CAUSALSIM: COUNTERFACTUAL IMPLICATION INVERSION AS A LOGICAL CONSISTENCY STRESS TEST FOR LARGE LANGUAGE MODELS

Youla Yang
Indiana University Bloomington
yangyoul@iu.edu

ABSTRACT

Large language models (LLMs) achieve strong performance on reasoning benchmarks, yet their structural logical consistency remains insufficiently understood. In particular, it is unclear whether models preserve valid implication direction when logical structure is minimally inverted while surface semantics remain nearly identical.

We introduce **CausalSim**, a benchmark for evaluating *counterfactual directional consistency* as a stress test of logical reasoning in LLMs. The benchmark consists of paired implication hypotheses ($A \rightarrow B$ vs. $B \rightarrow A$) that isolate sensitivity to implication reversal as a minimal structural perturbation.

We propose two evaluation metrics: the **Causal Advantage Index (CAI)**, measuring performance asymmetry under inversion, and **Balanced-CAI**, capturing cross-prompt logical consistency beyond raw accuracy.

Across six instruction-tuned LLMs, we observe systematic implication-direction asymmetries, demonstrating that high forward-direction accuracy does not guarantee structural logical robustness. Our findings position implication inversion as a minimal yet diagnostic probe of logical reasoning reliability in modern LLMs.

1 INTRODUCTION

Large language models (LLMs) demonstrate impressive performance across logical inference, numerical reasoning, and commonsense tasks. However, high benchmark accuracy does not necessarily imply structural logical robustness. Models may generate internally coherent explanations while relying on shallow correlational cues rather than preserving implication structure under minimal logical perturbations.

A minimal yet diagnostic perturbation is *implication inversion*: reversing the direction of a hypothesized implication while preserving premises and surface semantics. If a model correctly internalizes logical structure, it should systematically differentiate between $A \rightarrow B$ and $B \rightarrow A$ when only one direction is logically valid. Failure to do so reveals directional inconsistency and structural weakness in logical reasoning.

Most existing reasoning benchmarks evaluate performance under a single formulation of a task. Such protocols do not explicitly test cross-prompt logical consistency and therefore fail to expose asymmetric reasoning behaviors that emerge only under structural reversal.

To address this gap, we introduce **CausalSim**, a paired benchmark that contrasts model behavior under forward and inverted implication structures. By holding lexical content and prompt format constant, the benchmark isolates sensitivity to implication direction as a controlled stress test of logical reasoning.

We further introduce two metrics. The **Causal Advantage Index (CAI)** quantifies directional performance asymmetry under inversion. The **Balanced-CAI** measures accuracy jointly with cross-prompt consistency, providing a direction-aware assessment of structural logical robustness.

Empirical evaluation across six instruction-tuned LLMs reveals consistent implication-direction asymmetries. Even models with high overall accuracy frequently produce confident yet directionally inconsistent judgments under inversion.

These findings suggest that counterfactual directional consistency is a necessary dimension of logical reasoning evaluation for contemporary LLM systems.

2 METHODOLOGY

2.1 PAIRED IMPLICATION CONSTRUCTION

CausalSim is constructed around paired scenarios that differ only in the direction of a hypothesized implication:

$$H_c : A \to B, \quad H_a : B \to A. \tag{1}$$

The paired design preserves lexical content and framing while reversing only implication direction. This construction serves as a controlled test of implication preservation under counterfactual reversal. Observed performance differences therefore reflect structural logical robustness rather than prompt artifacts.

Each hypothesis is posed as a binary-choice decision (`A`/`B`). Forward and inverted variants are evaluated independently under identical deterministic decoding settings (`temperature=0`, `top_p=1.0`) to eliminate sampling variance.

2.2 WHY IMPLICATION INVERSION IS A LOGICAL STRESS TEST

Implication inversion is a minimal yet semantically consequential logical perturbation. Because LLMs are trained primarily on observational corpora, implication direction is often implicit or underspecified. Models may internalize associative regularities without encoding stable implication constraints.

Under forward formulation, associative priors may align with commonsense expectations; under inversion, the same priors can produce confident yet logically invalid conclusions.

We identify three recurring failure modes:

(1) Correlation-as-causation errors: inferring implication from co-occurrence regardless of direction;

(2) Narrative prior dominance: producing fluent but structurally invalid explanations under inversion;

(3) Directional default heuristics: assuming a preferred implication direction that is not revised under minimal structural change.

CausalSim therefore evaluates *cross-prompt logical consistency*—whether a model preserves implication validity under controlled directional reversal.

3 DATASET AND BENCHMARK

Availability. To preserve double-blind review, identifying links are omitted. The full benchmark, prompts, and evaluation scripts will be released upon acceptance.

Dataset Composition. The benchmark contains 120 paired scenarios (240 prompts total). Each pair includes a *forward* implication prompt and an *inverted* counterpart with nearly identical surface semantics. Scenarios are designed to emphasize logical structure and directionality rather than domain-specific memorization.

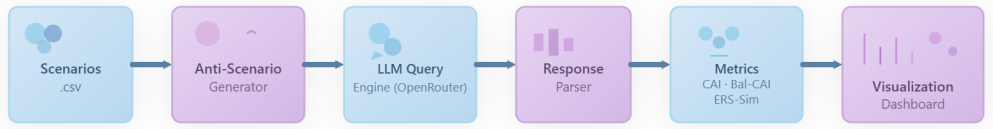

Figure 1: Evaluation pipeline for CausalSim. Paired forward and inverted implication prompts are issued under identical inference settings. Model outputs are parsed and aggregated to compute directional asymmetry (CAI) and cross-prompt consistency (Balanced-CAI).

Task Format. Each prompt asks the model to decide whether the stated implication is valid using a binary choice (`A`/`B`). The binary formulation reduces ambiguity and supports consistent aggregation across pairs.

Evaluation Protocol. Models are evaluated under identical deterministic decoding settings (`temperature=0`, `top_p=1.0`). Performance is reported using Accuracy (forward vs. inverted), the Causal Advantage Index (CAI; accuracy gap between directions), and Balanced-CAI, which incorporates cross-prompt consistency.

Limitations. CausalSim focuses on single-step implication judgments and does not evaluate multi-step deductive chains, formal proof construction, or long-horizon reasoning. The benchmark isolates structural consistency under direction reversal rather than full causal discovery.

4 Experimental Setup

4.1 Models

We evaluate six instruction-tuned large language models spanning both proprietary and open-weight families. The models cover a range of parameter scales and architectures to provide a representative assessment of contemporary LLM reasoning behavior.

4.2 Inference Protocol

All models are evaluated under deterministic decoding settings (`temperature=0`, `top_p=1.0`). Each prompt requests a binary decision (`A/B`). Forward and inverted prompts are evaluated independently to ensure controlled comparison.

4.3 Evaluation Metrics

We report:

Accuracy: correctness under forward vs. inverted implication direction;

Causal Advantage Index (CAI): directional performance gap between forward and inverted prompts;

Balanced-CAI: joint measure of accuracy and cross-prompt logical consistency;

SafeCons (optional diagnostic): fraction of pairs where the model is correct in *both* directions under the paired construction.

4.4 Reproducibility

All evaluation scripts are implemented in Python with fixed inference parameters. Raw outputs and parsed decisions are logged to enable full reproducibility.

5 Related Work

Logical reasoning and directionality in LLMs. A central question in LLM research is whether models perform genuine logical reasoning or rely on associative heuristics. Analyses of foundation models highlight that training on observational text can induce strong correlation-driven behavior that may fail under structurally meaningful perturbations (Kiciman et al., 2023). While much prior work studies causal reasoning and counterfactual evaluation, directionality reversal provides a particularly clean probe: it preserves surface semantics while flipping logical structure.

Benchmarks for causal/counterfactual reasoning. Recent benchmarks evaluate causal inference and counterfactual reasoning under interventions or variable manipulations (He et al., 2023; Jin et al., 2024; Liang et al., 2023; Zhou et al., 2024). However, many evaluations measure correctness under a single formulation and do not explicitly quantify *directional asymmetry* across paired forward/inverted prompts. CausalSim complements these efforts by using paired implication inversion to isolate directional robustness as an explicit evaluation target.

Prompting, self-consistency, and robustness. Chain-of-thought prompting and self-consistency sampling can improve multi-step reasoning accuracy by aggregating across generated rationales (Wang et al., 2022; Chen et al., 2023). Yet subsequent studies show that such techniques do not guarantee structural correctness and may remain brittle under rephrasing or distributional shift (Wu et al., 2023; Wei et al., 2023). Our paired construction probes robustness under a minimal structural perturbation and measures cross-prompt consistency directly.

Reliability and calibration. Beyond accuracy, research on reliability studies calibration, overconfidence, and robustness across tasks (Zhang et al., 2024; Mitchell et al., 2023; Lin et al., 2024). These works typically evaluate reliability over independently sampled prompts. In contrast, our setting links prompts via a controlled structural transformation, enabling a targeted measure of *logical consistency under inversion*.

Summary. Overall, CausalSim positions implication inversion as a minimal yet diagnostic stress test, and introduces direction-aware metrics to quantify whether LLMs preserve logical structure under controlled reversal.

6 Results

6.1 Directional Robustness under Implication Inversion

Figure 2 compares model accuracy under forward and inverted formulations of the same underlying scenarios. Across models, forward-direction accuracy is often substantially higher than accuracy under inversion, revealing systematic *directional asymmetry*. This gap indicates that strong performance under a single direction does not imply robust implication reasoning.

Several models achieve near-ceiling accuracy under forward formulations yet degrade sharply under inversion, producing large positive CAI values. Such behavior is consistent with direction-specific heuristics or correlation-driven associations that do not preserve implication validity when the direction is flipped.

Conversely, small or negative CAI values should be interpreted cautiously: a low directional gap can arise either from genuine structural robustness or from uniformly shallow behavior that performs similarly poorly in both directions. These results motivate evaluating directionality explicitly rather than relying on single-formulation accuracy.

Table 1 summarizes directional robustness and stability. Two patterns stand out. First, high forward accuracy does not guarantee stability under inversion: for example, LLaMA-3.1-8B attains near-ceiling Acc_f but extremely low SafeCons, indicating brittle, direction-dependent behavior. Second, negative CAI values (e.g., Qwen-2.5-7B and Gemma-2-9B) do not automatically imply stronger structural reasoning; such gaps may also reflect reversed heuristics or systematic misgeneralization. Overall, SafeCons varies substantially across models with comparable Acc_f, reinforcing that stability under direction reversal is a distinct dimension of logical reliability.

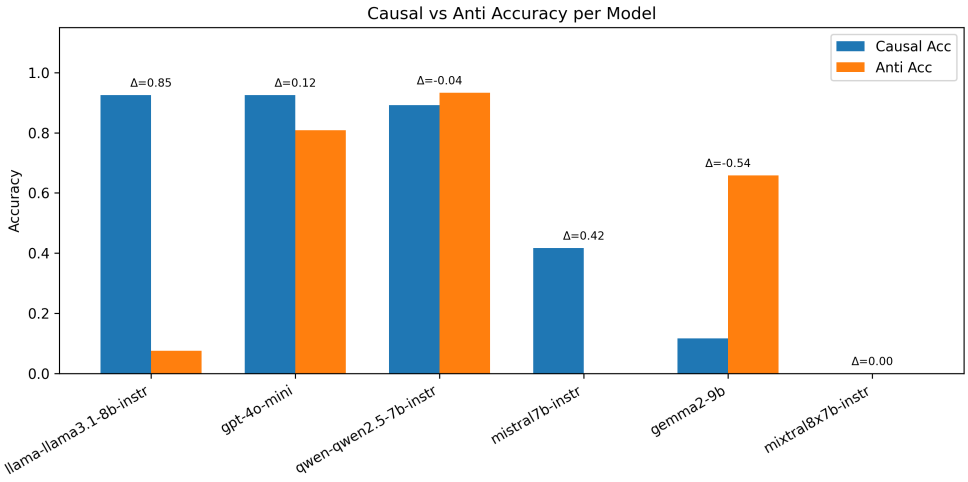

Figure 2: Accuracy under forward vs. inverted implication formulations across evaluated models. The difference between the two bars corresponds to the Causal Advantage Index (CAI), measuring sensitivity to direction reversal.

Table 1: Model-level directional robustness and stability. Acc_f and Acc_i denote accuracy under forward and inverted formulations. CAI is the directional gap ($\text{Acc}_f - \text{Acc}_i$). SafeCons measures the fraction of paired scenarios where the model is correct under *both* directions.

Model	Acc_f	Acc_i	CAI	SafeCons
LLaMA-3.1-8B-Instruct	0.93	0.08	**+0.85**	0.03
GPT-4o-mini	0.93	0.81	+0.12	0.78
Qwen-2.5-7B-Instruct	0.89	0.93	**–0.04**	0.89
Mistral-7B-Instruct	0.42	0.00	+0.42	0.00
Gemma-2-9B	0.12	0.66	**–0.54**	0.09
Mixtral-8×7B	0.00	0.00	0.00	0.00

6.2 Accuracy versus Cross-Prompt Consistency

Figure 3 plots forward accuracy against safe cross-prompt consistency (SafeCons), defined as the fraction of paired scenarios for which the model is correct under both forward and inverted formulations. The results reveal a dissociation between correctness and stability.

Models with similar forward accuracy can exhibit markedly different consistency profiles, indicating that single-direction accuracy is an unreliable proxy for logical robustness. Balanced-CAI captures this effect by jointly accounting for accuracy and consistency, penalizing models that appear strong under accuracy-centric evaluation but fail to maintain coherent behavior under inversion.

6.3 Model Agreement and Structural Heterogeneity

Figure 4 presents pairwise agreement across models. While partial clustering is observed among related model families, agreement remains limited overall. This heterogeneity suggests that current models do not share a stable inductive bias for implication directionality, even when trained with overlapping instruction-following data.

Moreover, agreement does not necessarily imply correctness: in multiple cases, models converge on directionally invalid judgments under inversion. This highlights that consensus can reflect shared training artifacts rather than robust logical inference.

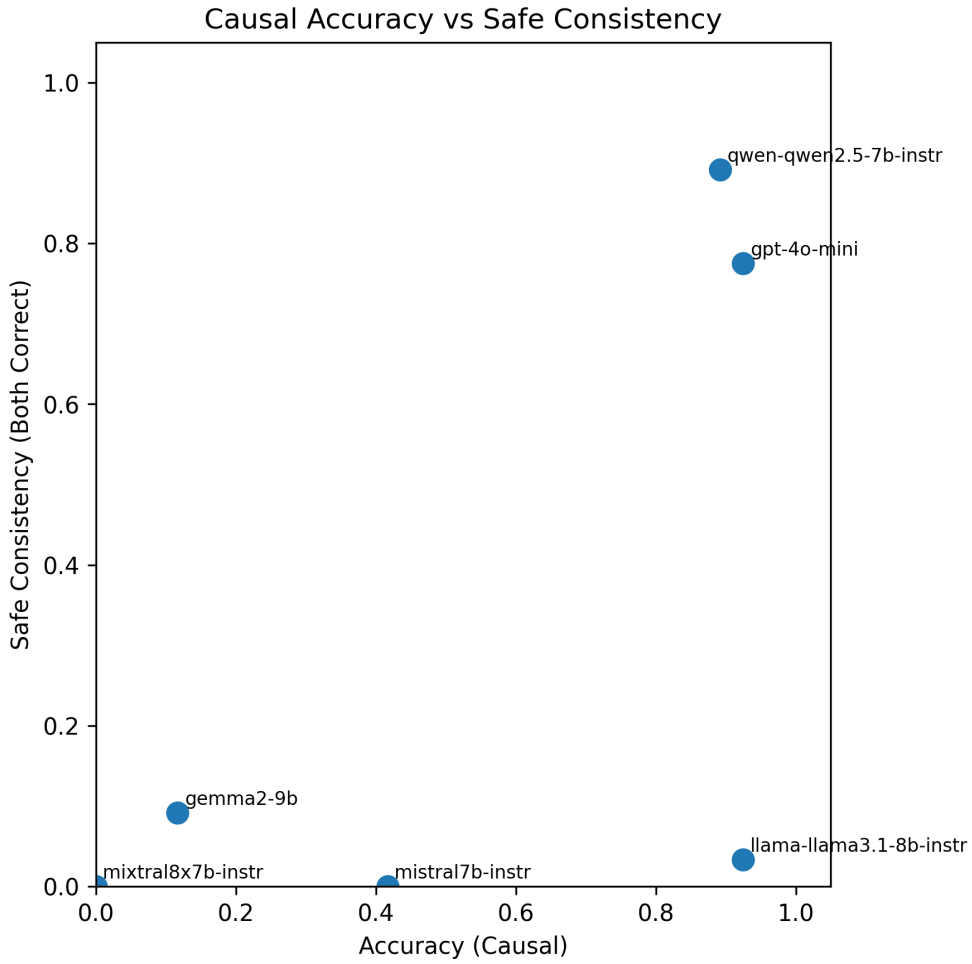

Figure 3: Relationship between forward-direction accuracy and safe cross-prompt consistency under inversion. Each point represents a model. High accuracy does not necessarily imply stable or consistent behavior under direction reversal.

6.4 Failure Pattern Analysis

We analyze prediction errors across paired forward/inverted prompts to identify recurring failure modes under direction reversal.

Correlation-driven implication. A prevalent failure mode arises when models treat correlation or co-occurrence as sufficient evidence for implication, producing the same decision in both directions even when only one direction is logically valid. This behavior yields high forward accuracy but collapses under inversion, contributing to large CAI values and low SafeCons.

Narrative prior dominance. We observe cases where fluent explanations override implication validity. Under inverted prompts, models may generate plausible narratives that align with common discourse patterns while violating the intended implication direction. This helps explain why some models show high forward accuracy yet low cross-prompt consistency.

Directional default heuristics. Several models appear to internalize default implication arrows for recurring templates. When presented with inverted counterparts, these defaults are not reliably revised, leading to systematic directional asymmetry. Conversely, small CAI values can also arise

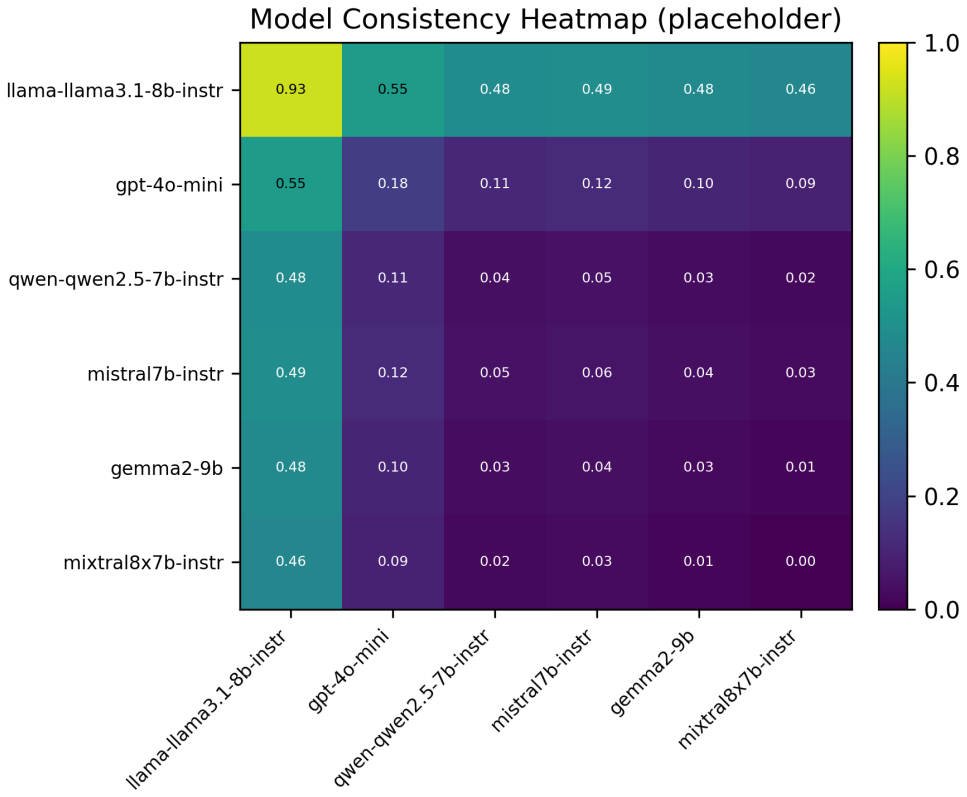

Figure 4: Pairwise agreement heatmap across model predictions. Brighter cells indicate higher agreement in implication judgments. Agreement varies substantially across model families.

from uniformly shallow behavior, reinforcing the need to interpret CAI jointly with consistency-based metrics (e.g., Balanced-CAI and SafeCons).

Summary. Together, these patterns explain why accuracy-centric evaluation can overestimate logical reliability. Implication inversion exposes brittle direction-dependent behaviors that remain hidden under standard single-formulation benchmarks, making cross-prompt consistency a useful diagnostic for logical reasoning robustness.

7 Discussion

Our results reveal a persistent gap between correlation-following behavior and robust causal reasoning in contemporary LLMs. While many models achieve high accuracy under forward causal phrasing, causal inversion systematically exposes brittle dependencies and directional instability. This pattern is consistent with models learning asymmetric heuristics that exploit surface-level regularities rather than internalizing stable causal abstractions.

Crucially, forward-direction accuracy does not guarantee robustness under minimal counterfactual perturbations. Models may produce confident and internally coherent explanations that invert causal direction, highlighting a disconnect between linguistic plausibility and causal validity. This motivates the use of **Balanced-CAI** as a robustness-oriented complement to accuracy-centric evaluation.

Risk-aware analysis further emphasizes the practical importance of directional robustness. Directional errors disproportionately affect high-impact scenarios, suggesting that accuracy alone is insufficient for trustworthy deployment in decision-support pipelines. Even under deterministic decoding, models can produce confidently incorrect causal judgments, and such errors become more consequential when weighted by domain-specific risk.

Finally, agreement analysis indicates partial convergence in causal judgments across models, but agreement does not imply correctness. The observed heterogeneity across architectures suggests the absence of shared causal priors and motivates future evaluation in ensemble and multi-agent settings.

8 Ethical Considerations

CausalSim-Enterprise is designed as an evaluation benchmark and does not involve the collection or use of personal, sensitive, or proprietary data. All scenarios are synthetic and abstracted to reflect general decision-making patterns rather than real-world individuals or organizations.

The benchmark is intended solely for diagnostic and research purposes. Risk annotations used in the Enterprise Risk Score (ERS) reflect heuristic estimates of potential downstream impact and are not intended to prescribe or automate real-world decisions. We caution against deploying model judgments evaluated under this benchmark directly in high-stakes settings without additional human oversight and domain validation.

By exposing failure modes under causal inversion, the benchmark aims to promote safer deployment practices and more transparent evaluation of LLM-based decision-support systems, rather than enabling autonomous decision-making without accountability.

9 Conclusion

We introduced CausalSim, a benchmark and evaluation framework designed to assess *directional logical robustness* in large language models. By constructing paired *forward* and *inverted* implication prompts with nearly identical surface semantics, the benchmark isolates sensitivity to direction reversal and enables controlled analysis of directional asymmetry.

We proposed complementary direction-aware metrics—the **Causal Advantage Index (CAI)** and **Balanced-CAI**—together with a consistency-based diagnostic (SafeCons) to characterize failure modes that are not observable under accuracy-centric evaluation alone. Experiments across six instruction-tuned models reveal systematic directional asymmetries, limited cross-prompt stability under inversion, and substantial variation in consistency even among models with similar forward accuracy.

Beyond individual model performance, agreement analysis highlights structural heterogeneity in implication judgments across architectures, suggesting the absence of a shared or stable inductive bias for implication directionality. Overall, these findings indicate that preserving logical structure under minimal directional perturbations remains an open and underexplored challenge, even for strong modern systems.

Broader impact and extensibility. CausalSim is designed to be lightweight, interpretable, and extensible. The paired direction-reversal paradigm can be readily applied to new domains, alternative logical relations, or longer-horizon settings. We hope this benchmark encourages the community to adopt directional robustness and cross-prompt consistency as first-class evaluation criteria, supporting the development of more reliable and logically grounded language models.

References

Zheng Chen, Xuezhi Wang, Yuhuai Zhou, et al. Universal self-consistency for large language model reasoning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Jiahui He et al. Causalbench: Benchmarking causal reasoning capabilities of language models. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Zhijing Jin et al. Evaluating causal reasoning in large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Emre Kiciman, Amit Sharma, et al. Causal reasoning and large language models. *Communications of the ACM*, 66(10):76–86, 2023.

Chen Liang et al. Evaluating counterfactual reasoning in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Sheng Lin et al. On the calibration of large language models. *arXiv preprint*, 2024.

Eric Mitchell et al. Trusteval: A benchmark for trustworthy language models. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Sharan Narang. Self-consistency improves chain-of-thought reasoning in language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Jason Wei et al. Chain-of-thought prompting fails under distribution shift. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Zhengxuan Wu et al. On the fragility of reasoning in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Yue Zhang et al. Trustllm: Trustworthiness evaluation of large language models. *arXiv preprint*, 2024.

Yuxuan Zhou et al. Counterfactual evaluation of language models under causal interventions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.