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

Anonymous Author(s)

Affiliation

Address

email

Abstract

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

1 Introduction

18 Machine learning achievements continue to break records and grab headlines, drawing attention from
19 both the public and the research community. However, the rapid proliferation of powerful models
20 and the increasing complexity of tasks continue to amplify existing challenges in reliable evaluation
21 of these models [1]. Between inflated expectations [2–4], opaque or misleading assessments [5], and
22 even the occasional mistake [6], the poor communication [7] and unreliable benchmarks [8–10] can
23 significantly undermine our understanding of the capabilities and limitations of these models [11, 12].
24 This risks a decline of public trust [13–15] and perhaps even an AI winter [16]. A key issue is that
25 many evaluations focus on performance alone [17], failing to account for the reasoning process
26 behind a model’s behavior. For instance, a model may arrive at the right answer for the wrong reasons,
27 making the performance alone an incomplete indicator of its capabilities.

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

32 One subfield that is particularly well-fitted for more causal analyses is the evaluation of reasoning
33 abilities in large language models (LLMs) [18, 19]. A cursory analysis of the recent NLP papers in
34 the ACL anthology reveals a dramatic rise in the attention to the reasoning capabilities of models, as
35 seen in Figure 1a. However, curiously, the subset of these papers that mention “causality” or “causal”
36 in the title or abstract is not growing in tandem (yet). In fact, the dendrogram in Figure 1b shows

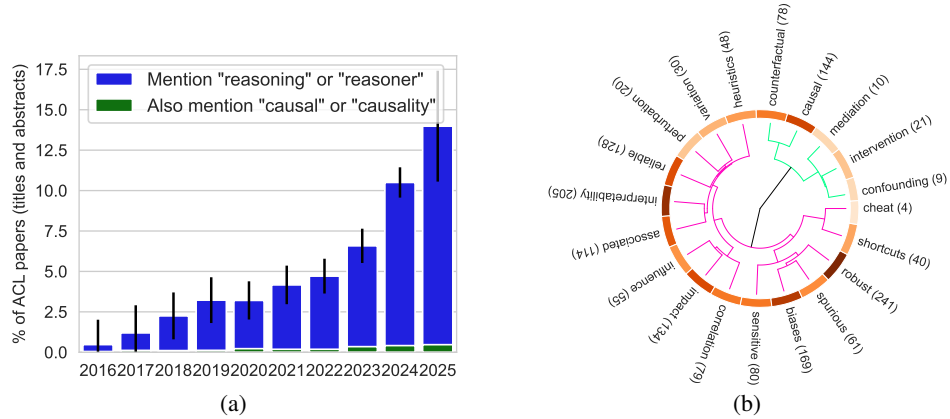


Figure 1: (a) Growth of reasoning papers in the ACL Anthology, among which the concept of “causality” is not growing at the same rate, suggesting that NLP is underutilizing causality. (b) This dendrogram shows the co-occurrences of causal and causality-adjacent terms of papers that contain “reasoning” 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 related concepts.

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

Although these issues often appear disparate, 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 1b.

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
- 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 [11, 13]. 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 [20], 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

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 [20] is unsatisfactory for 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.

68 mistakes which happen to cancel out, (2) it fails to pick up on the unethical situation implied by the
69 scenario, and (3) by implicitly condoning the criminal behavior, it does not consider the broader
70 consequences of the response. Crucially, if we only check for correctness, as is standard practice [18],
71 we would find no fault in the response.

72 The problem is that to demonstrate good reasoning abilities, a correct answer is insufficient. We
73 need to show that the model answers the question correctly *for the right reasons*. In other words,
74 our evaluation must verify that the model’s processing of the input information *leads to* the correct
75 answer consistently and reliably. This criterion makes a *causal* claim about the model’s reasoning
76 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, to meaningfully evaluate reasoning abilities, one must assess what influences *how* the model arrives at its predictions, which is inherently a causal inference problem.

77 2.1 “Here be dragons”¹

78 To get a qualitative sense of the myriad of issues, or *monsters*, that plague our benchmarks and
79 experiments, we will briefly survey recent approaches, including broad overviews into the nature of
80 reasoning tasks [18, 19] and the evaluation of LLMs [1, 21, 22]. For investigations of more specific
81 issues, we separate efforts into three clusters depending on whether the problem originates with the
82 (1) models, (2) datasets, or (3) evaluation procedures.

83 **Models** This line of work focuses on characterizing the reasoning failures and biases of language
84 models, which is nontrivial given their opaque behavior [23]. These failures range from well-defined
85 formal errors such as logical fallacies [24], red herrings [25], or invalid inferences [26] to broader
86 issues including sensitivity to superficial features [3, 22], overconfidence [11], hallucinations [27, 28],
87 and lack of robustness [29–31]. Some studies explore how models exhibit “content effects” [32],
88 absorbing and amplifying human biases [33, 34] including social and cultural biases [13, 35–40],
89 such as stereotyping [41].

90 **Datasets** Meanwhile, subtle variations of popular benchmarks, such as premise order in reasoning
91 tasks[42] or minor changes in problem parameters [43, 44], can cause large performance drops [11,
92 12], raising concerns not just about whether models genuinely reason [45], but also about exploitable
93 issues in the training data and benchmarks [9, 46]. These can be described as enabling cheating [47].

¹The vague and uneasy language researchers often use when alluding to biases or unresolved limitations in their evaluations is reminiscent of how medieval cartographers would fill the unknown edges of their maps with dragons.

94 heuristics [48, 49], or shortcuts [50–52], possibly due to sampling biases [53] or in certain cases even
 95 leakage between the training and testsets [47] which can result in memorization [54]. Poor dataset
 96 construction can lead to annotation artifacts [55, 56] such as priming effects [57], which degrade
 97 the quality and reliability of results [58] while also unintentionally reinforcing social biases [59] or
 98 cultural inequities [13, 60, 61].

99 **Evaluation** Even with well-constructed datasets, evaluation methodologies can introduce systematic
 100 errors [62] or lead to misleading conclusions [7]. For example, automated scoring systems can obscure
 101 obvious failures [6], while static benchmarks can emphasize surface-level accuracy at the cost of
 102 other important factors, such as generalization [17] or interpretability [63] or social costs [8, 13].
 103 While standardized leaderboards [64] and evaluation procedures [65] can enable more direct model
 104 comparisons, these benchmarks can gradually become less representative of real-world tasks [10, 66–
 105 68], introduce biases that favor certain model families [69], or inadvertently leak information from the
 106 test set [47] which can be difficult to detect due to closed-source models and proprietary datasets [1].
 107 Despite the diverse, at times redundant, terminology, we observe certain structural similarities in the
 108 approaches of these contributions. Terms like “ablation”, “perturbations”, “edits”, “flips”, “masking”
 109 can often be interpreted as interventional or counterfactual analyses, while “sensitivity”/“robustness”,
 110 “consistency”, “shortcut”, “leakage”, “bias”, etc. refer to how the model’s behavior is impacted by,
 111 for example, (seen or unseen) confounders.

Claim 2: The monsters are causal

Many of the recurring issues in benchmark evaluation, including biases, spurious correlations, or systematic failure modes, are often described in vague or ad hoc terms. However, these issues arise from specific causal features of the underlying data-generating process or evaluation procedure. Whether the factors are known or latent, their influences can be captured by an appropriate causal model to formulate precise, testable hypotheses and guide more principled experimental design.

112 3 Common Abstract Topologies

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 Link		• 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) used to formalize a wide variety of *monsters* that may lurk in benchmark or experiment analysis. The graphs use  for the independent variable,  for the dependent variable, and  for a third-variable factor.

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

- 117 • precisely communicate the assumptions that go into a benchmark, experiment, or analysis
- 118 • leverage the machinery of causal inference for more principled analyses
- 119 • understand the implications of our design choices including the strengths and limitations on both
- 120 technical and conceptual levels

121 To help make the process of constructing a causal graph more accessible and systematic, we identify
 122 a set of Common Abstract Topologies (CATs) that frequently appear in causal graphs. For motivation,
 123 we list some associated phenomena (see Table 1) in the context of evaluating reasoning abilities in
 124 large models, where these patterns may offer useful abstractions.

125 However, researchers may be hesitant to commit to a specific causal graph that fully captures all
 126 factors influencing their analysis [70]. In practice, causal graphs are often underdetermined by the
 127 available data and may hinge on subtle choices in how variables are defined or interpreted [60].
 128 As Loftus [63] point out, some researchers even avoid causal framing altogether, since it makes
 129 explicit assumptions that reviewers may challenge.

Claim 3: Instrumentalism is all you need for model design

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 incrementally refined, while providing precise falsifiable hypotheses at every step of the way.

130 We join Loftus [63], Janzing and Garrido [71] in advocating for a more pragmatic, instrumentalist
 131 attitude to causal modeling. In many cases, the same phenomenon can be represented by multiple
 132 causal graphs that differ in variable selection, construction, or level of abstraction [72–74]. Neverthe-
 133 less, as long as a proposed causal model does not directly conflict with the available data, it may be
 134 sufficient to produce actionable insights (such as more interpretable or explainable models).

135 Aside from the additional explanatory power, if a more formal treatment is necessary or desired,
 136 there is a vast world of tools and techniques to explore. The field of causal inference [70, 75–78]
 137 has developed a language for formalizing the effects of subtle design choices and their, potentially
 138 counterintuitive, consequences for the analysis. For example, Simpson’s paradox can be elegantly
 139 explained, to “resolve” the apparent paradox based on the appropriate causal assumptions of the
 140 problem (for a deep dive into this topic see Pearl [79] and Chapter 6 of Pearl [76]).

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.

141 Causal inference is valuable not only for formal analysis but also as a conceptual framework for
 142 understanding the structural assumption behind an approach or argument. By making the concepts
 143 and tools of causal inference more accessible, we aim to develop a practical guide to recognize
 144 familiar causal structures in the real world, as well as build an intuition for the implications of
 145 model design choices on analysis and interpretation. To this end, we present three simple CATs
 146 that correspond to the three causal interpretations of a statistical dependence between two variables
 147 according to Reichenbach’s common cause principle [80].

148 Here is a brief sketch of how CATs can be used to guide model or benchmark design:

- 149 1. **Identify a relationship of interest:** Select a measurable outcome variable (e.g., model accuracy)
 150 and a primary explanatory “stimulus” (e.g., input prompt, fine-tuning data, or model parameter).
- 151 2. **Enumerate additional influences:** Consider other factors likely to affect the outcome either
 152 directly or indirectly. Assess whether they are conceptually *upstream* (potential confounders),
 153 *downstream* (potential mediators), or *parallel* sources of variation (potential sources of spurious
 154 correlations) with respect to the stimulus. Based on these relationships, consider the respective
 155 CAT (confounding, mediation, or spurious links) to serve as a structural template.
- 156 3. **Refine the graph:** Adapt the structure based on available data and the specific research question.
 157 Variables and edges may be omitted or aggregated, provided the resulting model supports plausible
 158 hypotheses and does not contradict observed dependencies.
- 159 4. **Use the model to guide analysis:** Apply the graph to derive testable implications (e.g. identify
 160 estimable causal queries by causal inference), suggest experimental interventions, or motivate
 161 mitigation strategies (e.g., balancing, ablation, or regularization).

3.1 Confounding



Confounding occurs when there is a common cause between the stimulus and response 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. 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 2.

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) [75]. 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.

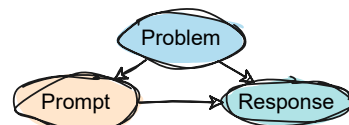


Figure 2: Sketch of a conceptual causal model treating the prompt (i.e. surface form) as a mediator between the underlying problem or task of interest and the model’s response.

3.3 Spurious Links



The third Common Abstract Topology (CAT) we discuss here describes a variety of graphical structures that give rise to spurious (i.e. non-causal) associations. More specifically, spurious links are statistical dependencies between variables that are not causally related (neither is an ancestor of the other), but are correlated due to an unblocked non-causal path in the underlying graph. This is often due to (1) a common cause (a confounder), which is usually unknown or too complex to be modeled explicitly, or due to (2) a common effect (a collider) which is conditioned on, thereby activating a backdoor path [75]. We denote non-causal links in a causal graph with a dashed curved edge, indicating that the observed association lacks a direct causal explanation.

Crucially, although spurious links are non-causal, models can still learn to exploit them for prediction, thereby effectively introducing an undesirable dependency (denoted by a dashed grey arrow) between the spurious feature and the outcome. This CAT is particularly relevant in machine learning, where models are typically trained on observational data only. In the absence of interventional or counterfactual signals, there is no principled way to distinguish causal from non-causal associations, making it easy for models to rely on spurious correlations in the data without appropriate inductive biases.

Another common source of spurious links, particularly in datasets, is due to selection bias, where the dataset includes only a subset of instances based on some unmodeled criteria. This selection process acts as a collider and can introduce spurious associations between otherwise independent variables.

Generally, it is not feasible to entirely eliminate spurious links, as seemingly innocent choices in variable construction and selection are invariably informed by the experimenter’s biases [60, 81]. Nevertheless, there is extensive causal inference machinery to address spurious correlations depending on the specific setting [82].

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



Several recent projects have used causal framing to identify or address confounding in LLM behavior. For example, Xia et al. [83] propose using a reward model as an instrumental variable to control for confounding in prompt biases. In a similar setting, Hüyük et al. [84] generate counterfactual samples to improve causal consistency. In human-LM collaboration, Zhang et al. [85] treat prior human and model actions as confounders influencing future decisions and introduce the Incremental Stylistic Effect to measure interventions in multi-turn interactions.

Meanwhile, an active area of research which could effectively be understood using the confounding CAT, investigates how dataset artifacts affect the performance of LLMs on mathematical reasoning tasks, such as GSM8K [20]. These studies vary premise order, subject matter, or input distributions [42, 43, 45, 69], often uncovering unexpected sensitivities.

Razeghi et al. [53] focus on the impact of token frequency in pretraining data on LLM performance in arithmetic tasks. A causal framing using the confounding CAT, shown in Figure 3a, captures their hypothesis that token frequency influences accuracy via a backdoor path.

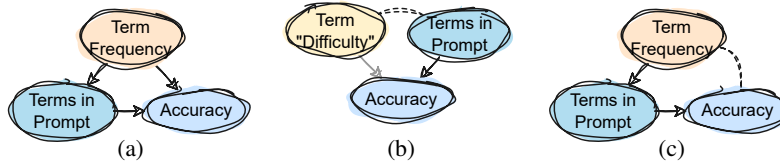


Figure 3: Causal framings of Razeghi et al. [53]: (a) confounding interpretation; (b) imagined alternative setting using a spurious link; (c) a graph combining CATs to avoid unsupported assumptions.

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. [53], 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 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. [53], we can expect to find a substantial correlation between the presence of certain numbers in the question and the model’s accuracy. To explain the results of our approach, we might phenomenologically posit a new property of numbers called “difficulty” which, the experiments demonstrate significantly affects the model’s accuracy, leading to the causal graph in Figure 3b using the spurious link CAT.

Verifying Causal Assumptions Notably, the graph in Figure 3a implies a falsifiable causal relationship between the term frequency and the accuracy which is, strictly speaking, not verified by the experiments of Razeghi et al. [53] which only show a correlation. Therefore, the hybrid causal graph in Figure 3c could be proposed where the correlation is instead due to a hypothesized 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 [86–90], but it is also useful in the LLMs comprehension of the underlying problem [91], augmentation of language models [92], embedding LLMs within larger programs [93], and the quantification of biases like, gender bias [37].

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. [94] 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: arguments 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.

255 This framing also highlights potential limitations, particularly regarding unmeasured confounders
 256 that could affect causal interpretations, as the authors’ “pattern-preserving” ablation does not fully
 257 isolate the induction heads’ effect. By considering mediation explicitly, we can better understand
 258 the underlying assumptions in their analysis and identify areas for further investigation, such as
 259 quantifying the natural indirect effect to understand the full impact of the induction heads on in-context
 260 learning abilities. In contrast, Stolfo et al. [87] propose a method for mechanistic interpretability of
 261 arithmetic reasoning in LLMs by editing the model’s parameters to characterize the information flow
 262 in the network.

263 4.3 Spurious Links

264 There are several recent projects that use causal models to characterize spurious links in, for example,
 265 factual knowledge [95], multi-modal models for fake news detection [96]. To avoid spurious features
 266 other projects design strategies to mitigate social biases [59], for finding useful demonstrations in
 267 few-shot learning [97] or to control NLP classifiers [98].

268 Chen et al. [96] develop a causal model to systematically quantify and remove two specific kinds of
 269 bias: psycholinguistic (use of emotional language) and image-only (ignoring text features). Note that
 270 the assumptions of the causal model address very specific types of bias using both interventional and
 271 counterfactual techniques.

272 Bansal and Sharma [98] presents a particularly interesting case as it addresses the same issue
 273 as Gardner et al. [57], but from a causal perspective. They both study the issue of label bias,
 274 specifically in “competency problems” [57], where an individual token in the prompt is not indicative
 275 of the label, but the model learns to rely on it, usually due to selection bias in the data collection.

276 The authors of Gardner et al. [57] propose a mitigation strategy based on “local edits” to individual
 277 tokens in the prompt to debias the benchmark. Using their statistical framing, the authors prove that
 278 the most promising strategy must apply local edits such that the label is flipped precisely half of the
 279 time, however, this result not only relies on some unrealistic assumptions, but also provides little
 280 insight into why this strategy is effective.

281 Translating this into a causal framing, we can gain a more general, and more intuitive result. Adopting
 282 a similar notation as Gardner et al. [57], the hypothesis can be formalized using the spurious links
 283 CAT where a single (text) feature, X_i , impacts the model’s prediction Y in the presence of the
 284 remaining input features $X_{\sim i}$ (stimulus) due to a label bias. Consequently, to remove the effect of X_i
 285 on the prediction, we need to design a strategy where the average causal effect (ACE) of X_i on Y
 286 conditioned on $X_{\sim i}$ is zero, which is, by definition:

$$\mathbb{E}(Y|X_{\sim i}, do(X_i = x'_i)) - \mathbb{E}(Y|X_{\sim i}, do(X_i = x_i)) = 0 \quad (1)$$

287 where $do(X_i = x'_i)$ denotes an intervention (i.e. edit) on feature X_i replacing x_i with some x'_i .

288 Note that Equation 1 is agnostic to the relationship between X_i and $X_{\sim i}$. If we make a “strong
 289 independence assumption” as in Gardner et al. [57], then we can directly recover their result that
 290 an edit on X_i should flip the label as often as not (i.e. making both terms equal). However, as
 291 the authors discuss, this assumption is not realistic as changing a single token may well affect the
 292 semantic meaning of the prompt beyond just the label (e.g. replacing “very” with “not” in a movie
 293 review). Here the causal framing provides us with a systematic way forward where, we can propose
 294 more realistic assumptions, and then use the rules of *do*-calculus to derive a corresponding mitigation
 295 strategy, such as the regularization term developed in Bansal and Sharma [98].

296 In summary, both approaches started with the same objective, but due to the purely statistical
 297 treatment, a cumbersome derivation still required an unrealistic assumption severely limiting the
 298 applicability of the method. The causal model not only provided a more intuitive motivation for the
 299 approach, but also offered a more powerful, principled method opening the door to further analysis,
 300 such as quantifying the impact of the label bias (e.g. by estimating the ACE), or controlling the effect,
 301 as described in Bansal and Sharma [98].

302 5 Alternative Views

303 We are hardly the first to point out systematic shortcomings of evaluation methodology, particularly
 304 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 [99], or more precisely defining LLMs [100] and how tasks relate to specific cognitive capabilities [66]. Raji et al. [8] 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 and Rumshisky [46] connect this to a proliferation of low-quality datasets.

Precisely this issue, that “benchmarking for NLU is broken” [9], can be addressed using causality. A causal framing can both provide a versatile way to define the underlying assumptions and design choices of a benchmark, while also offering principled methods for evaluating the benchmark’s external validity [101, 102].

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 [103]. This direction also coincides with an increasing practice in Natural Language Processing (NLP) to treat language models as agents [104, 105] or subjects in the social sciences [106–108]. Specifically, item response theory [109, 110] 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 [111]. Within the field of NLP there are also notable calls for more holistic evaluation schemes [9, 17, 69] and practical tools for improving the evaluation of language models [10, 112, 113] or even reintroducing principles from linguistic theory [114]. Lastly, there is growing interest in how LLMs acquire and apply causal knowledge [115, 116], including for causal reasoning [117], discovery [118], and even hypothesis generation [119] - this trend aligns with our core message: just as LLMs benefit from explicit causal modeling, so too 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, we illustrate how a causal framing can help understand the approach both conceptually in terms of how subtle design choices impact the causal interpretation, as well as technically where the grounded causal machinery can be used to derive a more general, actionable conclusions. Together, these examples demonstrate that even simple causal structures can offer insight into complex evaluation problems, guide experimental design, and refine our understanding of model behavior. We envision CATs as a practical tool, helping researchers quickly identify relevant causal models and choose appropriate inference strategies. Ultimately, causal models encourage more hypothesis-driven research which directly tackle key questions in a principled, transparent way, thereby leading to more robust progress across empirical machine learning.

References

- [1] Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. GPTEval: A Survey on Assessments of ChatGPT and GPT-4, December 2024.
- [2] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4, March 2023.
- [3] Tomer Ullman. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks, March 2023.
- [4] Katja Grace, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. Thousands of AI Authors on the Future of AI, January 2024.
- [5] Eric Martínez. Re-evaluating GPT-4’s bar exam performance. *Artificial Intelligence and Law*, March 2024. ISSN 1572-8382. doi: 10.1007/s10506-024-09396-9.
- [6] Raunak Chowdhuri, Neil Deshmukh, and David Koplow. No, GPT4 can’t ace MIT. <https://flower-nutria-41d.notion.site/No-GPT4-can-t-ace-MIT-b27e6796ab5a48368127a98216c76864>, June 2023.
- [7] Samuel R. Bowman. The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail, March 2022.
- [8] Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the Everything in the Whole Wide World Benchmark, November 2021.
- [9] Samuel R. Bowman and George E. Dahl. What Will it Take to Fix Benchmarking in Natural Language Understanding?, October 2021.
- [10] Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards, February 2024.
- [11] Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models, June 2024.
- [12] Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. Worse than Random? An Embarrassingly Simple Probing Evaluation of Large Multimodal Models in Medical VQA, October 2024.
- [13] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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, pages 610–623, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445922.
- [14] Ben Green and Lily Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the Machine Learning: The Debates Workshop*, 2018.
- [15] Lily Hu and Issa Kohler-Hausmann. What’s Sex Got To Do With Fair Machine Learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 513–513, January 2020. doi: 10.1145/3351095.3375674.
- [16] Luciano Floridi. Why the AI Hype is Another Tech Bubble. *Philosophy & Technology*, 37(4): 128, November 2024. ISSN 2210-5441. doi: 10.1007/s13347-024-00817-w.
- [17] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William

- 412 Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of
413 Language Models, October 2023.
- 414 [18] Jie Huang and Kevin Chen-Chuan Chang. Towards Reasoning in Large Language Models: A
415 Survey, May 2023.
- 416 [19] Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. Natural Language Reasoning, A
417 Survey, May 2023.
- 418 [20] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
419 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John
420 Schulman. Training Verifiers to Solve Math Word Problems, November 2021.
- 421 [21] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen,
422 Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu,
423 Qiang Yang, and Xing Xie. A Survey on Evaluation of Large Language Models, December
424 2023.
- 425 [22] Arash Hajikhani and Carolyn Cole. A Critical Review of Large Language Models: Sensitivity,
426 Bias, and the Path Toward Specialized AI, July 2023.
- 427 [23] Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings*
428 *of the National Academy of Sciences*, 120(6):e2218523120, February 2023. ISSN 0027-8424,
429 1091-6490. doi: 10.1073/pnas.2218523120.
- 430 [24] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu,
431 Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. Logical Fallacy Detection,
432 December 2022.
- 433 [25] Saeid Naeini, Raeid Saqr, Mozghan Saeidi, John Giorgi, and Babak Taati. Large Language
434 Models are Fixated by Red Herrings: Exploring Creative Problem Solving and Einstellung
435 Effect using the Only Connect Wall Dataset, November 2023.
- 436 [26] Abulhair Saparov and He He. Language Models Are Greedy Reasoners: A Systematic Formal
437 Analysis of Chain-of-Thought, March 2023.
- 438 [27] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. On the Origin of
439 Hallucinations in Conversational Models: Is it the Datasets or the Models? In Marine
440 Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of*
441 *the 2022 Conference of the North American Chapter of the Association for Computational*
442 *Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States, July
443 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.387.
- 444 [28] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu
445 Yao. Holistic Analysis of Hallucination in GPT-4V(ision): Bias and Interference Challenges,
446 November 2023.
- 447 [29] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large Language
448 Models Are Not Robust Multiple Choice Selectors, February 2024.
- 449 [30] Haoyu Wang, Guozheng Ma, Cong Yu, Ning Gui, Linrui Zhang, Zhiqi Huang, Suwei Ma,
450 Yongzhe Chang, Sen Zhang, Li Shen, Xueqian Wang, Peilin Zhao, and Dacheng Tao. Are
451 Large Language Models Really Robust to Word-Level Perturbations?, September 2023.
- 452 [31] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT Really Robust? A Strong
453 Baseline for Natural Language Attack on Text Classification and Entailment, April 2020.
- 454 [32] Gabriel Poesia, Kanishk Gandhi, Eric Zelikman, and Noah D. Goodman. Certified Reasoning
455 with Language Models, June 2023.
- 456 [33] Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Antonia Creswell, Dharshan
457 Kumaran, James L. McClelland, and Felix Hill. Language models show human-like content
458 effects on reasoning, July 2022.
- 459 [34] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal Parrots:
460 Large Language Models May Talk Causality But Are Not Causal, August 2023.
- 461 [35] Wolfgang Messner, Tatum Greene, and Josephine Matalone. From Bytes to Biases: Investigat-
462 ing the Cultural Self-Perception of Large Language Models, December 2023.

- [36] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social Biases in NLP Models as Barriers for Persons with Disabilities, May 2020.
- [37] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias, November 2020.
- [38] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study, March 2023.
- [39] Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. Investigating Cultural Alignment of Large Language Models, July 2024.
- [40] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 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.
- [41] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, CI ’23, pages 12–24, New York, NY, USA, November 2023. Association for Computing Machinery. ISBN 979-8-4007-0113-9. doi: 10.1145/3582269.3615599.
- [42] Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. Premise Order Matters in Reasoning with Large Language Models, May 2024.
- [43] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, October 2024.
- [44] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks, March 2024.
- [45] Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F. Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. Is Your Model Really A Good Math Reasoner? Evaluating Mathematical Reasoning with Checklist, October 2024.
- [46] Anna Rogers and Anna Rumshisky. A guide to the dataset explosion in QA, NLI, and commonsense reasoning. In Lucia Specia and Daniel Beck, editors, *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 27–32, Barcelona, Spain (Online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.coling-tutorials.5.
- [47] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don’t Make Your LLM an Evaluation Benchmark Cheater, November 2023.
- [48] Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic Without Algorithms: Language Models Solve Math With a Bag of Heuristics, May 2025.
- [49] Tom McCoy, Ellie Pavlick, and Tal Linzen. 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*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334.
- [50] Ruben Branco, António Branco, João António Rodrigues, and João Ricardo Silva. Short-cuttet Commonsense: Data Spuriousness in Deep Learning of Commonsense Reasoning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.113.
- [51] Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. How Pre-trained Language Models Capture Factual Knowledge? A Causal-Inspired Analysis, March 2022.

- [52] Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not All Neuro-Symbolic Concepts Are Created Equal: Analysis and Mitigation of Reasoning Shortcuts, December 2023.
- [53] Yasaman Razeghi, Robert L. Logan IV, Matt Gardner, and Sameer Singh. Impact of Pretraining Term Frequencies on Few-Shot Reasoning, May 2022.
- [54] Vitaly Feldman. Does Learning Require Memorization? A Short Tale about a Long Tail, January 2021.
- [55] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 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)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017.
- [56] Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.126.
- [57] Matt Gardner, William Merrill, Jesse Dodge, Matthew E. Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. Competency Problems: On Finding and Removing Artifacts in Language Data, December 2021.
- [58] Matthew Byrd and Shashank Srivastava. 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)*, pages 119–130, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.15.
- [59] Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. Prompting Fairness: Integrating Causality to Debias Large Language Models, March 2025.
- [60] Lily Hu and Issa Kohler-Hausmann. What’s Sex Got To Do With Fair Machine Learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 513–513, January 2020. doi: 10.1145/3351095.3375674.
- [61] Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models, March 2024.
- [62] Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mender-Dünner. Questioning the Survey Responses of Large Language Models, February 2024.
- [63] Joshua R. Loftus. Position: The Causal Revolution Needs Scientific Pragmatism. In *Proceedings of the 41st International Conference on Machine Learning*, pages 32671–32679. PMLR, July 2024.
- [64] Edward Beeching, Clémentine Fourier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. Available at https://huggingface.co/spaces/open-llmleaderboard/open_llm_leaderboard, 2023.
- [65] Aarohi Srivastava and et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. URL <http://arxiv.org/abs/2206.04615>.
- [66] David Schlangen. Language Tasks and Language Games: On Methodology in Current Natural Language Processing Research, August 2019.
- [67] Ali Shirali, Rediet Abebe, and Moritz Hardt. A Theory of Dynamic Benchmarks. URL <http://arxiv.org/abs/2210.03165>.
- [68] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking Benchmarking in NLP, April 2021.

- [69] Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A Careful Examination of Large Language Model Performance on Grade School Arithmetic, May 2024.
- [70] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl’s Hierarchy and the Foundations of Causal Inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36, pages 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition, March 2022. ISBN 978-1-4503-9586-1.
- [71] Dominik Janzing and Sergio Hernan Garrido. Phenomenological Causality, November 2022.
- [72] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Multi-level cause-effect systems. In Arthur Gretton and Christian C. Robert, editors, *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*, pages 361–369. JMLR.org, 2016. URL <http://proceedings.mlr.press/v51/chalupka16.html>.
- [73] P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, page ID 11, August 2017. URL <http://auai.org/uai2017/proceedings/papers/11.pdf>. *equal contribution.
- [74] Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in artificial intelligence*, pages 606–615. PMLR, 2020.
- [75] Judea Pearl. Causal inference in statistics: An overview. 2009.
- [76] Judea Pearl. *Book of Why*. Basic Books, New York, reprint edition edition, August 2020. ISBN 978-1-5416-9896-3.
- [77] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, USA, 2015. ISBN 0521885884.
- [78] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017. ISBN 978-0-262-03731-0.
- [79] Judea Pearl. Comment: understanding simpson’s paradox. In *Probabilistic and causal inference: The works of judea Pearl*, pages 399–412. 2022.
- [80] Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1956.
- [81] Wolfgang Pietsch. Aspects of theory-ladenness in data-intensive science. *Philosophy of Science*, 82(5):905–916, 2015.
- [82] Drago Plecko and Elias Bareinboim. A Causal Framework for Decomposing Spurious Variations, June 2023.
- [83] Yu Xia, Tong Yu, Zhankui He, Handong Zhao, Julian McAuley, and Shuai Li. Aligning as Debiasing: Causality-Aware Alignment via Reinforcement Learning with Interventional Feedback. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4684–4695, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.262.
- [84] Alihan Hüyük, Xinnuo Xu, Jacqueline Maasch, Aditya V. Nori, and Javier González. Reasoning Elicitation in Language Models via Counterfactual Feedback, March 2025.
- [85] Bohan Zhang, Yixin Wang, and Paramveer S. Dhillon. Causal Inference for Human-Language Model Collaboration, March 2024.
- [86] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and Mitigating Hallucination in Large Vision-Language Models via Modular Attribution and Intervention. In *The Thirteenth International Conference on Learning Representations*, October 2024.
- [87] Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis, October 2023.
- [88] Anshita Gupta, Debanjan Mondal, Akshay Krishna Sheshadri, Wenlong Zhao, Xiang Lorraine Li, Sarah Wiegrefe, and Niket Tandon. Editing Common Sense in Transformers, October 2023.

- [89] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT, January 2023.
- [90] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 small, November 2022.
- [91] Yujin Han, Lei Xu, Sirui Chen, Difan Zou, and Chaochao Lu. Beyond Surface Structure: A Causal Assessment of LLMs’ Comprehension Ability, November 2024.
- [92] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. Augmented Language Models: A Survey, February 2023.
- [93] Imanol Schlag, Sainbayar Sukhbaatar, Asli Celikyilmaz, Wen-tau Yih, Jason Weston, Jürgen Schmidhuber, and Xian Li. Large Language Model Programs, May 2023.
- [94] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.
- [95] Boxi Cao, Qiaoyu Tang, Hongyu Lin, Xianpei Han, and Le Sun. Does the Correctness of Factual Knowledge Matter for Factual Knowledge-Enhanced Pre-trained Language Models? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2327–2340, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.143.
- [96] Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. Causal Intervention and Counterfactual Reasoning for Multi-modal Fake News Detection. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.37.
- [97] Ruiyi Zhang and Tong Yu. Understanding Demonstration-based Learning from a Causal Perspective. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1465–1475, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.125.
- [98] Parikshit Bansal and Amit Sharma. Controlling Learned Effects to Reduce Spurious Correlations in Text Classifiers, June 2023.
- [99] Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema, October 2021.
- [100] Anna Rogers and Alexandra Sasha Luccioni. Position: Key Claims in LLM Research Have a Long Tail of Footnotes, June 2024.
- [101] Elias Bareinboim and Judea Pearl. 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.
- [102] Judea Pearl and Elias Bareinboim. External Validity: From Do-Calculus to Transportability Across Populations. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36, pages 451–482. Association for Computing Machinery, New York, NY, USA, 1 edition, March 2022. ISBN 978-1-4503-9586-1.
- [103] Oliver Wilhelm. Measuring reasoning ability. In *Handbook of Understanding and Measuring Intelligence*, pages 373–392. January 2005. ISBN 978-0-7619-2887-4. doi: 10.4135/9781452233529.n21.

- [104] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior, April 2023.
- [105] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as Agents, August 2023.
- [106] John J. Horton. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, January 2023.
- [107] Yan Leng and Yuan Yuan. Do LLM Agents Exhibit Social Behavior?, December 2023.
- [108] Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. *Perspectives on Psychological Science*, page 17456916231214460, January 2024. ISSN 1745-6916. doi: 10.1177/17456916231214460.
- [109] Frederic M Lord and Melvin R Novick. *Statistical theories of mental test scores*. IAP, 2008.
- [110] Frank B Baker. *The basics of item response theory*. ERIC, 2001.
- [111] Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 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)*, pages 4486–4503, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346.
- [112] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP models with CheckList, May 2020.
- [113] Saurabh Srivastava, Annarose M. B, Anto P V, Shashank Menon, Ajay Sukumar, Adwait Samod T, Alan Philipose, Stevin Prince, and Sooraj Thomas. Functional Benchmarks for Robust Evaluation of Reasoning Performance, and the Reasoning Gap. URL <http://arxiv.org/abs/2402.19450>.
- [114] 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.
- [115] Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. Understanding Causality with Large Language Models: Feasibility and Opportunities, April 2023.
- [116] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality, April 2023.
- [117] 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, January 2024.
- [118] Francesco Montagna, Max Cairney-Leeming, Dhanya Sridhar, and Francesco Locatello. Demystifying amortized causal discovery with transformers. URL <http://arxiv.org/abs/2405.16924>.
- [119] Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. 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.