 2024; Zhang et al., 2024b).

Several of these projects probe the robustness of the LLM's reasoning ability by systematically varying certain features such as the subjects or numbers involved (Mirzadeh et al., 2024), the order of the premises (Chen et al., 2024), or attempt to replicate the original data generative process (Zhang et al., 2024b) to test whether LLMs have overfit to the original dataset.

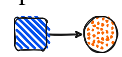
While these projects generally suggest that LLMs are sensitive to these factors, a more causal treatment can provide more precise conclusions. Let's take a closer look at one of the projects with a relatively specific target: Razeghi et al. (2022) investigate how much a language model's performance on quantitative reasoning tasks is affected by how often the numbers in the question occur in the model's training dataset. An intuitive causal framing for their approach using the "confounding" CAT is shown in Figure 4a. Note, that here the model's response is abstracted away since we are only interested in the response in so far as it affects the resulting accuracy.

Alternative Approach Here it is instructive to consider a hypothetical project where we design a benchmark to evaluate the math skills of a language model. Much like in Razeghi et al. (2022), our questions take the form "What is n_1 times n_2?" where n_1 and n_2 are numbers selected by some sampling strategy. However, we do not consider the training dataset of the model at all, and instead of following Razeghi et al. (2022), we sample numbers uniformly, which effectively removes the causal link between the term frequency and the numbers used in the question. Based on the findings of Razeghi et al. (2022), we can expect to find a substantial correlation between the presence of certain numbers in the question and the model's accuracy, even though the rules of arithmetic are obviously entirely agnostic to which numbers are used.

To explain the results of our approach, we might phenomenologically define a new property of numbers called "difficulty" which, we conclude significantly affects the model's accuracy, leading to the causal graph in Figure 4b using the "spurious correlation" CAT.

Verifying Causal Assumptions A notable consequence of committing to a causal graph as in Figure 4a is that it may imply certain falsifiable causal relationships that are not actually verified by the experiments. Specifically, the experiments of Razeghi et al. (2022) identify a significant correlation between the term frequency and the model's accuracy, rather than showing a causal relationship, as the authors helpfully state explicitly. Therefore, an alternative plausible causal graph as in Figure 4c may be posited for their approach where the term frequency is merely correlated with the model's accuracy by sharing a hither-to unknown confounder. This process illustrates how structurally distinct causal interpretations can be proposed to motivate certain experiments or approaches, and then how the results can be used to incrementally refine the causal graph.

4.2. Mediation



Mediation analysis guides the approaches of mechanistic interpretability (Stolfo et al., 2023; Gupta et al., 2023; Meng et al., 2023; Wang et al., 2022), but it is also useful in augmentation of language models (Mialon et al., 2023), embedding LLMs within larger programs (Schlag et al., 2023). and the quantification of biases like, gender bias (Vig et al., 2020).

A common setup for mechanistic interpretability is to study the impact of a specific component, such as an attention head or even a single parameter on the model behavior. Olsson et al. (2022) propose that transformers can learn simple, interpretable algorithms called "induction heads," which they hypothesize significantly contribute to in-context learning abilities. While mediation analysis is not explicitly used in their work, we can frame their approach as studying a mediation graph, where the tendency for a given model architecture (stimulus) to exhibit in-context learning (response) is mediated by induction heads. Their six supporting arguments can be interpreted through this causal lens: argu-

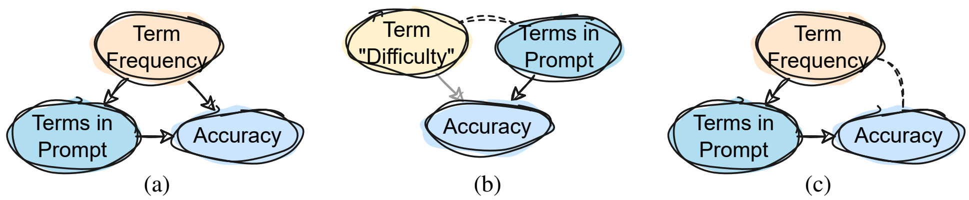
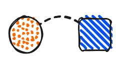


Figure 4. Various causal framings based on the approach of Razeghi et al. (2022). (a) A simple interpretation of their approach using the confounding CAT. (b) A causal framing for an alternative approach where we do not consider the term frequency, and instead observe a spurious correlation. (c) A more cautious causal framing that combines CATs to avoid claiming that the term frequency causally affects the model's accuracy (as is consistent with the authors' approach).

ments 1 and 2 establish links between stimulus, mediator, and response through co-occurrence and co-perturbation; argument 3, an ablation study, resembles controlled direct effect estimation; and arguments 4-6 examine the causal influence of the mediator on the response. This framing also highlights potential limitations, particularly regarding unmeasured confounders that could affect causal interpretations, as the authors' "pattern-preserving" ablation does not fully isolate the induction heads' effect. By considering mediation explicitly, we can better understand the underlying assumptions in their analysis and identify areas for further investigation, such as quantifying the natural indirect effect to understand the full impact of the induction heads on in-context learning abilities.

In contrast, Stolfo et al. (2023) propose a method for mechanistic interpretability of arithmetic reasoning in LLMs by editing the model's parameters to characterize the information flow in the network. Note that the level of abstraction for this approach is quite different from the causal model we proposed for Olsson et al. (2022), as the focus is on how information flows between individual model subcomponents, rather than how specific subcomponents affect the overall model's behavior.

4.3. Spurious Correlations



There are several recent projects that use causal models to characterize spurious correlations in, for example, factual knowledge (Cao et al., 2023a), multi-modal models for fake news detection (Chen et al., 2023), or to avoid spurious features by designing strategies for finding useful demonstrations in few-shot learning (Zhang & Yu, 2023) or control NLP classifiers (Bansal & Sharma, 2023).

Chen et al. (2023) develop a causal model to systematically quantify and remove two specific kinds of bias: psycholinguistic (use of emotional language) and image-only (ignoring text features). Note that the assumptions of the causal model address very specific types of bias using both interventional and counterfactual techniques.

Bansal & Sharma (2023) presents a particularly interesting case as it addresses the same issue as Gardner et al. (2021), but from a causal perspective. They both study the issue of label bias, specifically in "competency problems" (Gardner et al., 2021), where an individual token in the prompt is not indicative of the label, but the model learns to rely on it, usually due to selection bias in the data collection.

The authors of Gardner et al. (2021) propose a mitigation strategy based on "local edits" to individual tokens in the prompt to debias the benchmark. Using their statistical framing, the authors prove that the most promising strategy must apply local edits such that the label is flipped precisely half of the time.

Translating this into a causal framing, we can recover the same result quite intuitively. Adopting the same terms as Gardner et al. (2021), we now treat the input (text) features X as the stimulus, the model's response Y as the response, and the individual token X_i as the third variable, which our model has learned to rely on despite it being a spurious feature. Now, to remove the label bias for our model, we need the effect of an edit on $X_i = x_i'$ to be as likely to flip the label as not. This is equivalent to the average causal effect conditioned on X:

$$\mathbb{E}(Y|X, do(X_i = x_i')) - \mathbb{E}(Y|X, do(X_i = x_i)) = 0 \quad (1)$$

However, due to the non-causal treatment Gardner et al. (2021), need to make a "strong independence assumption," which is equivalent to, for the purposes of the mitigation strategy, assuming that the individual token X_i is completely independent of the prompt X. As the authors point out, this assumption is not very realistic, as changing a single token may well affect the semantic meaning of the prompt beyond just the label (e.g. replacing "very" with "not" in a movie review).

Meanwhile, Bansal & Sharma (2023) uses a causal graph matching the spurious correlation CAT and a condition analogous to Equation 1 to derive a causal regularization term for the model's training objective - without the need for the strong independence assumption.

In summary, both approaches started with the same objective, but due to the purely statistical treatment, a cumbersome derivation still required an unrealistic assumption severely limiting the applicability of the method. The causal model not only provided a more intuitive motivation for the approach, but also offered a more powerful, principled method for achieving the same goal.

5. Alternative Views

We are hardly the first to point out systematic shortcomings of evaluation methodology, particularly in NLP. One existing perspective focuses on improving the external validity of benchmarks to ensure that high performance on a benchmark actually translates to improved capabilities in the real world, such as with common sense reasoning (Elazar et al., 2021), or more precisely defining LLMs (Rogers & Luccioni, 2024) and how tasks relate to specific cognitive capabilities (Schlangen, 2019). Raji et al. (2021) argue that the common practice for certain "standard" benchmarks to become proxies for testing complex, high-level abilities, such as natural language understanding (NLU) leads to vague or unreliable results, while Rogers & Rumshisky (2020) connect this to a proliferation of low-quality datasets.

Precisely this issue, that "benchmarking for NLU is broken" (Bowman & Dahl, 2021), can be addressed using causality. Not only does a causal framing provide a versatile way to define the underlying assumptions and design choices of a benchmark, but it also offers principled methods for evaluating the benchmark's external validity (Bareinboim & Pearl, 2012; Pearl & Bareinboim, 2022).

In the context of evaluating the reasoning abilities of language models, a natural field to turn to is psychometrics, which has been studying the evaluation of human reasoning abilities for over a century (Wilhelm, 2005). This direction also coincides with an increasing practice in Natural Language Processing (NLP) to treat language models as agents (Park et al., 2023; Liu et al., 2023) or subjects in the social sciences (Horton, 2023; Leng & Yuan, 2023; Pellert et al., 2024). Specifically, item response theory (Lord & Novick, 2008; Baker, 2001) holds promise to develop tools to systematically quantify what information about the model's reasoning abilities can be extracted from a benchmark with respect to some population candidate models, and there are some projects applying this framework in the context of NLP (Rodriguez et al., 2021). Within the field of NLP there are also notable calls for more holistic evaluation schemes (Liang et al., 2023; Bowman & Dahl, 2021; Zhang et al., 2024b) and practical tools for improving the evaluation of language models (Ribeiro et al., 2020; Srivastava et al.; Alzahrani et al., 2024) or even reintroducing principles from linguistic theory (Lan, 2023).

There is also a growing interest in studying the causal knowledge learned by language models (Zhang et al., 2023; Kıcıman et al., 2023) and their causal reasoning abilities (Jin et al., 2024; Zečević et al., 2023; Liu et al., 2024) to help with causal discovery (Montagna et al.; Jiralerspong et al., 2024) or even hypothesis generation in psychology (Tong et al., 2024). This effort largely coincides with our message: just as an LLM may benefit from more explicit causal models, so can the research community.

6. Conclusion

The burgeoning research on large models, and, in particular, high-level reasoning tasks, faces a variety of challenges, or *monsters*, to reliably evaluate and improve models. Despite the wide variety of approaches and frameworks that have been developed to tackle these challenges, this variety obscures their shared structural features and recurring issues. By recognizing that monsters can often be effectively formulated in terms of causal assumptions underlying an experimental design or data generation process, we can unify our understanding using the language of causality.

A causal framing aids along several steps of the research process by guiding experimental design, formulating testable hypotheses, and interpreting results. Causal methods enable researchers to gain a clearer lens to understand how variables of interest interact, rather than merely optimizing for predictive performance on an artificial benchmark. We argue that causality offers a path toward deeper scientific insights, more transparent communication of assumptions, and stronger justifications for the conclusions drawn.

One stumbling block to adopting causal methods is that the restrictive assumptions and formalism may seem unapproachable at first. Additionally, researchers may hesitate to commit modeling assumptions to paper where they can be scrutinized. However, data-driven approaches which rely on implicit or vague assumptions along with results that may (inadvertently) be *interpreted* as causal contribute to confusion and unsupported claims, which hinder scientific progress. Causal methods, by contrast, encourage explicit modeling and critical thinking about the mechanisms that underlie empirical observations.

To make causality more accessible and practically applicable, we introduce Common Abstract Topologies (CATs) to faithfully describe the underlying structure of many issues that arise in designing and evaluating ML models. In the case studies in Section 4, we have shown how a causal framing can formalize a various common issues and help develop mitigate them. We envision CATs as a practical guide, helping researchers quickly identify relevant causal models and choose appropriate inference tools. Ultimately, causal models encourage more hypothesis-driven research which directly tackle key questions in a principled, transparent way, leading to more robust progress across empirical machine learning.

References

Language models and linguistic theories beyond words. *Nature Machine Intelligence*, 5(7):677–678, July 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00703-8.

AlKhamissi, B., ElNokrashy, M., AlKhamissi, M., and Diab, M. Investigating Cultural Alignment of Large Language Models, July 2024.

Alzahrani, N., Alyahya, H. A., Alnumay, Y., Alrashed, S., Alsubaie, S., Almushaykeh, Y., Mirza, F., Alotaibi, N., Altwairesh, N., Alowisheq, A., Bari, M. S., and Khan, H. When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards, February 2024.

Baker, F. B. *The basics of item response theory*. ERIC, 2001.

Bansal, P. and Sharma, A. Controlling Learned Effects to Reduce Spurious Correlations in Text Classifiers, June 2023.

Bareinboim, E. and Pearl, J. Transportability of Causal Effects: Completeness Results. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1):698–704, 2012. ISSN 2374-3468. doi: 10.1609/aaai.v26i1.8232.

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. On Pearl's Hierarchy and the Foundations of Causal Inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition, March 2022. ISBN 978-1-4503-9586-1.

Beckers, S., Eberhardt, F., and Halpern, J. Y. Approximate causal abstractions. In *Uncertainty in artificial intelligence*, pp. 606–615. PMLR, 2020.

Beeching, E., Fourrier, C., Habib, N., Han, S., Lambert, N., Rajani, N., Sanseviero, O., Tunstall, L., and Wolf, T. Open llm leaderboard. Available at `https://huggingface.co/spaces/open-llmleaderboard/open_llm_leaderboard`, 2023.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445922.

Binz, M. and Schulz, E. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, February 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2218523120.

Bowman, S. R. The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail, March 2022.

Bowman, S. R. and Dahl, G. E. What Will it Take to Fix Benchmarking in Natural Language Understanding?, October 2021.

Branco, R., Branco, A., António Rodrigues, J., and Silva, J. R. Shortcutted Commonsense: Data Spuriousness in Deep Learning of Commonsense Reasoning. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1504–1521, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.113.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of Artificial General Intelligence: Early experiments with GPT-4, March 2023.

Byrd, M. and Srivastava, S. Predicting Difficulty and Discrimination of Natural Language Questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 119–130, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.15.

Cao, B., Tang, Q., Lin, H., Han, X., and Sun, L. Does the Correctness of Factual Knowledge Matter for Factual Knowledge-Enhanced Pre-trained Language Models? In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2327–2340, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.143.

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., and Hershcovich, D. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study, March 2023b.

Chalupka, K., Eberhardt, F., and Perona, P. Multi-level cause-effect systems. In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, volume 51 of *JMLR Workshop and Conference Proceedings*, pp. 361–369. JMLR.org, 2016. URL `http://proceedings.mlr.press/v51/chalupka16.html`.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. A Survey on Evaluation of Large Language Models, December 2023.

Chen, X., Chi, R. A., Wang, X., and Zhou, D. Premise Order Matters in Reasoning with Large Language Models, May 2024.

Chen, Z., Hu, L., Li, W., Shao, Y., and Nie, L. Causal Intervention and Counterfactual Reasoning for Multi-modal Fake News Detection. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 627–638, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.37.

Chowdhuri, R., Deshmukh, N., and Koplow, D. No, GPT4 can't ace MIT. https://flower-nutria-41d.notion.site/No-GPT4-can-t-ace-MIT-b27e6796ab5a48368127a98216c76864, June 2023.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training Verifiers to Solve Math Word Problems, November 2021.

Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., and Yao, H. Holistic Analysis of Hallucination in GPT-4V(ision): Bias and Interference Challenges, November 2023.

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Creswell, A., Kumaran, D., McClelland, J. L., and Hill, F. Language models show human-like content effects on reasoning, July 2022.

Dominguez-Olmedo, R., Hardt, M., and Mendler-Dünner, C. Questioning the Survey Responses of Large Language Models, February 2024.

Dziri, N., Milton, S., Yu, M., Zaiane, O., and Reddy, S. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5271–5285, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.387.

Elazar, Y., Zhang, H., Goldberg, Y., and Roth, D. Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema, October 2021.

Feldman, V. Does Learning Require Memorization? A Short Tale about a Long Tail, January 2021.

Fleisig, E., Blodgett, S. L., Klein, D., and Talat, Z. The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2279–2292, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.126.

Gardner, M., Merrill, W., Dodge, J., Peters, M. E., Ross, A., Singh, S., and Smith, N. A. Competency Problems: On Finding and Removing Artifacts in Language Data, December 2021.

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., and Brauner, J. Thousands of AI Authors on the Future of AI, January 2024.

Green, B. and Hu, L. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the Machine Learning: The Debates Workshop*, 2018.

Gupta, A., Mondal, D., Sheshadri, A. K., Zhao, W., Li, X. L., Wiegreffe, S., and Tandon, N. Editing Common Sense in Transformers, October 2023.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017.

Hajikhani, A. and Cole, C. A Critical Review of Large Language Models: Sensitivity, Bias, and the Path Toward Specialized AI, July 2023.

Horton, J. J. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, January 2023.

Hu, L. and Kohler-Hausmann, I. What's Sex Got To Do With Fair Machine Learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 513–513, January 2020a. doi: 10.1145/3351095.3375674.

Hu, L. and Kohler-Hausmann, I. What's Sex Got To Do With Fair Machine Learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 513–513, January 2020b. doi: 10.1145/3351095.3375674.

Huang, J. and Chang, K. C.-C. Towards Reasoning in Large Language Models: A Survey, May 2023.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. Social Biases in NLP Models as Barriers for Persons with Disabilities, May 2020.

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, USA, 2015. ISBN 0521885884.

Janzing, D. and Garrido, S. H. Phenomenological Causality, November 2022.

Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment, April 2020.

Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R., and Schölkopf, B. Logical Fallacy Detection, December 2022.

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Adauto, F. G., Kleiman-Weiner, M., Sachan, M., and Schölkopf, B. CLadder: Assessing Causal Reasoning in Language Models, January 2024.

Jiralerspong, T., Chen, X., More, Y., Shah, V., and Bengio, Y. Efficient Causal Graph Discovery Using Large Language Models, February 2024.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality, April 2023.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., and Williams, A. Dynabench: Rethinking Benchmarking in NLP, April 2021.

Kotek, H., Dockum, R., and Sun, D. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, pp. 12–24, New York, NY, USA, November 2023. Association for Computing Machinery. ISBN 979-8-4007-0113-9. doi: 10.1145/3582269.3615599.

Leng, Y. and Yuan, Y. Do LLM Agents Exhibit Social Behavior?, December 2023.

Li, S., Li, X., Shang, L., Dong, Z., Sun, C., Liu, B., Ji, Z., Jiang, X., and Liu, Q. How Pre-trained Language Models Capture Factual Knowledge? A Causal-Inspired Analysis, March 2022.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. Holistic Evaluation of Language Models, October 2023.

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., and Tang, J. AgentBench: Evaluating LLMs as Agents, August 2023.

Liu, X., Feng, Y., and Chang, K.-W. CASA: Causality-driven Argument Sufficiency Assessment, April 2024.

Loftus, J. R. Position: The Causal Revolution Needs Scientific Pragmatism. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 32671–32679. PMLR, July 2024.

Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores*. IAP, 2008.

Mao, R., Chen, G., Zhang, X., Guerin, F., and Cambria, E. GPTEval: A Survey on Assessments of ChatGPT and GPT-4, December 2024.

Marconato, E., Teso, S., Vergari, A., and Passerini, A. Not All Neuro-Symbolic Concepts Are Created Equal: Analysis and Mitigation of Reasoning Shortcuts, December 2023.

Martínez, E. Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*, March 2024. ISSN 1572-8382. doi: 10.1007/s10506-024-09396-9.

McCoy, T., Pavlick, E., and Linzen, T. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and Editing Factual Associations in GPT, January 2023.

Messner, W., Greene, T., and Matalone, J. From Bytes to Biases: Investigating the Cultural Self-Perception of Large Language Models, December 2023.

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., and Scialom, T. Augmented Language Models: A Survey, February 2023.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, October 2024.

Montagna, F., Cairney-Leeming, M., Sridhar, D., and Locatello, F. Demystifying amortized causal discovery with transformers. URL http://arxiv.org/abs/2405.16924.

Motoki, F., Pinho Neto, V., and Rodrigues, V. More human than human: Measuring ChatGPT political bias. *Public Choice*, 198(1-2):3–23, January 2024. ISSN 0048-5829, 1573-7101. doi: 10.1007/s11127-023-01097-2.

Naeini, S., Saqur, R., Saeidi, M., Giorgi, J., and Taati, B. Large Language Models are Fixated by Red Herrings: Exploring Creative Problem Solving and Einstellung Effect using the Only Connect Wall Dataset, November 2023.

Naous, T., Ryan, M. J., Ritter, A., and Xu, W. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models, March 2024.

Nezhurina, M., Cipolina-Kun, L., Cherti, M., and Jitsev, J. Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models, June 2024.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative Agents: Interactive Simulacra of Human Behavior, April 2023.

Patel, A., Bhattamishra, S., and Goyal, N. Are NLP Models really able to Solve Simple Math Word Problems?, April 2021.

Pearl, J. Causal inference in statistics: An overview. 2009.

Pearl, J. *Book of Why*. Basic Books, New York, reprint edition edition, August 2020. ISBN 978-1-5416-9896-3.

Pearl, J. Comment: understanding simpson's paradox. In *Probabilistic and causal inference: The works of judea Pearl*, pp. 399–412. 2022.

Pearl, J. and Bareinboim, E. External Validity: From Do-Calculus to Transportability Across Populations. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36, pp. 451–482. Association for Computing Machinery, New York, NY, USA, 1 edition, March 2022. ISBN 978-1-4503-9586-1.

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., and Strohmaier, M. AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. *Perspectives on Psychological Science*, pp. 17456916231214460, January 2024. ISSN 1745-6916. doi: 10.1177/17456916231214460.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017. ISBN 978-0-262-03731-0.

Pietsch, W. Aspects of theory-ladenness in data-intensive science. *Philosophy of Science*, 82(5):905–916, 2015.

Plecko, D. and Bareinboim, E. A Causal Framework for Decomposing Spurious Variations, June 2023.

Poesia, G., Gandhi, K., Zelikman, E., and Goodman, N. D. Certified Reasoning with Language Models, June 2023.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. AI and the Everything in the Whole Wide World Benchmark, November 2021.

Razeghi, Y., Logan IV, R. L., Gardner, M., and Singh, S. Impact of Pretraining Term Frequencies on Few-Shot Reasoning, May 2022.

Reichenbach, H. *The direction of time*, volume 65. Univ of California Press, 1956.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond Accuracy: Behavioral Testing of NLP models with CheckList, May 2020.

Rodriguez, P., Barrow, J., Hoyle, A. M., Lalor, J. P., Jia, R., and Boyd-Graber, J. Evaluation Examples are not Equally Informative: How should that change NLP Leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346.

Rogers, A. and Luccioni, A. S. Position: Key Claims in LLM Research Have a Long Tail of Footnotes, June 2024.

Rogers, A. and Rumshisky, A. A guide to the dataset explosion in QA, NLI, and commonsense reasoning. In Specia, L. and Beck, D. (eds.), *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pp. 27–32, Barcelona, Spain (Online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.coling-tutorials.5.

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. ID 11, August 2017. URL http://auai.org/uai2017/proceedings/papers/11.pdf. *equal contribution.

Saparov, A. and He, H. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought, March 2023.

Schlag, I., Sukhbaatar, S., Celikyilmaz, A., Yih, W.-t., Weston, J., Schmidhuber, J., and Li, X. Large Language Model Programs, May 2023.

Schlangen, D. Language Tasks and Language Games: On Methodology in Current Natural Language Processing Research, August 2019.

Shirali, A., Abebe, R., and Hardt, M. A Theory of Dynamic Benchmarks. URL http://arxiv.org/abs/2210.03165.

Srivastava, A. and et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. URL http://arxiv.org/abs/2206.04615.

Srivastava, S., B, A. M., P V, A., Menon, S., Sukumar, A., T, A. S., Philipose, A., Prince, S., and Thomas, S. Functional Benchmarks for Robust Evaluation of Reasoning Performance, and the Reasoning Gap. URL http://arxiv.org/abs/2402.19450.

Stolfo, A., Belinkov, Y., and Sachan, M. A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis, October 2023.

Tong, S., Mao, K., Huang, Z., Zhao, Y., and Peng, K. Automating psychological hypothesis generation with AI: When large language models meet causal graph. *Humanities and Social Sciences Communications*, 11 (1):896, July 2024. ISSN 2662-9992. doi: 10.1057/s41599-024-03407-5.

Ullman, T. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks, March 2023.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., Huang, J., Singer, Y., and Shieber, S. Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias, November 2020.

Wang, H., Ma, G., Yu, C., Gui, N., Zhang, L., Huang, Z., Ma, S., Chang, Y., Zhang, S., Shen, L., Wang, X., Zhao, P., and Tao, D. Are Large Language Models Really Robust to Word-Level Perturbations?, September 2023.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 small, November 2022.

Wilhelm, O. Measuring reasoning ability. In *Handbook of Understanding and Measuring Intelligence*, pp. 373–392. January 2005. ISBN 978-0-7619-2887-4. doi: 10.4135/9781452233529.n21.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks, March 2024.

Xia, Y., Yu, T., He, Z., Zhao, H., McAuley, J., and Li, S. Aligning as Debiasing: Causality-Aware Alignment via Reinforcement Learning with Interventional Feedback. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4684–4695, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.262.

Yan, Q., He, X., Yue, X., and Wang, X. E. Worse than Random? An Embarrassingly Simple Probing Evaluation of Large Multimodal Models in Medical VQA, October 2024.

Yu, F., Zhang, H., Tiwari, P., and Wang, B. Natural Language Reasoning, A Survey, May 2023.

Zečević, M., Willig, M., Dhami, D. S., and Kersting, K. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal, August 2023.

Zhang, B., Wang, Y., and Dhillon, P. S. Causal Inference for Human-Language Model Collaboration, March 2024a.

Zhang, C., Bauer, S., Bennett, P., Gao, J., Gong, W., Hilmkil, A., Jennings, J., Ma, C., Minka, T., Pawlowski, N., and Vaughan, J. Understanding Causality with Large Language Models: Feasibility and Opportunities, April 2023.

Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., Zhao, T., Raja, P., Slack, D., Lyu, Q., Hendryx, S., Kaplan, R., Lunati, M., and Yue, S. A Careful Examination of Large Language Model Performance on Grade School Arithmetic, May 2024b.

Zhang, R. and Yu, T. Understanding Demonstration-based Learning from a Causal Perspective. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1465–1475, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.125.

Zheng, C., Zhou, H., Meng, F., Zhou, J., and Huang, M. Large Language Models Are Not Robust Multiple Choice Selectors, February 2024.

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., and Han, J. Don't Make Your LLM an Evaluation Benchmark Cheater, November 2023.

Zhou, Z., Liu, S., Ning, M., Liu, W., Wang, J., Wong, D. F., Huang, X., Wang, Q., and Huang, K. Is Your Model Really A Good Math Reasoner? Evaluating Mathematical Reasoning with Checklist, October 2024.