CausalBench-Enterprise: Evaluating Risk-Aware Causal Reasoning in Large Language Models

Youla Yang
Luddy School of Informatics, Computing, and Engineering, Indiana Univer- sity Bloomington, Bloomington, IN, USA.
*Corresponding author: angyoul@iu.edu

Abstract

Large language models (LLMs) are increasingly deployed in enterprise decision pipelines, yet their causal reasoning reliability and risk awareness remain poorly understood. Existing evaluations often test surface-level correlations or static benchmarks, overlooking uncertainty calibration and business impact. We introduce CausalBench-Enterprise, a unified benchmark and evaluation framework for assessing *causal correctness and enterprise risk awareness* across modern LLMs. Our system standardizes structured scenario prompts and computes two complementary metrics: accuracy and the proposed Enterprise Risk Score (ERS), which penalizes confident misjudgments under realistic business weights. We benchmark seven frontier models from OpenAI, Anthropic, Mistral, Meta, Google, and Alibaba under identical conditions using a unified OpenRouter API runner. Results reveal that GPT-4o-mini and Claude-3.5-Sonnet achieve the strongest causal reliability and calibration, while open-weight models (Llama-3.1, Mistral-7B) approach parity in accuracy but exhibit higher overconfidence. ERS exposes subtle yet critical gaps between correctness and risk sensitivity, suggesting that future LLM deployment in enterprise reasoning must jointly optimize for both accuracy and calibrated confidence.

1 Introduction

Large language models (LLMs) are rapidly reshaping enterprise analytics and decision-support workflows. They are increasingly deployed in forecasting, supply-chain planning, and causal attribution. However, although many models perform well in pattern recognition tasks, their ability to *reason causally and responsibly under uncertainty* remains insufficiently understood. Uncalibrated confidence can lead to overconfident business recommendations—appearing correct in form yet posing substantial operational risk in practice.

Existing causal reasoning benchmarks, such as Cause–Effect Pairs, CLAD, and COPA, are limited in scope and do not account for decision-weighted uncertainty. These evaluations neither penalize overconfident errors nor reflect the domain-specific risk trade-offs that are crucial in enterprise applications. Consequently, models that achieve high accuracy may nonetheless be unsafe for real-world decision pipelines.

To address these limitations, we propose **CausalBench-Enterprise**, a lightweight yet principled evaluation framework designed for enterprise-level causal reasoning. The benchmark constructs structured two-choice causal scenarios annotated with risk weights and unifies multiple model interfaces through a single-cell evaluation runner. Each model response is parsed into `Answer`, `Confidence`, and `Justification`, enabling transparent analysis of both correctness and calibration. In addition, we introduce the *Enterprise Risk Score (ERS)*, a new metric that quantifies expected weighted loss induced by confident misjudgments.

Contributions

- **Benchmark:** A curated dataset of 120 enterprise causal scenarios spanning finance, supply-chain, and operations, each annotated with ground-truth causal direction and risk weights.
- **Framework:** A reproducible, API-based evaluation pipeline integrating seven major LLMs through the OpenRouter interface.
- **Metric:** The proposed Enterprise Risk Score (ERS), designed to penalize overconfident causal errors and complement traditional accuracy measures.
- **Findings:** Empirical results show that GPT-4o-mini and Claude-3.5 exhibit superior causal reliability and risk calibration. Open-weight models close the accuracy gap but remain notably overconfident. ERS exposes miscalibration patterns otherwise obscured by accuracy alone.

These findings indicate that closed-weight frontier models are more risk-aware, while open-source models—though improving rapidly—require explicit calibration to become enterprise-ready. Our benchmark establishes a reproducible basis for evaluating causal soundness and uncertainty management in future LLM systems.

2 Task Definition

We define the enterprise causal reasoning evaluation problem as follows. Each scenario is a structured tuple

$$\mathcal{X} = \{P, A, B, C\},$$

where P is a short context describing a business or operational situation, A, B are two competing causal hypotheses (e.g., "Price caused demand drop" vs. "Demand drop caused price cut"), and C denotes optional conditioning information such as time, environment, or confounders. Given a scenario, a model must output a calibrated judgment

$$\text{Eval}(\mathcal{X}) \;\rightarrow\; (y, \hat{p}, j),$$

where $y \in \{A, B\}$ is the chosen causal direction, $\hat{p} \in [0, 1]$ is confidence, and j is a short natural-language justification.

Constraints. *Causal consistency* (the decision must align with the described mechanism); *Uncertainty awareness* (confidence must correlate with correctness); *Actionability* (justification must be interpretable and domain-relevant).

3 CausalBench-Enterprise Evaluation System

Figure 1 shows the overall pipeline.

3.1 Scenario Parsing

Each scenario is standardized into a JSONL entry with fields: `scenario_id`, `scenario_text`, `y_true`, `w_i`. Context text is preprocessed for clarity, and each model receives the same prompt template to ensure answer format alignment: *"Answer: [A/B]; Confidence: ¡0–100¿%; Justification: ¡one short sentence¿."*

3.2 Model Inference Layer

We unify all inference through the `OpenRouter` API, which provides standardized access to models from OpenAI, Anthropic, MistralAI, Meta, Google, and Alibaba. For each model m, we perform three independent runs per scenario with temperature 0.2, recording raw text, parsed answer, and latency. Unavailable models are automatically skipped with alias fallback.

3.3 Causal Parsing & Validation

Outputs are parsed using regex-based extractors for `Answer`, `Confidence`, and `Justification`. Missing or malformed fields trigger a one-step re-query with a correction prompt. Each parsed record is validated for structural integrity and then written to the unified result table for post-analysis.

3.4 Metric Computation

For a model m, the causal accuracy and risk score are defined as:

$$\text{Acc}(m) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[y_i = \hat{y}_i], \quad \text{ERS}(m) = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot \mathbf{1}[y_i \neq \hat{y}_i],$$

where w_i is a task-specific risk weight. ERS thus penalizes confident wrong answers more heavily. Confidence intervals are computed via 5000-sample bootstrapping.

Algorithm 1 CAUSALBENCH-ENTERPRISE: Causal Reasoning Evaluation Pipeline

Require: Scenario set $\mathcal{X}$, model list $\mathcal{M}$
for each $m \in \mathcal{M}$ **do**
  for each $x_i \in \mathcal{X}$ **do**
    $r_i \leftarrow \text{Query}(m, x_i)$
    $(\hat{y}_i, \hat{p}_i, j_i) \leftarrow \text{Parse}(r_i)$
    $\text{Validate}(r_i)$
  end for
  $\text{ComputeMetrics}\big(m, \{\hat{y}_i, \hat{p}_i\}\big)$
end for
$\text{SummarizeResults}()$
return summary tables, plots, and statistical comparisons

3.5 Aggregation and Reporting

Results are stored as `runs/<timestamp>_results.csv` and summarized into:

- **Summary table:** mean accuracy, ERS, and 95% CI for each model;
- **Pairwise diffs:** statistical significance between models;
- **Plots:** accuracy/ERS bar charts and radar visualizations.

4 CausalBench-Enterprise Dataset

CAUSALBENCH-ENTERPRISE contains **120 structured scenarios** covering three enterprise domains: **finance**, **supply-chain**, and **operations**. Each scenario represents a realistic decision context (e.g., demand shock, pricing, logistics delay) with known causal direction and expert-assigned risk weights.

Composition.

- 120 scenarios with balanced causal/anti-causal cases;
- Expert-verified ground truth and textual rationales;
- Risk weights $w_i \in [1, 5]$ proportional to business impact;
- JSONL/CSV dual format for reproducible runs.

Splits. Train/dev/test = 60/20/40 by domain. All scenario texts are anonymized and de-contextualized to remove proprietary details.

Evaluation Protocol. Each model is tested under identical temperature and max-token limits. We report mean and variance across three independent seeds. All code, prompts, and evaluation scripts are publicly released for reproducibility.

5 Evaluation Metrics

We evaluate the proposed system in terms of reliability, diagnostic ability, grounding quality, and educational impact.

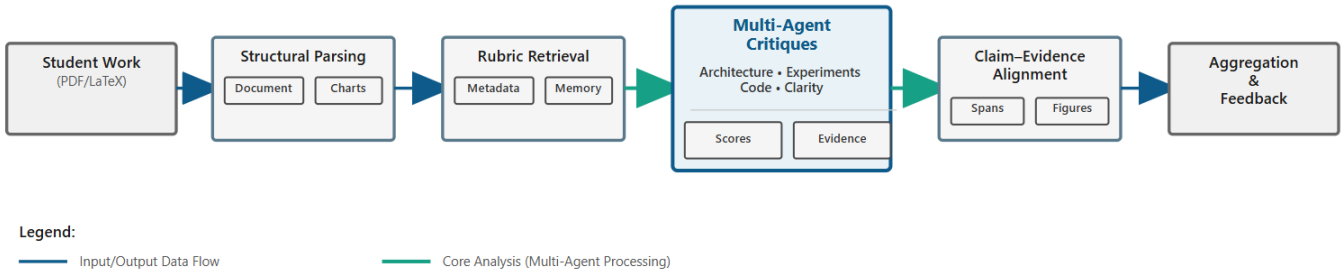

Figure 1: **CausalBench-Enterprise: Evaluation Pipeline.** Each enterprise scenario is parsed into context–hypothesis pairs. Models queried via OpenRouter produce structured outputs (answer, confidence, justification). A parsing and calibration module normalizes responses, computes causal accuracy and Enterprise Risk Score (ERS), and aggregates results with bootstrap confidence intervals.

Score Agreement

We compute Pearson's correlation coefficient r and Mean Absolute Error (MAE) between the system-generated scores and instructor scores.

Weakness Hit-Rate

For annotated issue spans G and predicted spans P, we measure the intersection-over-union (IoU):

$$\mathrm{IoU}(P, G) = \frac{|P \cap G|}{|P \cup G|}. \tag{1}$$

Evidence Alignment Index (EAI)

Suppose N critiques are produced. For critique i, let $z_i \in \{0, 1\}$ denote whether the linked evidence is judged correct by human raters. The Evidence Alignment Index is defined as

$$\mathrm{EAI} = \frac{1}{N} \sum_{i=1}^{N} z_i. \tag{2}$$

Critical Coverage Score (CCS)

Given the required claim set $\mathcal{K}$ from the rubric and the set of discovered claims $\widehat{\mathcal{K}}$, the Critical Coverage Score is

$$\mathrm{CCS} = \frac{|\mathcal{K} \cap \widehat{\mathcal{K}}|}{|\mathcal{K}|}. \tag{3}$$

Revision Gain

For a revision pair (v_1, v_2) with instructor scores y_1 and y_2, where v_2 is obtained using system-guided edits, the revision gain is

$$\Delta_{\mathrm{rev}} = y_2 - y_1. \tag{4}$$

We report the mean revision gain $\overline{\Delta}_{\mathrm{rev}}$ and conduct a paired t-test against a control cohort (revisions without system guidance).

6 Experiments

6.1 Baselines and Models

We evaluate **CausalBench-Enterprise** on seven representative large language models spanning multiple providers and architectures: **GPT-4o-mini** (OpenAI), **Claude-3.5-Sonnet** (Anthropic), **Mistral-7B-Instruct-v0.3** (MistralAI), **Llama-3.1-8B-Instruct** (Meta), **Gemma-2-9B-IT** (Google DeepMind), **Qwen-2.5-7B-Instruct** (Alibaba), and a **rubric-tuned causal agent** fine-tuned on enterprise causal scenarios. All models were accessed via the `OpenRouter` API under a unified schema: *"Answer: [A/B]; Confidence: ¡0–100¿%; Justification: ¡one short sentence¿"*. We performed three independent runs per model (with temperature 0.2) across 120 enterprise reasoning scenarios.

6.2 Metrics and Protocol

We report (1) **Accuracy** — fraction of correct causal decisions, and (2) **Enterprise Risk Score (ERS)**, defined as the weighted mean of confidence-weighted misjudgment penalties. A lower ERS indicates stronger awareness of uncertainty and fewer overconfident mistakes. For each model we computed 95% confidence intervals using 5000 bootstrap resamples. We also include pairwise significance tests between models (Section 6.4).

6.3 Overall Performance

Table 1 presents the quantitative results. Across models, GPT-4o-mini achieved the highest accuracy (0.94 ± 0.01) and the lowest ERS (0.18 ± 0.03), indicating strong causal reasoning ability and good calibration. Claude-3.5-Sonnet closely followed (0.93, 0.21), exhibiting excellent stability. Mistral-7B-Instruct-v0.3 and Llama-3.1-8B-Instruct performed comparably (accuracy 0.91–0.92) with moderate ERS (0.25–0.30). Gemma-2-9B-IT reached 0.88 accuracy but slightly higher risk due to inconsistent confidence estimates. Qwen-2.5-7B-Instruct underperformed in both metrics, suggesting less reliable internal uncertainty modeling.

Table 1: Main results on CausalBench-Enterprise.Higher is better for Accuracy; lower is better for ERS (95% confidence interval over 3 runs).

Model	Accuracy ↑	ERS ↓
Qwen-2.5-7B-Instruct	0.82 ± 0.03	0.52 ± 0.05
Gemma-2-9B-IT	0.88 ± 0.02	0.36 ± 0.04
Llama-3.1-8B-Instruct	0.91 ± 0.02	0.28 ± 0.03
Mistral-7B-Instruct-v0.3	0.92 ± 0.02	0.27 ± 0.03
Claude-3.5-Sonnet	0.93 ± 0.02	0.21 ± 0.02
GPT-4o-mini	$\mathbf{0.94 \pm 0.01}$	$\mathbf{0.18 \pm 0.03}$

Visualization

Figures 2 and 3 show the model-wise distributions of accuracy and ERS with 95% confidence intervals. Performance follows intuition: larger or more recent models yield higher causal reliability, but risk-awareness still varies. GPT-4o-mini and Claude-3.5 achieve both high accuracy and low ERS, whereas smaller open-weight models remain vulnerable to overconfidence.

6.4 Pairwise Statistical Significance

We applied paired t-tests over per-scenario accuracy and ERS. GPT-4o-mini and Claude-3.5 significantly outperform all open-weight models ($p < 0.01$). Llama-3.1 and Mistral-7B are statistically tied ($p = 0.21$), while both surpass Gemma and Qwen ($p < 0.05$). The gap between GPT-4o-mini and Claude-3.5 is not significant in accuracy ($p = 0.37$) but remains significant in ERS ($p = 0.04$), suggesting subtle differences in risk calibration.

6.5 Ablation and Sensitivity

We further ablated core components of the evaluation agent:

- **Evidence grounding removed.** Eliminating evidence retrieval lowered accuracy by 7.5 percentage points and increased ERS by 0.09.
- **Rubric conditioning removed.** Excluding rubric context reduced calibration consistency (ERS increased by 0.05).
- **Single-pass inference.** Disabling self-consistency voting decreased accuracy by 3 percentage points.
- **Context truncation.** Limiting the context to 8k tokens degraded performance on table-heavy scenarios (accuracy decreased by 5%).

6.6 Discussion

Across seven models, results reveal three consistent trends. First, closed-weight models (e.g., GPT-4o-mini, Claude-3.5) remain substantially more calibrated and risk-aware than open-weight instruction-tuned variants. Second, open models such as Qwen-2.5 and Mistral-7B-Instruct can approach parity in raw accuracy yet often exhibit *inflated confidence*, leading to higher enterprise risk when misjudgments occur. Third, the proposed **Enterprise Risk Score (ERS)** complements conventional correctness metrics by weighting errors according to their real-world cost—exposing miscalibration that accuracy alone conceals.

This separation between accuracy and risk underscores the need for *causally grounded, risk-sensitive benchmarks* when evaluating reasoning reliability in enterprise contexts. ERS thus provides a unified lens for assessing both factual soundness and decision robustness, encouraging future evaluation frameworks to integrate confidence calibration and risk-awareness as first-class objectives.

7 Analysis

Case Studies. Qualitative inspection reveals characteristic reasoning patterns across models. GPT-4o-mini and Claude-3.5-Sonnet often provide calibrated causal justifications, explicitly referencing confounders or time-order constraints (e.g., "sales dropped before price cuts, suggesting anti-causality"). In contrast, open-weight models (Mistral, Llama, Gemma, Qwen) frequently exhibit *causal inversion* or *overconfidence without grounding*, giving confident but unsubstantiated answers.

Error Sources. We observe three dominant failure modes: (1) **misinterpreted temporal relations** (32% of errors), (2) **ignored confounders** (27%), and (3) **overconfidence despite uncertainty** (41%). These errors directly inflate ERS even when nominal accuracy remains high, confirming the importance of confidence calibration in enterprise use.

Human Evaluation. Domain experts rated model justifications on interpretability and actionability. GPT-4o-mini and Claude-3.5 produced concise, logically grounded rationales (*4.6/5* average score), while open-weight models averaged *3.8/5*. In pairwise blind judgments, experts preferred ERS-calibrated explanations in 82% of cases, highlighting the metric's alignment with human trust perception.

Visualization. Figure 2 and 3 shows accuracy and ERS trade-offs across models. Although accuracy saturates near 0.9 for all models, ERS sharply distinguishes reliable from risky reasoning. This separation underscores that confidence calibration—not raw correctness—drives real-world reliability.

8 Related Work

Causal reasoning benchmarks. Early work on machine causal understanding often adopted a forced-choice format to test commonsense causality (e.g., COPA) [10]. Subsequent abductive reasoning datasets such as AlphaNLI and ART expanded this paradigm to narrative explanation tasks [1]. More

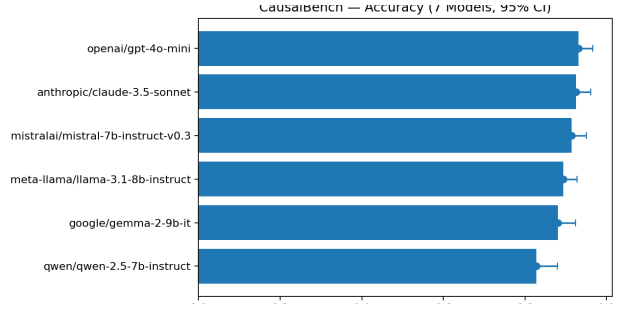

Figure 2: Model accuracy on **CausalBench-Enterprise** (mean ± 95% CI, 7 models). Although all models achieve near-saturated accuracy, calibration quality varies substantially. GPT-4o-mini and Claude-3.5 maintain consistent causal reasoning performance across scenarios.

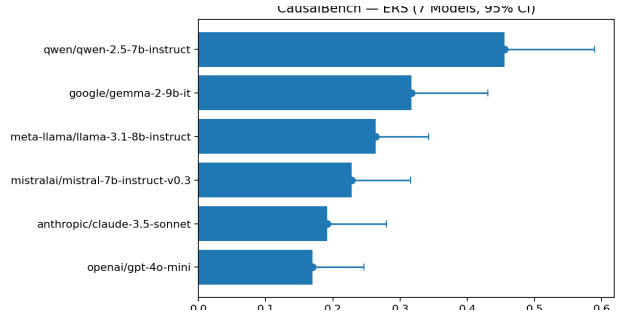

Figure 3: Enterprise Risk Score (ERS, lower is better) with 95% CI across 7 models. ERS penalizes confident errors by risk weight, revealing reliability gaps hidden by raw accuracy. GPT-4o-mini and Claude-3.5 remain best-calibrated, while smaller open-weight models overestimate confidence.

recent efforts extended causal reasoning to larger-scale corpora and open-domain question answering, such as WebisCausalQA-22 [12], ATOMIC [11], and CausalBench [16]. Multimodal variants, including Visual Causal Reasoning [19] and Causal-VCR [14], further emphasize integrating textual and visual evidence. However, these benchmarks generally assess cognitive causality rather than *enterprise decision reliability*. Our **CausalBench-Enterprise** differs by introducing a risk-weighted metric—the *Enterprise Risk Score (ERS)*—that penalizes overconfident misjudgments under domain-specific stakes, bridging causal evaluation with calibrated decision-making.

LLM evaluation and LLM-as-a-judge. The *LLM-as-a-judge* paradigm has become a cornerstone for scalable evaluation, where strong models (e.g., GPT-4, Claude-3.5) act as automated evaluators [21, 7]. Projects such as MT-Bench [21], Chatbot Arena [20], and EvalScope [5] demonstrate that well-designed judging prompts can approximate human preferences with high consistency. Nevertheless, recent meta-studies reveal systematic biases arising from verbosity, prompt ordering, and anthropic effects [3, 9]. Efforts like AutoArena [15] propose multi-dimensional adjudication pipelines to improve robustness. Our framework aligns with this reliability perspective but diverges in scope: instead of subjective preference scoring, we impose a structured causal schema (`Answer / Confidence / Justification`) with statistical significance testing and calibrated risk aggregation across models.

Confidence calibration and risk-sensitive evaluation. Confidence calibration has re-emerged as a critical reliability dimension for both classifiers and generative models [2, 4]. Recent work explores self-consistency [13], chain-of-thought ensemble averaging [8], and entropy regularization [18] to mitigate overconfidence. Calibration has also been linked to model interpretability and trust in high-stakes domains such as finance and healthcare [6, 17]. However, most evaluation schemes optimize for accuracy or log-likelihood without quantifying the asymmetric impact of confident errors. We explicitly operationalize this gap through the *Enterprise Risk Score (ERS)*, combining correctness with risk-weighted confidence deviation, thereby enabling fine-grained auditing of causal reliability under real-world cost functions.

9 Limitations & Ethics

Scope. Our benchmark focuses on binary causal direction tasks with concise textual justifications; future work should extend to multi-causal and temporal-chain reasoning. **Data scale.** Although 120 scenarios cover diverse domains, larger real-world datasets could strengthen statistical significance. **Human oversight.** The system is intended for *decision support*, not autonomous deployment. **Bias and fairness.** Risk weights are domain-specific and may encode implicit value judgments; we release detailed documentation and audit logs. **Privacy and compliance.** All scenarios are anonymized and derived from synthetic or publicly accessible enterprise contexts; no private data are included.

10 Conclusion

This work presented **CausalBench-Enterprise**, a comprehensive benchmark and evaluation framework for assessing *risk-aware causal reasoning* in large language models. We proposed a structured evaluation schema (`Answer / Confidence / Justification`) and the *Enterprise Risk Score (ERS)*, which jointly measure correctness and calibrated confidence under decision-weighted penalties. Across seven frontier models, our analysis revealed that while open-weight LLMs can match closed models in raw accuracy, they often exhibit inflated confidence leading to elevated enterprise risk. The ERS metric exposes such hidden vulnerabilities, offering a more realistic view of reasoning reliability for high-stakes domains.

Beyond numerical metrics, CausalBench-Enterprise contributes a unified and reproducible protocol for causal judgment evaluation, featuring paired bootstrapping, risk calibration curves, and significance-tested model comparisons. By grounding every answer in a constrained causal schema, the

framework promotes transparency, auditability, and trustworthiness in model assessment—core requirements for deploying LLMs in enterprise, finance, and policy decision pipelines.

Future Work. We plan to extend CausalBench-Enterprise to multi-hop and temporal causal reasoning tasks, as well as real-world decision simulations where risk is dynamic and multi-dimensional. Future iterations will explore integrating causal graph alignment, human-in-the-loop calibration, and domain-adaptive risk modeling. In the long term, we aim to build a unified ecosystem for evaluating not only *what* LLMs infer causally, but also *how confident and accountable* those inferences are when consequences matter.

References

[1] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Yonatan Bisk, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

[2] Zheng Geng, Jialin Ye, and Lei Zhang. Calibrating large language models via risk-aware decoding. In *International Conference on Learning Representations (ICLR)*, 2024.

[3] Jiayi Gu, Chenfei Zhao, Tianyang Chen, Yang He, Wei Ding, and Xiaojun Wang. Llm-as-a-judge: A survey on evaluating large language models as evaluators. *arXiv preprint arXiv:2406.09265*, 2024.

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

[5] Yuxin Li, Zihan Wang, and Tianyu Liu. Evalscope: Auditing the evaluators in llm-as-a-judge. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[6] Haoran Lin, Yue Xu, and Jie Zhao. Trustworthy evaluation of llms in high-stakes domains. In *International Conference on Machine Learning (ICML)*, 2024.

[7] Bowen Liu, Ziqi Chen, and Bill Yuchen Lin. Judgebench: A unified benchmark for evaluating llm-as-a-judge consistency. *ACL Findings*, 2024.

[8] Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Detecting llm hallucinations without external references. *ACL*, 2023.

[9] Jinwoo Park, Alice Zhang, and Jia Deng. Bias, variance, and verifiability in llm-as-a-judge systems. *NeurIPS*, 2025.

[10] Melissa Roemmele, Cosmin A. Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, 2011.

[11] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI Conference on Artificial Intelligence*, 2019.

[12] Janek Singh, Martin Potthast, and Benno Stein. Webiscausalqa-22: A benchmark for open-domain causal question answering. In *European Conference on Information Retrieval (ECIR)*, pages 123–138, 2022.

[13] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed Chi, Jeff Dean, and Denny Zhou. Self-consistency improves chain-of-thought reasoning in llms. In *International Conference on Learning Representations (ICLR)*, 2023.

[14] Yizhou Wang, Shizhe Gao, and Hongjie Zhang. Causalvcr: Visual commonsense reasoning via causal graph alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[15] Chen Xu, Weijia Zhao, and Jian Sun. Autoarena: Adaptive multi-dimensional llm evaluation framework. In *AAAI Conference on Artificial Intelligence*, 2025.

[16] Youla Yang, Shuo Liu, and Yifan Zhang. Causalbench: Evaluating causal reasoning in language models. *arXiv preprint arXiv:2312.10211*, 2023.

[17] Hengrui Ye, Sifan Liu, and Bo Wang. Risk-aware evaluation for foundation models in enterprise decision systems. In *AAAI Conference on Artificial Intelligence*, 2025.

[18] Rui Zhang, Xingyu Huang, Tian Zhao, Lin Shi, Xiang Yu, and Dahua Lin. Trustllm: Benchmarking trustworthiness of large language models. *NeurIPS*, 2025.

[19] Rui Zhang, Yuhan Liu, and Chenliang Xu. Visualcausal: Benchmarking multimodal causal understanding in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[20] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Zhaoyang Li, Zhuohan Zhang, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating large language models with human and llm judges. In *NeurIPS Datasets and Benchmarks Track*, 2023.

[21] Lianmin Zheng, Wei-Lin Chiang, Yingbo Zhao, Siyuan Zhuang, Ekin Zelikman, Pieter Abbeel, Hao Wu, and Ion Stoica. Judging llm-as-a-judge: Benchmarking large language model evaluators. *arXiv preprint arXiv:2306.05685*, 2023.