CounterBench: Evaluating and Improving Counterfactual Reasoning in Large Language Models

Yuefei Chen[1], **Vivek K. Singh**[1], **Jing Ma**[2], **Ruixiang Tang**[1]
[1]Rutgers University
[2]Case Western Reserve University
{chen.yuefei, vivek.k.singh, ruixiang.tang}@rutgers.edu, jing.ma5@case.edu

Abstract

Counterfactual reasoning is widely recognized as one of the most challenging and intricate aspects of causality in artificial intelligence. In this paper, we evaluate the performance of large language models (LLMs) in counterfactual reasoning. In contrast to previous studies that primarily focus on commonsense causal reasoning, where LLMs often rely on prior knowledge for inference, we specifically assess their ability to perform counterfactual inference using a set of formal rules. To support this evaluation, we introduce a new benchmark dataset, **CounterBench**, comprising 1.2K counterfactual reasoning questions. The dataset is designed with varying levels of difficulty, diverse causal graph structures, distinct types of counterfactual questions, and multiple nonsensical name variants. Our experiments demonstrate that counterfactual reasoning poses a significant challenge for LLMs, with most models performing at levels comparable to random guessing. To enhance LLM's counterfactual reasoning ability, we propose a novel reasoning paradigm, **CoIn**, which guides LLMs through iterative reasoning and backtracking to systematically explore counterfactual solutions. Experimental results show that our method significantly improves LLM performance on counterfactual reasoning tasks and consistently enhances performance across different LLMs.

Introduction

Counterfactual reasoning, residing at the pinnacle of Pearl's Causal Hierarchy (Pearl and Mackenzie 2018), underpins the "what if" inquiries essential to human cognition and decision-making across critical fields such as healthcare, business, public administration, and science (Gvozdenović et al. 2021; Kyrimi et al. 2025; Kasirzadeh and Smart 2021; Koonce, Nelson, and Shakespeare 2011; Gow, Larcker, and Reiss 2016; Loi and Rodrigues 2012). For example, a consumer who declined an extended warranty may later wonder, "What if I had purchased it, could I have avoided the repair costs?" This illustrates how counterfactual reasoning guides decision-making by evaluating missed opportunities and alternative outcomes (Krishnamurthy and Sivaraman 2002). While traditional causal inference methods (Sharma and Kiciman 2020; Chen et al. 2020; Feder et al. 2022) have enhanced the predictive accuracy, robustness, and explainability of NLP models, recent progress in LLMs has further enriched our ability to capture nuanced causal dependencies (Liu et al. 2024a; Petroni et al. 2019; Liang et al. 2024; Tarassow 2023; Ma 2024; Liu et al. 2024b). These advancements not only demonstrate sophisticated reasoning in tasks ranging from writing to programming but also pave the way toward emulating human-like intelligence and achieving artificial general intelligence (Li and Li 2024; Alwin 2023; Sahota 2023; Bubeck et al. 2023).

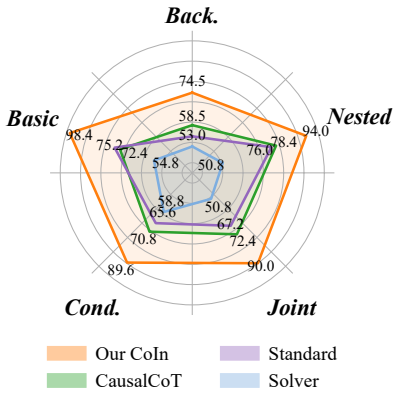


Figure 1: Comparison of accuracy scores on the CounterBench dataset across different strategies: our proposed CoIn paradigm versus baseline approaches (Standard, CausalCoT (Jin et al. 2023), and Solver (Hua et al. 2024)), evaluated using Gemini-1.5-flash. Our CounterBench dataset includes five kinds types. Basic focuses on exploring how a single change in a causal variable. Joint involves simultaneous changes in multiple causes, Nested involves stepwise hypothetical assumptions about multiple variables. Conditional evaluates counterfactuals under observed conditions. And Backdoor involves counterfactual reasoning in the presence of backdoor paths that create confounding between the treatment variable and the outcome.

Despite recent advancements, progress in counterfactual reasoning using LLMs remains constrained by two primary challenges. First, there is currently no dedicated benchmark dataset for rigorously evaluating LLMs' performance on counterfactual tasks, making it difficult to measure the models' capacity to capture nuanced causal relationships. Sec-

ond, even with advanced prompting techniques, Causal CoT (Jin et al. 2023) and other iterative methods, LLMs often struggle to produce logically consistent, contextually appropriate counterfactuals (Ma 2024; Jin et al. 2023; Kıcıman et al. 2023; Zečević et al. 2023). In response, this paper focuses on two key questions:

How well do LLMs handle counterfactual reasoning? The absence of a standardized benchmark dataset has impeded rigorous empirical evaluation of LLMs' capabilities in capturing intricate causal relationships within complex counterfactual reasoning tasks. To address this, we present CounterBench, a comprehensive evaluation framework designed to assess counterfactual reasoning through 1.2K questions encompassing various domains and reasoning types. By systematically evaluating five key dimensions, it demands genuine reasoning beyond pattern recognition or memorized responses. Our experiments expose notable performance limitations in LLMs, even those equipped with advanced inference techniques. Most models like GPT-4o and Deepseek-V3 achieve accuracy of approximately **50%**, equivalent to random guessing. Furthermore, our evaluation of state-of-the-art inference strategies shows only marginal improvements over baseline performance for most models. The models consistently struggle with maintaining logical coherence during multi-step reasoning processes and accurately handling causal relationships in complex scenarios.

How to improve LLMs' counterfactual reasoning abilities? To advance large language models' counterfactual reasoning capabilities, this paper presents CoIn (**Co**unterfactual **In**ference), a novel approach that explicitly tackles the critical challenges of multi-step inference, which remain unresolved by previous methods. CoIn embeds a tailored search algorithm into the reasoning process, guiding LLMs through abduction, action, iterative prediction, and backtracking validation to systematically formalize and explore counterfactual paths. This structured mechanism dynamically assesses the promise of each inference step, enabling reversion to more promising points and ensuring logical consistency, particularly in long-chain causal dependencies. This systematic process substantially improves the accuracy of counterfactual analysis. Experiments on CounterBench demonstrate that CoIn achieves an accuracy of **89.9%**, delivering a nearly **20%** improvement over Gemini-1.5-flash compared to alternative strategies (see Figure 1). The framework of this work is illustrated in Figure 2. The contributions of this work are summarized as follows:

- We build a comprehensive dataset, **CounterBench**. The dataset contains over 1200 long-chain complex counterfactual reasoning questions. The dataset spans multiple difficulty levels, diverse causal graph structures, various types of counterfactual questions, and a wide range of nonsensical variant name combinations.
- We benchmark LLMs with various inference strategies on **CounterBench**, and results reveal that most existing models (e.g., GPT-4o and Deepseek-V3) exhibit limited capabilities in performing counterfactual inference tasks.
- We propose a novel reasoning paradigm **CoIn** guides LLMs through abduction, action, iterative prediction, and backtracking validation to systematically formalize and explore counterfactual reasoning paths. It achieves nearly 90% accuracy on several state-of-the-art LLMs evaluated on CounterBench, representing a 20% improvement over the previous best baseline.

CounterBench

To evaluate the counterfactual reasoning capabilities of LLMs, we introduce a comprehensive benchmarking dataset specifically designed to measure their ability to handle complex causal reasoning tasks. This section details the structure of the dataset, the methodology for query generation, and the benchmarking results analysis.

Dataset Structure

The dataset consists of two main components: a set of counterfactual queries and corresponding binary answers. Formally, the dataset is defined as $\mathcal{D} := \{(q_i, a_i) | i = 1, 2, ..., N\}$, where each q_i is a counterfactual query, and $a_i \in \{\text{yes}, \text{no}\}$ represents the correct answer. Each query is derived from a deterministic Structural Causal Model (SCM) $M = \langle U, V, f \rangle$, where U is the set of exogenous variables with assignments u, V is the set of endogenous variables, and f is the set of structural equations (Pearl 2009). For each $V_i \in V$, we have $V_i = f_i(Pa(V_i), U_i)$, where $Pa(V_i) \subseteq V$ denotes the parents of V_i, and U_i refers to the subset of exogenous variables from U that directly influence the value of V_i. Intervening on a set of variables $X \subseteq V$ and setting them to x modifies the model to M_x, which deterministically defines the values of intervened variables given u. The dataset includes five types of counterfactual queries:

Basic Counterfactual. The basic counterfactual type addresses simple "what-if" scenarios. In this scenario, it is formalized as $Y_x(u)$, which serves as a potential outcome expression. In causal reasoning, potential outcomes refer to the hypothetical results observed when a variable is set to a particular value (Holland 1986). In $Y_x(u)$ expression, Y is the outcome variable,x is the value considered in the hypothetical scenario, and u denotes the context. To illustrate, consider a lawn irrigation system that only activates when the weather is sunny and the soil is dry. In this example, $Y_x(u)$ describes the system's potential behavior when the weather condition x is imposed while the soil condition (context u) remains unchanged. Consequently, when asking whether the system would activate if the weather changed to cloudy, the relevant counterfactual outcome is $Y_{cloudy}(u)$.

Joint Counterfactual. This type involves a counterfactual scenario in which multiple variables are set simultaneously. Formally, it is expressed as $Y_{x,z}(u)$, representing the outcome Y after setting $X = x$ and $Z = z$. For instance, a lawn irrigation system will activate if the weather is sunny, but it also requires dry soil as a trigger. Suppose Z represents the weather condition and X represents the soil moisture condition. The query asks if the irrigation system will activate when the weather changes to cloudy and the sensor detects that the soil is moist meanwhile. The relevant counterfac-

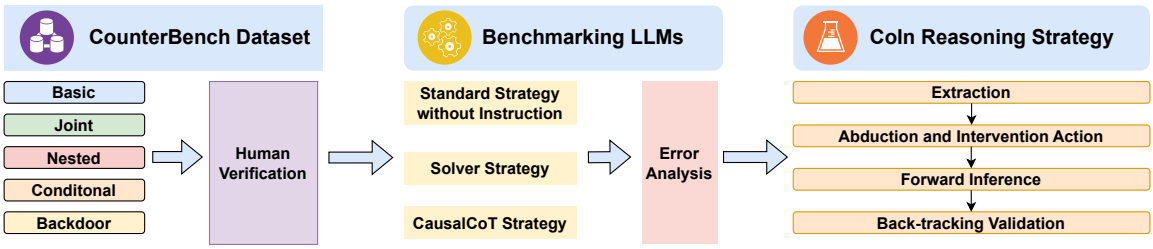


Figure 2: Illustration of the framework. We create CounterBench, a dataset featuring five types of counterfactual questions (basic, joint, conditional, nested, and backdoor). Based on this dataset, we benchmark state-of-the-art LLMs using various inference strategies, conduct comprehensive error analysis, and propose our CoIn reasoning framework featuring systematic inference with validation mechanisms.

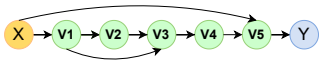
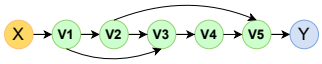
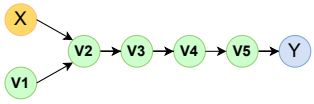
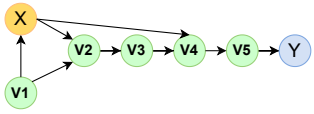
Type	Query Template Example	Causal graph
Basic	We know that X causes V1, V1 causes V2, V2 causes V3, and V3 causes V4, V4 causes V5, V5 causes Y. Would Y occur if not X instead of X?	
Joint	We know that X causes V1, V1 causes V2, V2 and V1 together cause V3, V3 causes V4, V4 and X together cause V5, and V5 causes Y. Would Y occur if not X and not V3?	
Nested	We know that X causes V1, V1 causes V2, V2 and V1 together cause V3, V3 causes V4, V4 and V2 together cause V5, and V5 causes Y. Assume not X, and based on this assumption, further suppose not V4. Would Y occur?	
Conditional	We know that X and V1 together cause V2, V2 causes V3, V3 causes V4, V4 causes V5, V5 causes Y. We observed V1. Would Y occur if not X instead of X?	
Backdoor	We know that V1 causes X, X and V1 together cause V2, V2 causes V3, V3 and X together cause V4, V4 causes V5, and V5 causes Y.	

Table 1: Illustrative Counterfactual Query Types

	Standard						CausalCoT					
Models	**Basic**	**Cond.**	**Joint**	**Nested**	**Back.**	**Avg.**	**Basic**	**Cond.**	**Joint**	**Nested**	**Back.**	**Avg.**
GPT-3 (Davinci-002)	56.8	50.2	48.8	51.6	52.5	51.9	51.2	41.9	51.2	51.6	50.5	49.3
GPT-3 (Babbage-002)	50.0†	50.0†	50.0†	50.0†	47.5	49.6	3.6*	7.6*	1.2*	19.6*	18.5*	9.8*
GPT-3.5	49.6	51.2	50.4	50.0	52.0	50.6	43.6	50.4	53.6	50.0	47.5	49.1
GPT-4o mini	50.0†	50.0†	50.0†	50.0†	52.5	50.4	57.2	66.4	60.0	63.2	50.0	59.8
GPT-4o	50.4	54.4	50.4	54.8	54.0	52.8	80.4	**72.4**	**80.8**	**81.6**	60.5	**75.8**
Claude-3 (Sonnet)	50.4	48.8	50.0	50.8	**59.5**	51.6	59.2	52.0	64.4	60.0	**65.5**	59.0
Claude-3.5 (Haiku)	28.4	24.0	43.6	54.0	51.0	39.8	60.4	65.6	67.2	66.0	61.0	64.2
Gemini-1.5-flash	**75.2**	**65.6**	**67.2**	**76.0**	53.0	**68.0**	72.4	70.8	72.4	78.4	58.5	71.0
Gemini-1.5-flash-8b	50.0†	50.0†	50.0†	50.0†	52.5	50.4	66.8	67.2	65.2	65.2	58.5	64.8
Deepseek-V3	50.4	50.4	50.0	50.0	60.5	51.9	**80.8**	70.4	76.4	77.6	63.5	74.2

Table 2: Model accuracy of standard method and CausalCoT across different reasoning categories. Note: * The average accuracy is only 9.8% because most of responses are not "Yes" or "No" but "incomprehensible", which means LLM cannot follow instruction of CausalCoT instruction well to infer. More details will be explained in the Appendix H. † indicates that the LLM predicts all questions as either "Yes" or "No", leading to a 50% accuracy.

tual outcome is $Y_{cloudy,moist}(u)$. This scenario examines the combined effect of both actions happening simultaneously.

Nested Counterfactual. Nested counterfactual involves sequential dependencies between variables. This is repre-

sented as $Y_{Z_x}(u)$, where an intervention on X affects Z, which in turn impacts Y. For example, if the weather had been cloudy, which is a counterfactual weather state, and under this scenario, the sensor detected moist soil instead of dry soil, would the irrigation system activate? In this framework, Z represents the weather condition, X is the soil moisture reading. Z depends on the counterfactual value of X through the system's structural causal relationships.

Conditional Counterfactual. This type introduces observed conditions into the counterfactual world (Pearl 2009). Formally, it is written as $Y_x(u) \mid Z_x(u) = z$, asking how Y would change if X was set to x while $Z = z$ being observed as a condition. For example, a lawn irrigation system will activate if the weather is sunny, but it also requires dry soil as a trigger. Now the weather is observed as sunny, the query evaluates whether the lawn irrigation system still activate or not if the sensor detects moist soil instead of dryness. Here, $Z = z$ represents the weather is observed as sunny, which is a given condition for reasoning.

Backdoor Counterfactual. This type involves counterfactual reasoning in the presence of backdoor paths that create confounding between the treatment variable and the outcome (Pearl 2009). Formally, it addresses queries of the form $Y_x(u)$ when there exist backdoor paths from X to Y through confounders. In such scenarios, the causal effect cannot be directly identified without controlling for the confounding variables along the backdoor paths. For example, a manager considers evaluating whether a new marketing campaign would increase sales if it is implemented. However, both the decision to launch the campaign and the sales outcome might be influenced by seasonal demand patterns. Here, the backdoor path runs from the marketing campaign through seasonal demand to sales, creating a spurious association. The counterfactual query "Would sales increase if we launched the campaign?" requires accounting for this confounding by either controlling for seasonal effects or using other identification strategies. In our dataset, backdoor counterfactuals test whether LLMs can distinguish between genuine causal effects and spurious correlations when reasoning about alternative scenarios.

Query Generation and Quality Assessment

Each query consists of background information and a specific question. Table 1 illustrates how samples are generated using various deterministic counterfactual query types. The background is constructed with causal graphs and story templates, and variable names are replaced by nonsensical, artificially generated words (e.g., "Kelp," "Ziklo") to prevent models from relying on memorized knowledge. In this way, we force LLMs to engage in causal reasoning rather than using prior knowledge in pretraining data. The dataset also features balanced distributions in multiple dimensions, with binary responses evenly split between 50% "Yes" and 50% "No." This balance extends across different question types and difficulty levels, ensuring a uniform response distribution within each category. The dataset consists of 1,200 questions, categorized into five distinct types, with each type containing 200 or 250 questions. Within each type, there is an equal distribution of answers, comprising 100 "Yes" responses and 100 "No" responses or 125 "Yes" and 125 "No". Additionally, the dataset is stratified based on five levels of difficulty, determined by the number of events present in each question, ranging from 5 to 9. Each difficulty level includes 240 questions, maintaining a balanced distribution of answers with 120 "Yes" and 120 "No". We also conduct a human evaluation on these queries, with further details provided in Appendix B.

Benchmarking LLMs on CounterBench

We conducted comprehensive experiments to systematically evaluate the performance of current LLMs on counterfactual reasoning tasks, demonstrating their capabilities using state-of-the-art reasoning techniques.

Models. The tested LLM models include GPT-3.5 turbo, GPT-4o, GPT-4o mini, Davinci-002, Babbage-002 (OpenAI 2024), Claude 3.5 Haiku, Claude 3 Sonnet (Anthropic 2024), Deepseek-V3 (DeepSeek 2024) and Gemini-1.5-Flash and Gemini-1.5-Flash-8B (Google 2024).

Reasoning Strategies. In our baseline evaluations, we employed two distinct reasoning strategies to assess these models. The first relied on standard prompting methods without specialized instructions. The second used the advanced CausalCoT approach (Jin et al. 2023), an extension of the Chain-of-Thought prompting paradigm (Wei et al. 2022). By integrating a systematic derivation process, including causal graph extraction, query type classification, data collection, and formalization, CausalCoT ensures robust logical consistency and high reasoning accuracy.

Evaluation Settings. Within our evaluation framework, responses are classified into three distinct categories, "Yes", "No", and "Incomprehensible". The latter encompassing responses that are either ambiguous or lack clear meaning, typically manifesting when no answer is detected, such as NULL returns or mere query echoes. During inference, we set the temperature at 0. We employ inference accuracy as our primary performance metric.

Experimental Results

As shown in Table 2, the results indicate that without specific instructions, most LLMs struggle with counterfactual reasoning, performing no better than random guessing in terms of accuracy. Specifically, for model GPT-4o mini, we observed consistent predictions of either "Yes" or "No," resulting in a 50.0% accuracy in the first four kinds of questions. Among all tested models, Gemini-1.5-flash achieved the highest baseline performance with an accuracy of 68.0%. Although the CausalCoT approach is designed to enhance the causal reasoning capabilities of LLMs, our empirical findings suggest that it does not significantly improve their performance in counterfactual reasoning tasks. Most models showed minimal or no improvement, as exemplified by GPT-3.5 Turbo. The best performance model in the CausalCoT is GPT-4o, achieved an accuracy of only 75.8%.

Error Analysis. To systematically analyze the limitations of existing approaches, we conducted an error analysis on responses generated by CausalCoT. Our analysis focuses on

three key components: causal data collection, inference process, and conclusion derivation. Through careful examination of each component, we identified three primary categories of errors: **Wrong causal relationships:** This error occurs when LLMs cannot construct accurate causal graphs or extract known values from background information. **Wrong inference process:** This happens when LLMs, despite correctly identifying causal relationships, make incorrect predictions of the target event Y. **Wrong conclusion:** This type of error arises when LLMs reach contradictory final answers, even after correctly computing the value of Y. The distribution of these error categories is illustrated in Figure 3. Notably, 86% of errors occur in the inference process, revealing that even with well-constructed causal graphs, LLMs struggle significantly with deriving accurate predictions through reasoning.

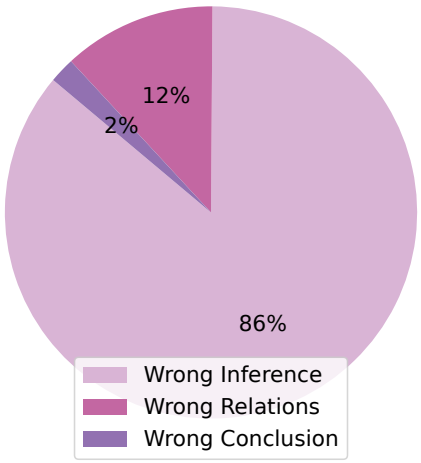


Figure 3: Error Analysis of CausalCoT.

Proposed Reasoning Strategy

As discussed in previous section, the primary challenge for large language models is to minimize incorrect inferences, which are a major source of errors. To address this challenge, we propose **CoIn** (Counterfactual Inference), a systematic reasoning framework that guides large language models through structured problem-solving instead of relying on intuitive shortcuts or memorized patterns. Our approach transforms counterfactual queries into a five-phase algorithmic process, mirroring how humans naturally approach "what if" questions (Sel et al. 2023): first understanding what actually happened, then imagining the alternative scenario, systematically working through the consequences, and finally double-checking the logic. This structured approach significantly reduces reasoning errors by breaking down complex problems into manageable steps with built-in validation. An example of proposed paradigm is provided in Appendix A.

The CoIn framework consists of five key phases: **Extraction:** Extract Counterfactual Information from the given natural language facts. **Abduction:** Infer the underlying conditions from observed facts; **Intervention Action:** Apply the hypothetical changes specified in the query; **Forward Inference:** Systematically trace through the causal consequences; **Back-tracking Validation:** Verify the logical consistency of the entire reasoning chain. Each phase serves a specific purpose in ensuring accurate counterfactual reasoning, and they provide a robust methodology for handling complex causal dependencies together. Below, we describe each phase in detail, explaining its role and how it contributes to the overall process.

Extraction

In the first phase, we focus on systematically gathering all relevant information explicitly stated in the scenario. The process begins with constructing the causal graph by identifying relationships between events and representing them in a clear "$event\ 1 \rightarrow event\ 2$" format, which eliminates potential ambiguities. Next, we collect the given values for each variable from both background information and questions, where these values indicate whether specific events occur or not. Crucially, this phase maintains strict adherence to explicitly stated information, avoiding any unsupported inferences or assumptions in favor of a rigorous and unbiased data collection process.

Abduction

This phase focuses on inferring the posterior constraints over the exogenous noise variables, equivalently, constraints over parent assignments that make the observed factual world consistent with the structural equations. For each observed variable V with value v_{obs}, we invert its structural equation $V := f_V(\text{Parents}(V), U_V)$ to obtain either a unique solution for U_V or a feasible set over U_V given the parents. In deterministic logical models, this is often conveniently carried out by deducing parent assignments that must hold for v_{obs} to be true. The resulting values are stored as the factual world knowledge base and will be held fixed in subsequent Intervention Action and Forward Inference, ensuring counterfactuals are evaluated in the same world.

Intervention Action

In this phase, the framework applies the counterfactual interventions described in the query. This involves modifying the original set of causal rules by replacing the equations for the intervened variables with constant values, resulting in an updated set of rules. The interventions are incorporated into the knowledge base, effectively making a precise alteration to the causal graph. This phase captures the core "what if" element of the query, allowing the framework to simulate hypothetical worlds in a controlled way. It focuses on specific changes, which streamlines the exploration by limiting the search to paths directly affected by the intervention, similar to how efficient searches eliminate unnecessary branches.

Forward Inference

During this iterative phase, the framework predicts values for unobserved variables by selecting nodes in the causal

graph whose parent variables are already known in the knowledge base and evaluating their updated equations. Beginning with the intervened variables and the inferred noise terms, it gradually computes the effects on downstream variables until it reaches the target variable Y. If a node's value cannot be calculated due to missing information about its parents, the framework chooses another suitable node and continues the process until Y is determined. This forward progression mimics a depth-first exploration of causal chains, enabling the framework to dynamically construct and assess potential outcomes. By focusing on nodes that can be computed immediately, it navigates the dependency graph efficiently, steering clear of unproductive paths and promoting a methodical advancement toward the solution. The details of this phase is in the Appendix G.

Back-tracking Validation

To confirm that the predicted values are logically consistent, this final phase retraces the steps through the knowledge base and re-evaluates the equation for each non-noise variable using the predicted values. For every such variable V, it recalculates the expected value based on the updated equation and checks if it matches the previously stored value. If any mismatch occurs, the framework signals an error, highlighting a potential issue in the earlier reasoning that may need reevaluation. This validation serves as a protective measure against errors that could accumulate during the process, akin to retracing a path to confirm its validity.

Experiments

Experiment Setup

We adopted the same LLMs as mentioned in Section 3 for our experiments. To establish baselines, we implemented both CausalCoT (Jin et al. 2023) and standard solver strategies. The latter integrates external tools into the chain-of-thought process, as described in (Hua et al. 2024). Specifically, this approach combines LLMs with Structural Causal Model (SCM) tools (Pearl 2009) for causal inference. The study introduces CausalTool, a suite of 10 inference tools designed for various causal tasks. It leverages LLMs to classify causal questions, extract causal graphs and formalized data, and route them to the appropriate tools for inference, with the final answer generated by the LLM. [1] During inference, the temperature is set to zero.

Main Result

The comprehensive performance comparison across all models is presented in Table 3. Our approach demonstrates notable improvements over existing methods across the model spectrum, with particularly noteworthy performance gains achieved by smaller language models, including GPT-4o mini, Claude-3.5 Haiku, and Gemini-1.5-flash-8b. For instance, our method enables GPT-4o mini to achieve an accuracy of 79.9%, surpassing the performance of several larger models without CoIn enhancement. As detailed in Table 4, taking GPT-4o mini as an example, CoIn achieves superior performance across all five types of counterfactual questions, with particularly better results on basic questions compared to more complex variants. Additionally, state-of-the-art LLMs such as GPT-4o, Gemini-1.5-flash, and Deepseek-V3 achieve remarkable accuracy approaching or exceeding 90% when augmented with our method. Taken GPT-4o as example, Our strategy improves the accuracy of the model from 75.8% to 89.4%, demonstrating CoIn's effectiveness in guiding LLMs through algorithm to explore paths step-by-step. The results indicate that contemporary LLMs, when equipped with our strategy, can effectively resolve most formal complex counterfactual problems. The details of all performance of our results are presented in Appendix C. Moreover, in Appendix D, we conducted error analysis. The analysis reveals a substantial reduction in errors of inference process. Moreover, we also examine the impact of complex causal relationships on outcomes. We found that accuracy decreases as the number of variants increases for CausalCoT and CoIn. The details are shown in the next section.

Model	Standard	CausalCoT	Solver	Ours
GPT-3 (Davinci-002)	**51.9**	49.3	50.1	49.6
GPT-3 (Babbage-002)	**49.6**	9.8	47.9	45.8
GPT-4o mini	50.4	59.8	47.2	**79.9**
GPT-4o	52.8	75.8	51.4	**89.4**
GPT-3.5 turbo	50.6	49.1	49.6	**58.9**
Claude-3 (Sonnet)	51.6	59.0	51.8	**89.8**
Claude-3.5 (Haiku)	39.8	64.2	48.3	**79.1**
Gemini-1.5-flash	68.0	71.0	52.8	**89.9**
Gemini-1.5-flash-8b	50.4	64.8	50.3	**83.9**
Deepseek-V3	51.9	74.2	49.3	**91.8**

Table 3: Model accuracy on CounterBench. We report the average accuracy for four inference strategies: Standard, CausalCoT, Solver, and CoIn.

Methods	Basic	Cond.	Joint	Nested	Back.	Avg.
Standard	50.0	50.0	50.0	50.0	52.5	50.4
CausalCoT	57.2	66.4	60.0	63.2	50.0	59.8
Solver	35.2	54.4	50.4	50.0	45.5	47.2
Ours	**82.8**	**79.2**	**80.0**	**80.4**	**76.5**	**79.9**

Table 4: Accuracy of GPT-4o mini across five query types in the CounterBench.

Validating Generalization Ability

In this section, we evaluate the generalization capability of CoIn using the CLADDER dataset (Jin et al. 2023). CLADDER is a dataset focus on the causal reasoning questions. We utilize the CLADDER dataset to determine if the proposed method can be extended beyond the CounterBench dataset. Unlike our dataset, which focuses on formal rules, CLADDER includes examples that utilize common sense knowledge rather than causal inference abilities to answer queries.

[1] Since the source code for CausalTool is not publicly available, we re-implemented its counterfactual inference procedure based on descriptions in the original paper.

Specifically, it encompasses both commonsense and anti-commonsense scenarios, allowing us to explore whether CoIn remains effective under the influence of pretraining knowledge in LLMs. We conduct experiments on the counterfactual subset of CLADDER. We applied the Gemini-1.5 and Gemini-1.5-8b models to both commonsense and anti-commonsense queries, with the results displayed in Figure 4 and detailed examples in Appendix E. These results demonstrate that performance is consistently stable across different reasoning paradigms, suggesting that pretraining knowledge has a limited impact on the CoIn counterfactual reasoning capabilities. Furthermore, our method achieves an accuracy of 78.98%, outperforming both CausalCoT at 64.77% and the Standard method at 64.20%. This performance underscores CoIn's generalizability and its potential for broader application in various counterfactual reasoning tasks.

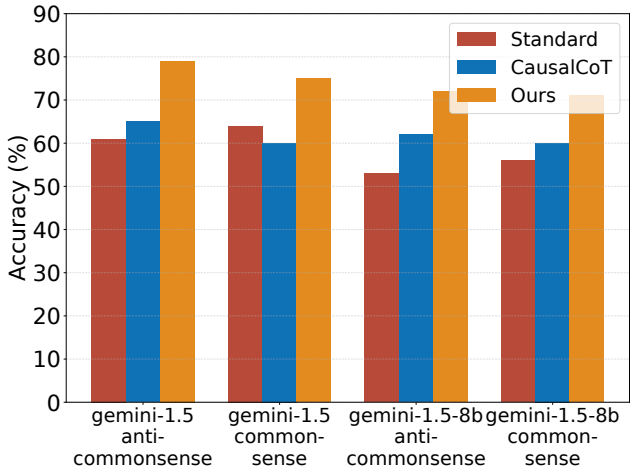


Figure 4: Accuracy comparison between Standard, CoIn, and CausalCoT method in Anti-commonsense and Commonsense Dataset.

Related Work

Counterfactual Reasoning. Counterfactual reasoning explores how outcomes change when certain variables are altered from their historical states. In Structural Causal Models (SCMs), Pearl's (Pearl 2009) "surgery" and do-calculus provide systematic ways to infer intervention outcomes, highlighting deep causal knowledge required for accurate inference. Counterfactuals can be deterministic or probabilistic: deterministic settings yield predictable outcomes from given conditions, while probabilistic models incorporate inherent uncertainties. These methods have gained traction in domains like social sciences, where they assess alternative policy outcomes and study causal mechanisms in observational data (Morgan 2015), and in medicine, where they enable personalized treatment and decision support (Johansson, Shalit, and Sontag 2016; Shalit, Johansson, and Sontag 2017; Louizos et al. 2017; Yoon, Jordon, and Van Der Schaar 2018). In artificial intelligence, counterfactual reasoning is crucial for interpretability and fairness, enabling models to generate alternative scenarios and assess decision-making robustness. Although recent efforts extend counterfactual reasoning to LLMs (Jin et al. 2023), significant challenges persist, particularly regarding complex variable relationships in high-dimensional text data. Consequently, bridging the gap between textual complexity and robust causal inference remains a focal point for future research.

LLMs in Counterfactual Learning. With the rapid evolution of LLMs, the research community has increasingly focused on their ability to perform causal inference (Zhang et al. 2023; Ashwani et al. 2024). A prominent example is Causal Agent, an agent-based LLM framework that merges an LLM with causal tools for complex tasks (Han et al. 2024). While it excels at identifying causal associations and conducting interventions, it largely omits counterfactual reasoning, limiting its applicability to more advanced scenarios. Current efforts to integrate counterfactual reasoning into LLMs typically follow two paths. First, commonsense-based approaches leverage background knowledge to imagine scenarios that defy established facts (Ning et al. 2024; Chatzi et al. 2024; Musi and Palmieri 2024; Vicuna 2023), such as positing alternative historical outcomes. Second, graph-based methods employ formal causal graphs and external Python packages for computations, as seen in CausalTool (Hua et al. 2024). Although these methods effectively incorporate structured causal information, they often offload key calculations outside the LLM.

Conclusion

In this work, we develop and extend CounterBench, a counterfactual reasoning dataset with five problem types for LLM evaluation. Our findings reveal that most LLMs perform near-randomly, with state-of-the-art methods showing minimal improvement. To address these challenges, we propose CoIn, a reasoning paradigm inspired by formal causal inference principles and planning strategies. CoIn guides LLMs through iterative thinking and backtracking to explore reasoning paths more effectively. Our approach significantly enhances counterfactual reasoning capabilities of LLMs.

Acknowledgments

We acknowledge the support of the Ph.D. student Canyu Gao from Rutgers University's School of Public Affairs and Administration in organizing human dataset evaluation. We further acknowledge the computational resources provided through ACCESS, an NSF-funded advanced program.

References

Alwin. 2023. Understanding Causal AI: Bridging The Gap Between Correlation And Causation. https://www.alwin.io/causal-ai. Accessed: 2025-01-06.

Anthropic. 2024. Claude. https://www.anthropic.com/api. Accessed: 2025-01-06.

Ashwani, S.; Hegde, K.; Mannuru, N. R.; Sengar, D. S.; Jindal, M.; Kathala, K. C. R.; Banga, D.; Jain, V.; and Chadha, A. 2024. Cause and effect: Can large language models truly

understand causality? In *Proceedings of the AAAI Symposium Series*, volume 4, 2–9.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Chatzi, I.; Benz, N. C.; Straitouri, E.; Tsirtsis, S.; and Gomez-Rodriguez, M. 2024. Counterfactual token generation in large language models. *arXiv preprint arXiv:2409.17027*.

Chen, H.; Harinen, T.; Lee, J.-Y.; Yung, M.; and Zhao, Z. 2020. Causalml: Python package for causal machine learning. *arXiv preprint arXiv:2002.11631*.

DeepSeek. 2024. DeepSeek: AI-Powered Search Engine. Accessed: 2025-02-15.

Feder, A.; Keith, K. A.; Manzoor, E.; Pryzant, R.; Sridhar, D.; Wood-Doughty, Z.; Eisenstein, J.; Grimmer, J.; Reichart, R.; Roberts, M. E.; et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10: 1138–1158.

Google. 2024. Gemini. https://gemini.google.com/. Accessed: 2025-01-06.

Gow, I. D.; Larcker, D. F.; and Reiss, P. C. 2016. Causal inference in accounting research. *Journal of Accounting Research*, 54(2): 477–523.

Gvozdenović, E.; Malvisi, L.; Cinconze, E.; Vansteelandt, S.; Nakanwagi, P.; Aris, E.; and Rosillon, D. 2021. Causal inference concepts applied to three observational studies in the context of vaccine development: from theory to practice. *BMC Medical Research Methodology*, 21: 1–10.

Han, K.; Kuang, K.; Zhao, Z.; Ye, J.; and Wu, F. 2024. Causal agent based on large language model. *arXiv preprint arXiv:2408.06849*.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396): 945–960.

Hua, Z.; Xing, S.; Jiang, H.; Wei, C.; and Wang, X. 2024. Improving Causal Inference of Large Language Models with SCM Tools. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 3–14. Springer.

Jin, Z.; Chen, Y.; Leeb, F.; Gresele, L.; Kamal, O.; Zhiheng, L.; Blin, K.; Adauto, F. G.; Kleiman-Weiner, M.; Sachan, M.; et al. 2023. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh conference on neural information processing systems*.

Johansson, F.; Shalit, U.; and Sontag, D. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*, 3020–3029. PMLR.

Kasirzadeh, A.; and Smart, A. 2021. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 228–236.

Kıcıman, E.; Ness, R.; Sharma, A.; and Tan, C. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Koonce, L.; Nelson, K. K.; and Shakespeare, C. M. 2011. Judging the relevance of fair value for financial instruments. *The Accounting Review*, 86(6): 2075–2098.

Krishnamurthy, P.; and Sivaraman, A. 2002. Counterfactual thinking and advertising responses. *Journal of Consumer Research*, 28(4): 650–658.

Kyrimi, E.; Mossadegh, S.; Wohlgemut, J. M.; Stoner, R. S.; Tai, N. R.; and Marsh, W. 2025. Counterfactual reasoning using causal Bayesian networks as a healthcare governance tool. *International Journal of Medical Informatics*, 193: 105681.

Li, J.; and Li, X. 2024. Relation-First Modeling Paradigm for Causal Representation Learning toward the Development of AGI. arXiv:2307.16387.

Liang, W.; Zhang, Y.; Wu, Z.; Lepp, H.; Ji, W.; Zhao, X.; Cao, H.; Liu, S.; He, S.; Huang, Z.; et al. 2024. Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*.

Liu, J.; Cao, S.; Shi, J.; Zhang, T.; Nie, L.; Hu, L.; Hou, L.; and Li, J. 2024a. How Proficient Are Large Language Models in Formal Languages? An In-Depth Insight for Knowledge Base Question Answering. In *Findings of the Association for Computational Linguistics ACL 2024*, 792–815.

Liu, X.; Xu, P.; Wu, J.; Yuan, J.; Yang, Y.; Zhou, Y.; Liu, F.; Guan, T.; Wang, H.; Yu, T.; et al. 2024b. Large language models and causal inference in collaboration: A comprehensive survey. *arXiv preprint arXiv:2403.09606*.

Loi, M.; and Rodrigues, M. 2012. A note on the impact evaluation of public policies: the counterfactual analysis.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

Ma, J. 2024. Causal inference with large language model: A survey. *arXiv preprint arXiv:2409.09822*.

Morgan, S. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

Musi, E.; and Palmieri, R. 2024. The Fallacy of Explainable Generative AI: evidence from argumentative prompting in two domains. In *CEUR Workshop Proceedings*, volume 3769, 59–69.

Ning, X.; Lin, Z.; Zhou, Z.; Wang, Z.; Yang, H.; and Wang, Y. 2024. Skeleton-of-thought: Prompting LLMs for efficient parallel generation. In *The Twelfth International Conference on Learning Representations*.

OpenAI. 2024. Models. https://platform.openai.com/docs/models. Accessed: 2025-01-06.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic books.

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Sahota, N. 2023. Causal AI: Bridging the Gap Between Correlation and Causation. https://www.neilsahota.com/causal-ai-bridging-the-gap-between-correlation-and-causation/. Accessed: 2025-01-06.

Sel, B.; Al-Tawaha, A.; Khattar, V.; Jia, R.; and Jin, M. 2023. Algorithm of thoughts: Enhancing exploration of ideas in large language models. *arXiv preprint arXiv:2308.10379*.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, 3076–3085. PMLR.

Sharma, A.; and Kiciman, E. 2020. DoWhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*.

Tarassow, A. 2023. The potential of LLMs for coding with low-resource and domain-specific programming languages. *arXiv preprint arXiv:2307.13018*.

Vicuna. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. https://vicuna.lmsys.org/. Accessed: 2023.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yoon, J.; Jordon, J.; and Van Der Schaar, M. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*.

Zečević, M.; Willig, M.; Dhami, D. S.; and Kersting, K. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.

Zhang, C.; Bauer, S.; Bennett, P.; Gao, J.; Gong, W.; Hilmkil, A.; Jennings, J.; Ma, C.; Minka, T.; Pawlowski, N.; et al. 2023. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*.