CoBe: Counterfactual Benchmark for Text Editing

Arvind Raghavan* **Inwoo Hwang*** **Shreyas Havaldar***
Kai-Zhan Lee **Aurghya Maiti** **Yushu Pan**
Jeffrey Wu **Mingxuan Li** **Elias Bareinboim**
Causal Artificial Intelligence Lab
Columbia University

Abstract

Counterfactual text editing, defined as the ability to edit a text by incorporating a hypothetical change in the scenario being described, is a core reasoning skill, particularly relevant in domains like medicine, law and business. This task requires causal understanding of the underlying scenario. By contrast, the task of *associational text generation* relies on correlational reasoning. Given a text and an altered condition, it generates similar text that co-occurs with the alternative condition in the training corpus. Despite the safety-critical requirement of counterfactual reasoning skills, prior benchmarks do not clearly distinguish between these two different modes of reasoning. We introduce Counterfactual Benchmark (`CoBe`): a new testbed designed to assess the native ability of language models to perform counterfactual edits to a given scenario. Our dataset consists of 2377 questions, each testing counterfactual reasoning across a variety of domains in science, human behavior, and common-sense. We test a suite of frontier and smaller models on our benchmark, with frontier models only achieving an avg. score of 53.94%. [2]

1 Introduction

Counterfactual reasoning is a vital tool for understanding the world and the consequences of performing alternative actions in different scenarios [6, 27, 36, 46, 48]. Counterfactual questions such as *"would the patient have succumbed to an infection had she not been moved to the crowded ward?"* or *"would oil prices be lower today if the UAE had quit OPEC one month earlier?"* are integral to the way humans think, communicate, and construct explanations for why events occurred. They form the basis of our intuitions of ethics, regret, credit, and responsibility [23]. Counterfactuals are acknowledged to be important for personalized decision-making [7, 43], mediation analysis [4, 47, 53], fairness analysis [49, 64], and explainability [24, 35]. This mode of reasoning is thus also important in the context of large language models (LLMs), given their increasing prominence as embedded tools in modern machine learning pipelines.

Counterfactual logic belongs to the third level of Judea Pearl's celebrated *ladder of causation*, also known as the Pearl Causal Hierarchy [8, 48]. Counterfactual text editing is a specific manifestation of a broader causal reasoning capability. Causal intelligence spans many activities, including causal understanding, explanation, decision-making, generalization, representation learning, and causal simulation [6, Ch. 1]. Within the Pearl Causal Hierarchy, counterfactual reasoning occupies the third layer, involving queries about what would have happened under a hypothetical intervention, given that some possibly conflicting facts actually occurred [6, Ch. 5, 6]. The task studied here focuses only on a small but concrete slice of this space: whether a language model can edit a textual scenario in a way that respects the causal semantics of a counterfactual intervention.

*Equal contribution.
[2]We provide anonymized link to the sample data.

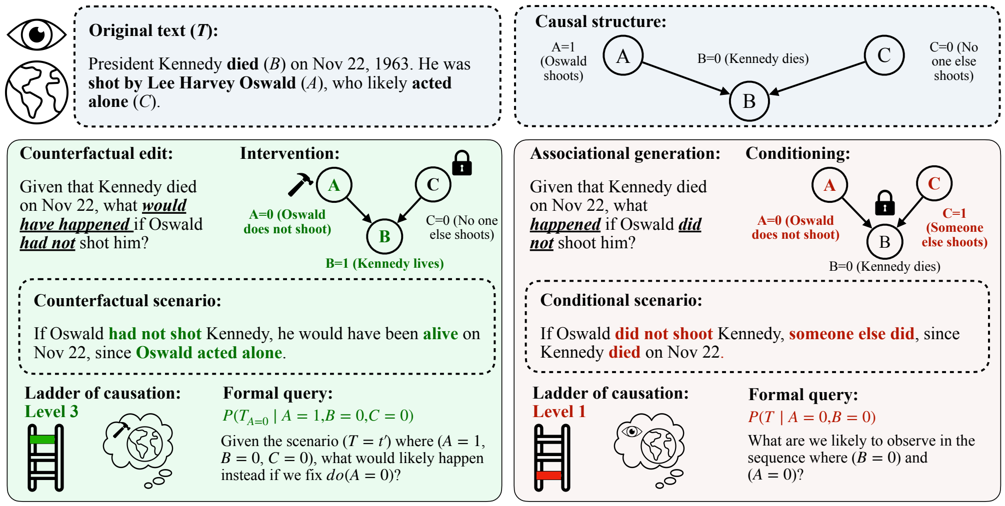

Figure 1: Counterfactual edits to a text, belonging to Level 3 of Pearl's *ladder of causation*, situate a factual scenario (including fixing background factors) and reason about what would happen as a result of a hypothetical intervention. By contrast, associational text generations belong to Level 1, and generate alternative scenarios correlated with the altered condition (possibly changing background factors). **Note**: Counterfactual scenarios need not always be intuitive or likely in the real world.

Counterfactual text editing. Building on Pearl's structural account of counterfactuals, we define this natural language process (NLP) task to involve three steps: *(1) situating ourselves in the factual events described in a text, and updating our beliefs about background factors that may have produced these events; (2) imagine a hypothetical intervention (described in conversational language) on some events which occurred in the factual scenario; (3) reason about the downstream causal effects of enacting this change in the original scenario (given our updated beliefs about background factors), and produce edited text reflecting this understanding*.

More formally, we are given a scenario described in text $T = t$ (e.g., the passage describing the shooting of Kennedy by Oswald, in Fig. 1). The Oswald/Kennedy example is a classical illustration of the contrast between indicative and subjunctive conditionals, going back at least to Adams and later discussions in the Stalnaker-Lewis tradition [2, 36, 55]. We use it here as a simple entry point into the structural account of counterfactual text editing. The scenario involves events A, B, C, where some events potentially cause others. In our example, Oswald shoots, $A = 1$, no one else shoots, $C = 0$, and Kennedy dies, $B = 0$. An example of a counterfactual edit is to ask *given that Kennedy died and Oswald acted alone, what would have happened had Oswald not shot him?* (Ans: He lives.) This counterfactual text edit involves the three steps: **(1)** we situate ourselves in the given scenario ($A = 1, B = 0, C = 0$) and update our beliefs about any background factors that could have produced this set of events, such as facts about Kennedy's public schedule and security environment; **(2)** imagining an external intervention that miraculously stops Oswald from shooting, written symbolically as ($do(A = 0)$); and **(3)** producing text $T = t'$ that reasonably reflects the downstream effects of this edit in the original scenario. In this example, we are thus sampling from the distribution $P(T_{A=0} \mid A = 1, B = 0, C = 0)$. Notably, this edit should hold fixed any background factors and other events that are not causally downstream of the edited events. This protocol is inspired by the 3-step process of *Abduction-Intervention-Prediction* formalized by Judea Pearl [46].

Associational text generation. By contrast, associational reasoning belongs to the first level of the ladder of causation. This task instead *generates text by merely conditioning on an alternative event to ask what scenarios are statistically likely to co-occur with the given change*. For instance, an associational query asks *given that Kennedy died, what did happen if Oswald did not shoot him?* (Ans: Someone else shot him.) We are thus generating alternative text by sampling from $P(T \mid A = 0, B = 0)$, possibly changing background conditions and prior events to maintain correlation.

Counterfactual text editing is highly useful both for probing model behavior and for evaluation and data augmentation [10, 38, 60, 62]. However, a growing body of work documents that LLMs

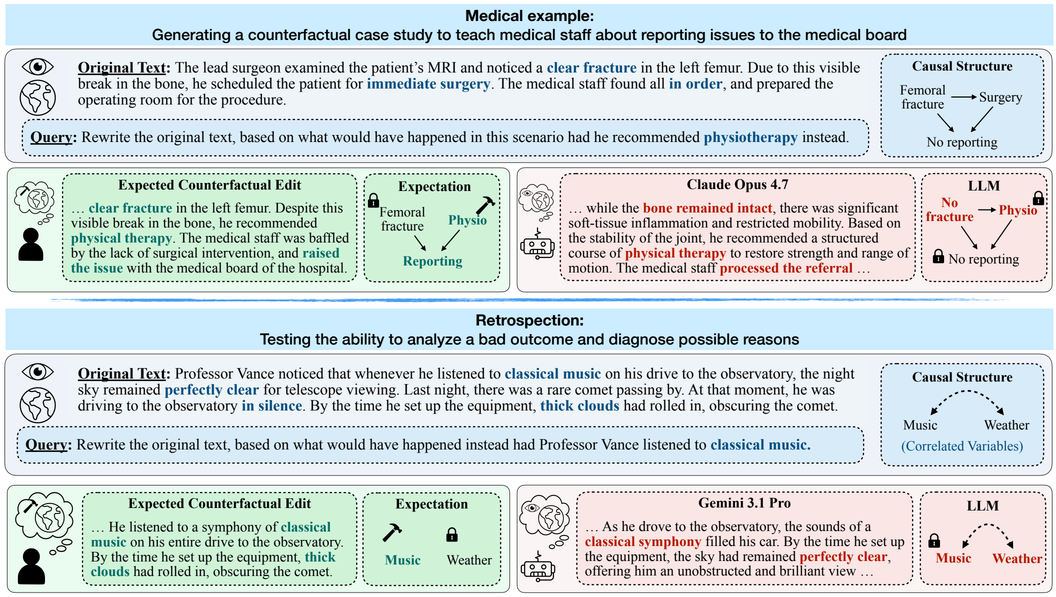

Figure 2: Frontier LLMs were given two scenarios and were prompted, in conversational language, to incorporate a counterfactual edit. Models failed to correctly account for causal dependencies among events in the original scenario.

rely predominantly on *associational* signals rather than genuine causal structure, learning surface heuristics from their pretraining distribution [32, 61] and often learning shortcuts to predict "right but for the wrong reasons" [15, 19, 40]. This gap directly limits their causal reasoning: state-of-the-art models perform near chance on formal counterfactual inference [12, 17], conflate temporal precedence with causation, and fail to derive causal relations from counterfactuals, a deficiency that is notably observed to not improve with scale [32, 33]. This gap is not surprising, since *counterfactual* and *associational* reasoning are fundamentally different modalities and it is a known impossibility result that Level 3 questions cannot be answered in general with Level 1 reasoning alone [6, 8]. Furthermore, recent work [50] showcases that even Level 1 reasoning of LLMs is brittle, which supports the hypothesis that Level 3 questions are also going to be hard to answer accurately.

This gap has serious implications for AI safety and trustworthiness. Models lacking counterfactual *"what if?"* capabilities produce unreliable self-explanations that humans cannot use to predict behavior on perturbed inputs [11, 18]. In safety-critical applications such as medicine, law, and policy, decisions hinge on hypothetical reasoning about treatments, alternative actions, and contested causes. Reliance on spurious correlations and heuristic shortcuts risks systematic, hard-to-predict errors [33, 34]. Even when attempting to leverage counterfactuals to mitigate harms, naive approaches can backfire: for instance, naively augmented data can introduce new biases [30, 42, 57]. In Fig. 2, we illustrate two instances from different task domains, where frontier models mistake a natural language counterfactual edit query for a associational generation query and produce misaligned output.

In light of this motivation, we introduce **Co**unterfactual **Be**nchmark (CoBe): a benchmark for testing the native ability of LLMs to produce causally-valid counterfactual edits to text, when prompted to do so in *conversational language* (i.e. without needing prompt engineering). The key highlights of our dataset are as follows:

- We curate a set of 2377 **samples** (1007 public), each containing, in natural language, (i) a short text describing a scenario, and (ii) a set of three prompts requesting, in colloquial language, a **counterfactual edit** to the paragraph.
- We provide the **evaluation criteria** for each question to determine the validity of LLM responses, offering a principled and causally-grounded way of evaluating counterfactual texts, verified by human experts.
- Our dataset covers questions in **diverse domains**: Health & Medicine, Science & Technology, Economics & Finance, Arts & Entertainment, Engineering, History & Geography, and general Human Behavioral scenarios.

- We investigate a comprehensive suite of flagship and open-source models, with results summarized in Sec. 4.3. Notably, **frontier models only score an avg. of 53.94%**, highlighting significant room for improvement.
- We taxonomize the types of failure modes we observed, and characterize the idiosyncratic error profiles of different models. For instance, we observe that Gemini-3.1-Pro is heavily optimized for correlational consistency, and GPT-5.4-Pro struggles to incorporate background information from the scenario in its edits.

Our goal is to test the native ability of LLMs to *counterfactually edit text with conversational prompts articulated by lay users, without needing technical jargon or specialized prompts*. We provide, to the best of our knowledge, the first rigorous testbed, grounded in formal causal semantics and with high-quality evaluation criteria, for this capability. To the best of our knowledge, we are also the first to characterize the distinct error-profiles of different models on this important task.

Further, our evaluation criteria allow for multiple valid responses that respect counterfactual semantics, and works in the domain of narrative stories expressed in natural language. This allows `CoBe` to be seamlessly integrated as a training signal into pipelines designed to improve LLM reasoning. As we point out in the next section, several of pre-existing methods either (a) bypass narrative stories and focus on formally-verifiable domains like math and code; or (b) rely on very simple causal narratives with single crowdsourced answers that may leave such training methods brittle.

2 Related Work

We discuss below relevant prior work grouped under common themes. In Table 1, we compare these to our benchmark, `CoBe`, across five dimensions: does the benchmark (1) ground itself in rigorous 3-step definition of counterfactual text editing (see Sec. 1 for a definition); (2) test understanding of narrative stories (vs. mathematics or coding questions phrased in text); (3) use complex causal structures, (e.g., involving unobserved confounding); (4) evaluate text generation (vs. multiple choice selection); and (5) include evaluation criteria verified by experts or formal solvers (vs. crowd-sourced).

Table 1: Comparison to prior benchmarks.

Prior work	Ctf-editing 3 steps	Narrative scenarios	Complex graphs	Generative task	Expert/formal verification
General benchmarks	✗	✓	–	✓	✓
TimeTravel [51]	✓	✓	✗	✗	✗
CRASS [17]	✗	✓	–	✗	✗
CLadder [31]	✓	✗	✗	✓	✓
`CoBe` **(our work)**	✓	✓	✓	✓	✓

General benchmarks. It is now standard for newly-released model cards to describe performance on general benchmarks like MMLU [25], BigBench [54], GSM8K [14], MATH [26], ARC-AGI [13], HellaSwag [63], GPQA [52], BigBenchHard [56] etc., which test a combination of domain knowledge, and abstract and common-sense reasoning skills. Given that counterfactual logic is a key component of reasoning, some of these baselines do occasionally test this skill, albeit in an indirect and non-rigorous way, pointing to a gap in the literature of a test for making valid and formally-grounded counterfactual edits to a text.

Counterfactual story rewriting. The closest related benchmark is *TimeTravel* [51] which tests the ability to rewrite a simple story based on a counterfactual edit. This and related benchmarks like *ART* [9] and *PASTA* [21] are built on top of the seminal *ROCstories* dataset [41], which contains a corpus of 5-sentence stories that are restricted to follow a simple chain-like causal structure, often with a single event/actor, making interventions easy to assess. Further, these works rely on crowdsourced answers instead of expert-evaluation. Finally, and crucially, *TimeTravel* evaluates responses by computing BLEU score w.r.t a single correct answer for each counterfactual edit. As later work notes, evaluation metrics used in *TimeTravel* only correspond weakly with counterfactual validity [37]. Our benchmark's evaluation criteria are designed to permit multiple valid answers, making `CoBe` a seamless addition to in-context learning and post-training pipelines for improving model reasoning.

Counterfactual editing (other domains). A related line of work tests some aspects of causal reasoning in LLMs in Pearl's structural AIP sense, albeit in domains other than narrative rewriting. Notably, *CLadder* [31] and *CounterBench* [12] test symbolic causal reasoning capabilities using diverse graphs and formal causal queries, translated into natural language. *Executable Counterfactuals* [58] tests this capability in the domain of code and math problems. Such works currently sidestep natural language stories as evaluation in prose is challenging. This is the gap addressed by `CoBe`.

Table 2: Dataset statistics.

Statistic	Value
No. of scenarios	2377
No. of domains	7
Avg. words/scenario	38
Avg. sentences/scenario	2.3
Avg. clauses/scenario	5.33
Avg. clauses/sentence	2.5

Table 3: Domain types.

Domain	%
Health and Medicine	6.6%
Engineering	12.2%
History and Geography	6.9%
Economics and Finance	9.0%
Arts and Entertainment	16.8%
Science and Technology	18.4%
Human Activities and Behavior	30.0%

Table 4: Graph types.

Graph	%
Chain-like	50.6%
Collider-like	23.9%
Correlated	4.8%
Diamond-like	3.8%
Fork-like	5.7%
Hydrid	11.1%

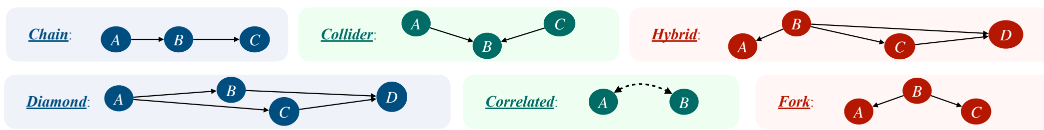

Figure 3: The types of underlying graphs that determine the interaction of variables in a story.

3 `CoBe`: A Dataset for Reasoning Counterfactually about Text

Our aim is to curate a high quality dataset, containing textual scenarios that meet the following criteria: (a) each scenario describes a set of events using conversational language; (b) the scenarios span a diverse range of domains, some of which may require factual knowledge to understand; (c) the scenarios involve a variety of causal dependency templates among events; and (d) each story is paired with a *counterfactual edit* request, phrased in conversational language, and accompanied by formally-grounded evaluation criteria that check the counterfactual validity of a response (instead of checking for any particular answer). In this section, we discuss dataset construction, and provide statistics of key dimensions.

3.1 Dataset Construction

As a first step, the authors manually crafted a core set of text examples. This core set was designed to capture a wide variety of scenarios, with no restrictions on sentence structure or storyline, except that the scenario needed to describe a set of events that have a clear causal relation among them (e.g., it should be self-evident that event A in the scenario is/is not caused by event B). This core set was subsequently filtered to remove scenarios where causal relations were ambiguous. Each scenario in the core set contains the following attributes:[3]

- `Variation text`: following the notation in Sec. 1, this is the text $T = t$, describing the original scenario prior to counterfactual edit.
- `Domain`: a metadata tag used to group scenarios into domains, summarized in Table 3.
- `Query`: this field contains three ways of asking a counterfactual *"what if?"* question, described in more detail below.
- `Representative answer`: a sample of a valid response to the query, for reference (multiple valid answers are possible).
- `Evaluation criteria`: the criteria we use to judge whether a response is a counterfactually valid edit to the text, described in more detail in Sec. 3.2.

Query. Following the notation in Sec. 1, given a textual scenario $T = t$ containing some event(s) $A = a'$, a counterfactual edit query asks what would the text read if we were to hypothetically intervene in the original scenario to fix $A = a$. I.e., it asks for a sample drawn from $P(T_{A=a} \mid A = a')$. We phrase this request in three ways per scenario, using the following syntax:

- (A) *"Rewrite the original text, based on what would have happened instead had...<intervention>."*
- (B) *"Based on the preceding text, rewrite the scenario to reflect what happens if ...<intervention had occurred>."*

[3] An example scenario in JSON format can be found in App. A.2.

- (C) *"Rewrite the above passage to illustrate what would have occurred had ...<intervention>."*

These phrasings were crafted to elicit the natural interpretation of a counterfactual edit, following standard subjunctive *"would have"* grammar [28, 59] (see also [2, 36, 55] for a seminal discussion of why the subjunctive mood intuitively evokes counterfactual reasoning). By averaging scores for responses to all three queries per scenario, we reduce any ambiguity that the question is intended to be interpreted counterfactually, in the sense of the 3-step definition discussed in Sec. 1. We also verify human agreement with our ground truth labels in App. C.3.

Augmentation. Each example in the core set was subsequently augmented to generate variations of the same theme across different domains using LLM assistance to increase diversity of sentence structures, clause reordering, and domains. Each augmented set was manually verified by the authors for coherence and unambiguity. In total, we curated a set of 2377 high-quality textual scenarios, with key statistics shown in Table 2.

Graph variety. Unlike previous counterfactual text editing datasets, `CoBe` ensures variety across different causal structures that connect events in each scenario, as listed in Fig. 3 and Tab. 4.

Notation. Each scenario in the dataset is numbered as `XXXvYY`, where `XXX` corresponds to the original core set example ID that was used as an inspiration to generate this variation (`v`), and `YY` represents the variation ID for that particular sample among all the other variations generated for the `XXX`'th core set ID, indexed starting from 1, for example `307v8`.

3.2 Evaluation Criteria and Error Types

For each scenario in the dataset, the `Evaluation criteria` are crafted to check whether an LLM's response to the query is a valid counterfactual edit. This is a concrete implement of the 3 steps of the counterfactual text editing task formally defined in Sec. 1. We discuss even more fine-grained failure modes in Sec. B.1.

Each string in `Evaluation criteria` checks for one of five error types in the LLM response. Any of these five errors renders the response an invalid counterfactual text edit, as it violates the 3 steps involved in the task definition (Sec. 1). Examples of each error type are linked in parantheses:

1. **Wrongly editing events that are not downstream**: The first string in `Evaluation criteria` lists all the events in the original scenario that should not change in the LLM response, since they are not causally downstream of the intervention (e.g., 116v1).
2. **Not editing downstream variables**: The second string lists in `Evaluation criteria` all the events in the original scenario that ought to change in the LLM response, since they would be causally affected by the intervention (e.g., 325v12).
3. **Wrong connectors**: This is a system-level check, whether the response fails to appropriately update connector words like *because of, since, but, fortunately* etc. For instance, not changing *due to* to *despite* when the counterfactual edit implies a flip (e.g., 114v4).
4. **Quantitative inconsistency**: In certain scenarios, the intervention should logically result in edits where a particular quantity is going up or down. This is a special case of error type 2, where the direction of a causal effect should align with the counterfactual edit (e.g., 315v17).
5. **Latent factors inconsistency**: In certain scenarios, there are unobserved latent variables implied by the text, which need to be held fixed across the counterfactual edit. They are never downstream of the intervention, so should not be changed (e.g., 126v7).

4 Evaluation

4.1 Model Suite

We benchmark three recent flagship models, *GPT-5.4-Pro* [44], *Claude-Opus-4.7* [3], *Gemini-3.1-Pro* [22]. We also include several open-source models to get a better understanding of how much of a role scale plays in developing causal reasoning abilities. We test *Llama3-8B* [16], *Mistral-7B* [29],

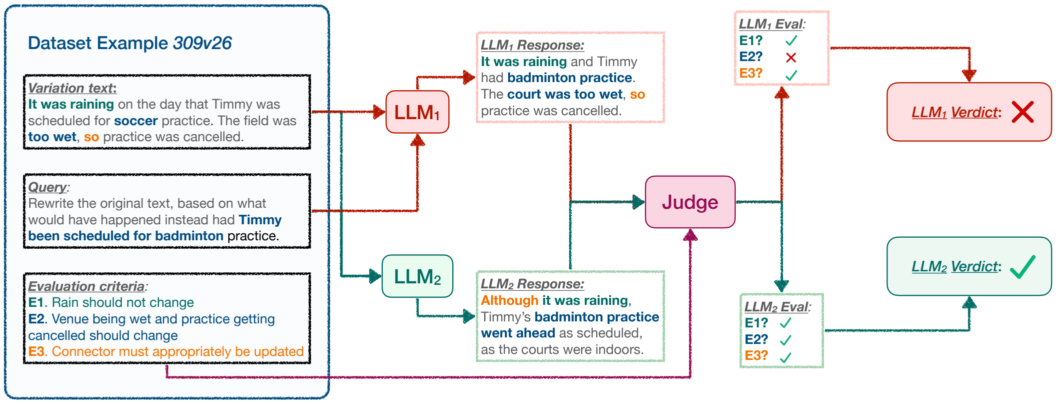

Figure 4: An end-to-end pipeline for evaluating two LLMs on an example from the `CoBe` dataset.

Table 5: Performance of models averaged across three queries.

Metric	GPT	Gemini	Claude	Llama3	Mistral	Qwen	Phi	Gemma
Acc. (%)	55.70$\pm$1.64	53.71 $\pm$2.02	52.42$\pm$2.52	33.55$\pm$1.60	34.08$\pm$4.23	36.32$\pm$2.62	43.92$\pm$1.83	48.48$\pm$2.32

Qwen-7B [5], *Phi-4* [1], and *Gemma-27B* [20]. Overall, the total number of models benchmarked is eight, incorporating various scale, training paradigms, and modeling techniques. For brevity, we refer to models by name only hereafter.

4.2 Pipeline

We first generate responses from the models for each combination of `Variation text` and `Query`, as depicted in Fig. 4. Then, we evaluate the counterfactual text based on five different criteria in `Evaluation criteria`. Specifically, we use Claude-Opus-4.7 to determine whether the response aligns with each evaluation criterion. Finally, we determine the correctness of the response by labeling it as correct if it aligns with all evaluation criteria, and wrong otherwise.

We note that despite the proliferation of complex mathematical metrics, there is a deep vulnerability in prior approaches to evaluating counterfactual texts to assess open-ended generation over an expansive space. Pattern matching or semantic similarity metrics like *BLEU* [45], *ROUGE* [39], and *MoverScore* [65] among others rarely correlate strongly with human judgments of validity [60]. Naive human evaluators are highly sensitive to pragmatic and semantic flaws, personal biases and superstitions, logical leaps, and subtle narrative contradictions. The task of evaluating counterfactuals is challenging due to the inherent nature of task allowing for a distribution of answers to be correct as opposed to a single ground truth. As discussed in Sec. 3.2, our benchmark pipeline offers a principled way of assessing the correctness of the counterfactual texts.

Human agreement. We conduct human agreement experiment to assess the reliability of our evaluation pipeline, with three independent expert human raters. Our LLM-Judge shows a high agreement with human raters, providing sufficient statistical confidence to scale the evaluator across the broader benchmark. See App. C.3 for details.

4.3 Results

Overall results (Tab. 5). We investigate the benchmark performance of the models averaged across three different query phrasings described in Sec. 3.1. We observe that all models fail to show strong performance, where the highest accuracy is only 55.7%, highlighting the brittle nature of the counterfactual reasoning capabilities of modern LLMs.

An intriguing observation is that the performance of Gemma-27B is close to flagship models (e.g., 4% gap with Claude), despite it has significantly smaller number of parameters. This implies that the gains on scaling the weights, the training cost and the model development cycle by several magnitudes might not bring improvements on counterfactual reasoning at the same magnification. Our benchmark

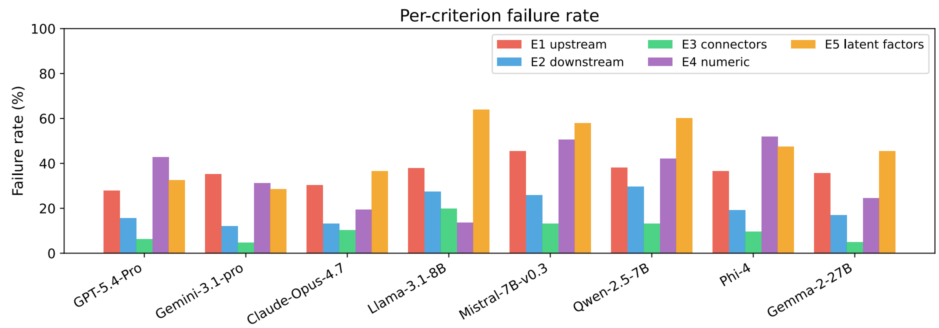

Figure 5: Failure rates across different evaluation criteria.

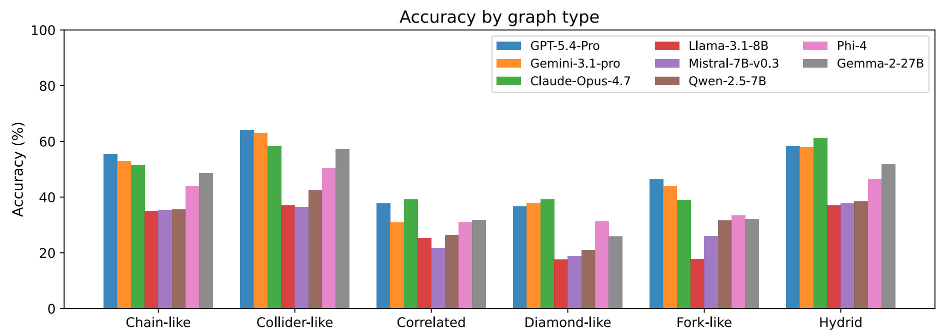

Figure 6: Accuracy across different graph types.

suggests that there need to be an explicit effort in developing systems that are natively designed to reason better, and innovations beyond scaling are required to cross the next threshold of performance.

Performance across different evaluation criteria (Fig. 5). To analyze different types of errors in model responses, we examine the failure rates for five evaluation criteria described in Sec. 3.2. The general trend of models being prone to making incorrect edits to the non-descendant variables (E1 in Fig. 5), perhaps due to mistaking correlational relationships it has observed in the training data as cause-and-effect relationships, seems consistent across the board. Comparatively, the errors due to not editing the downstream variables (E2) or using the wrong connectors (E3) are generally lower.

Among the frontier models, GPT does the worst on interventions when numerical changes are expected (E4), compared to Gemini and Claude, latter of which does remarkably well, as does Llama. On investigating the errors due to not keeping latent factors invariant across the factual and counterfactual world, (E5) we observe the superiority of large-scale frontier models compared to the open source models. We hypothesize this is due to larger models having a better understanding of unobserved factors in the given context, as opposed to the models with the fewer parameters.

Performance across different graph types (Fig. 6). The differences in accuracy of different models across varied graph types show that the models are fallible to misinterpret correlated graphs as cause-and-effect graphs, illustrating the behavior of LLMs that primarily rely on associational knowledge. Furthermore, the performance is also noticeably lower for Diamond-like and Fork-like graphs, indicating that models struggle with consistently propagating effects of interventions across different downstream branches.

Performance across different query phrasings (Fig. 7) and domains (Fig. 8). We investigate how different query phrasings and different domains impact the performance of models. By looking at three paraphrased versions of the same query, and comparing results per query-type illustrates that there is negligible difference in performance across different query phrasings. This implies that the high failure rates of models are likely due to their inherent lack of counterfactual reasoning capabilities, rather than misinterpreting the query. Similarly, we observe that the differences of the model performances across varied domains are insignificant. This corroborates that the failure of the models in counterfactual text editing does not primarily stem from lack of specific domain knowledge, but rather innate inability to generate appropriate counterfactual edits across the board.

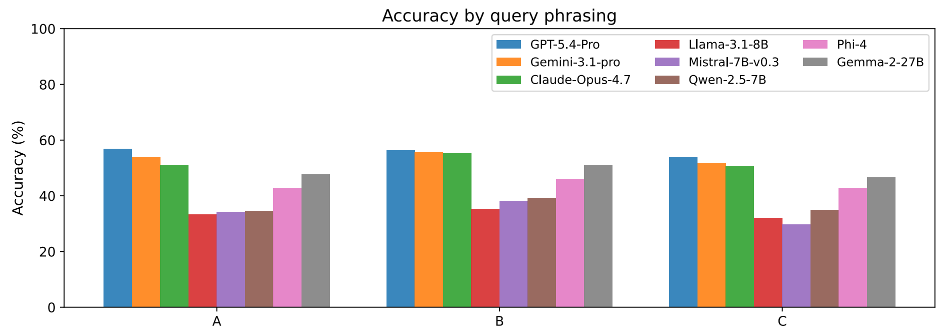

Figure 7: Accuracy across different query phrasings.

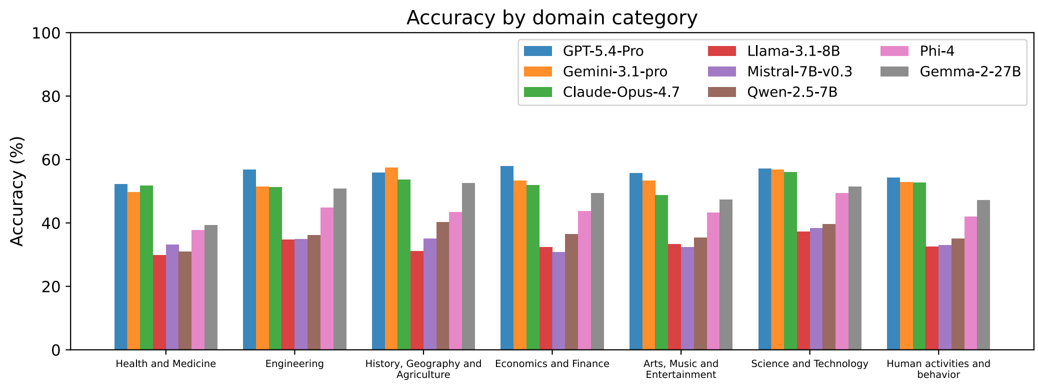

Figure 8: Accuracy across different domains.

Model biases. Among the flagship models, Claude seems to be performing the best on numerical changes, but struggles on maintaining the latent state constant for the factual and counterfactual worlds. Furthermore, while Gemini is exceptional at maintaining coherent grammatical structure, it suffers from making slightly more upstream changes. GPT seems to be most error prone when dealing with scenarios involving updates in numerical quantities but does better than average on refraining from changing the non-descendant variables, which suggests it might actually be less reliant on purely observational data. Different characteristics of models when sliced across different axes reveal insights that would have been hard to gain otherwise from a black box setting. This is one of the examples of how `CoBe` could be better used for understanding model behavior.

5 Limitations and Future Work

The task of counterfactual text editing is one specific example of counterfactual reasoning, and solving our benchmark should not be read as the sign of achieving general counterfactual reasoning capabilities. Yet, we believe our benchmark would serve as a stepping stone towards building models capable of counterfactual reasoning and opens the door for fruitful future research. Another limitation is that the form of manual creation of samples could introduce human biases, where we aimed to address such biases by ensuring the diversity of dataset in terms of domains, query phrasing, etc. A natural future direction is extending the benchmark to different domains, e.g., agentic environments.

6 Conclusion

We introduce `CoBe`, a principled and causally-grounded approach to benchmarking the causal reasoning capabilities of LLMs via the task of counterfactual text editing. We construct a diverse and expansive dataset, covering scenarios from varied domains, with different underlying graph structures and inducing novel failure modes. The dataset is structured in an easy-to-use format that should aid researchers in investigating these blind-spots and developing better systems. We hope `CoBe` is a valuable asset to the research community, and prompts the development of more causally-aware models.

Acknowledgments and Disclosure of Funding

This research is supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

[1] Marah Abdin et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[2] Ernest W. Adams. Subjunctive and indicative conditionals. *Foundations of Language*, 6(1): 89–94, 1970. doi: 10.2307/2272204.

[3] Anthropic. Claude opus 4.7 model release, 2026. URL https://www.anthropic.com. Accessed: 2026-04-28.

[4] C Avin, Ilya Shpitser, and Judea Pearl. Identifiability of Path-Specific Effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence {IJCAI-05}*, pages 357–363, Edinburgh, UK, 2005. Morgan-Kaufmann Publishers.

[5] Jinze Bai et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[6] Elias Bareinboim. *Causal Artificial Intelligence: A Roadmap for Building Causally Intelligent Systems*. 2025. URL https://causalai-book.net/.

[7] Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2015.

[8] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl's Hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 507–556. Association for Computing Machinery, New York, NY, USA, 1st edition, 2022.

[9] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations*, 2020.

[10] Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. Zero-shot LLM-guided Counterfactual Generation: A Case Study on NLP Model Evaluation . In *2024 IEEE International Conference on Big Data (BigData)*, pages 1243–1248, Los Alamitos, CA, USA, December 2024. IEEE Computer Society. doi: 10.1109/BigData62323.2024.10825537. URL https://doi.ieeecomputersociety.org/10.1109/BigData62323.2024.10825537.

[11] Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen Mckeown. Do models explain themselves? Counterfactual simulatability of natural language explanations. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7880–7904. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24bl.html.

[12] Yuefei Chen, Vivek K Singh, Jing Ma, and Ruixiang Tang. Counterbench: Evaluating and improving counterfactual reasoning in large language models. *arXiv preprint arXiv:2502.11008*, 2025.

[13] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[14] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

[15] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding. *Commun. ACM*, 67(1):110–120, December 2023. ISSN 0001-0782. doi: 10.1145/3596490. URL https://doi.org/10.1145/3596490.

[16] Abhimanyu Dubey et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[17] Jörg Frohberg and Frank Binder. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.229/.

[18] Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. Faithful explanations of black-box nlp models using llm-generated counterfactuals, 2023. URL https://arxiv.org/abs/2310.00603.

[19] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780, 2020. URL https://arxiv.org/abs/2004.07780.

[20] Gemma Team et al. Gemma 2: Improving open models with competitive results. *arXiv preprint arXiv:2408.00118*, 2024.

[21] Sayontan Ghosh, Mahnaz Koupaee, Isabella Chen, Francis Ferraro, Nathanael Chambers, and Niranjan Balasubramanian. Pasta: A dataset for modeling participant states in narratives. *Transactions of the Association for Computational Linguistics*, 11:1283–1300, 11 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00600. URL https://doi.org/10.1162/tacl_a_00600.

[22] Google DeepMind. Gemini 3.1 pro - model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-04-28.

[23] A Gopnik, C N Glymour, D M Sobel, L E Schulz, T Kushnir, and D Danks. A theory of causal learning in children: {C}ausal maps and {B}ayes nets. *Psychological Review*, 111(1):3–32, 2004.

[24] Joseph Y Halpern and Judea Pearl. Actual Causality. Technical Report R-266, University of California Los Angeles, Cognitive Systems Lab, Los Angeles, 1999.

[25] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv*, 2020. doi: 10.48550/arxiv.2009.03300.

[26] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[27] D Hume. *An Enquiry Concerning Human Understanding*. Open Court Press, LaSalle, 1748.

[28] Michela Ippolito. *Subjunctive Conditionals: A Linguistic Analysis*. 09 2013. ISBN 9780262019484. doi: 10.7551/mitpress/9780262019484.001.0001.

[29] Albert Q Jiang et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[30] Kyohoon Jin, Juhwan Choi, JungMin Yun, Junho Lee, Soojin Jang, and YoungBin Kim. CoBA: Counterbias text augmentation for mitigating various spurious correlations via semantic triples. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10260–10278, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.520. URL https://aclanthology.org/2025.emnlp-main.520/.

[31] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: assessing causal reasoning in language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[32] Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. Llms are prone to fallacies in causal inference, 2024. URL https://arxiv.org/abs/2406.12158.

[33] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on machine learning research*, 2024. URL https://par.nsf.gov/biblio/10574854.

[34] Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Hae Won Park, Samir Tulebaev, and Cynthia Breazeal. Medical hallucination in foundation models and their impact on healthcare. *medRxiv*, 2025. doi: 10.1101/2025.02.28.25323115. URL https://www.medrxiv.org/content/early/2025/03/03/2025.02.28.25323115.

[35] Kai Zhan Lee, Drago Plecko, and Elias Bareinboim. Causal explanations through counterfactual variable attributions. Technical Report R-135, Columbia Causal AI Laboratory, May 2025. URL https://causalai.net/r135.pdf. Columbia CausalAI Laboratory, Technical Report (R-135).

[36] D Lewis. *Counterfactuals*. Harvard University Press, Cambridge, MA, 1973.

[37] Dandan Li, Ziyu Guo, Qing Liu, Li Jin, Zequn Zhang, Kaiwen Wei, and Feng Li. Click: Integrating causal inference and commonsense knowledge incorporation for counterfactual story generation. *Electronics*, 12(19), 2023. ISSN 2079-9292. doi: 10.3390/electronics12194173. URL https://www.mdpi.com/2079-9292/12/19/4173.

[38] Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. Prompting large language models for counterfactual generation: An empirical study. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13201–13221, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1156/.

[39] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

[40] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

[41] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N16-1098.

[42] S Chandra Mouli, Yangze Zhou, and Bruno Ribeiro. Bias challenges in counterfactual data augmentation, 2022. URL https://arxiv.org/abs/2209.05104.

[43] Scott Mueller and Judea Pearl. Personalized decision making – a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050, 2023.

[44] OpenAI. Gpt-5.4 release information, 2026. URL https://openai.com. Accessed: 2026-04-28.

[45] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

[46] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2000. ISBN 978-0-521-89560-6.

[47] Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. Morgan Kaufmann, San Francisco, CA, 2001.

[48] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018. ISBN 978-0-465-09760-9.

[49] Drago Plecko and Elias Bareinboim. Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends in Machine Learning*, 17(3):304–589, Jan 2024.

[50] Drago Plecko, Patrik Okanovic, Shreyas Havaldar, Torsten Hoefler, and Elias Bareinboim. Epidemiology of large language models: A benchmark for observational distribution knowledge, 2025. URL https://arxiv.org/abs/2511.03070.

[51] Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[52] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

[53] Donald B Rubin. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31:161–170, 2004.

[54] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, and Amanda Dsouza. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research, May/2022, https://openreview.net/forum?id=uyTL5Bvosj*, 2022. doi: 10.48550/arxiv.2206.04615.

[55] R C Stalnaker. A Theory of Conditionals. In N. Rescher, editor, *Studies in Logical Theory*, volume No.\ 2, Am. Blackwell, Oxford, 1968.

[56] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

[57] Ewoenam Kwaku Tokpo and Toon Calders. Fairflow: An automated approach to model-based counterfactual data augmentation for nlp. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part VII*, page 160–176, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-70367-6. doi: 10.1007/978-3-031-70368-3_10. URL https://doi.org/10.1007/978-3-031-70368-3_10.

[58] Aniket Vashishtha, Qirun Dai, Hongyuan Mei, Amit Sharma, Chenhao Tan, and Hao Peng. Executable counterfactuals: Improving LLMs' causal reasoning through code. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Lm46gJA0q8.

[59] Kai von Fintel. Subjunctive conditionals. In Gillian Russell and Delia Graff Fara, editors, *The Routledge Companion to Philosophy of Language*, pages 466–477. Routledge, New York, 2012.

[60] Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. A survey on natural language counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4798–4818, 2024.

[61] Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. Causality for large language models, 2024. URL https://arxiv.org/abs/2410.15319.

[62] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 6707–6723, 2021.

[63] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[64] Junzhe Zhang and Elias Bareinboim. Fairness in Decision-Making-The Causal Explanation Formula. In *AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1016/j.energy.2007.09.003.

[65] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1053. URL https://aclanthology.org/D19-1053/.

Appendix

A Reference

In this section, we first provide several examples of LLM responses, along with the evaluation criteria used for the scenario, representative answer, and judgment reasoning (App. A.1). Then, we provide a sample JSON format in our dataset (App. A.2).

A.1 Examples

From App. A.1.7 to App. A.1.9, we highlight some of the examples that were evaluated as invalid counterfactual edits. In App. A.1.10, we show an example that was evaluated as valid counterfactual edit.

A.1.1 Sample Failure 1

Example 311v26

Text: Alonso is bad at backgammon, but he rolled good dice, so he won the game.
Query: *Rewrite the original text, based on what would have happened instead had Alonso been good at backgammon and lost the game.*
Claude Opus 4.7 Output: *Alonso is good at backgammon, but he rolled bad dice, so he lost the game.*
Sample Representative Answer: *Alonso is good at backgammon, and he rolled good dice, but he lost the game.*
Evaluation Criteria: *Should NOT change: He rolled good dice.*
Should change: Alonso is bad at backgammon. He won the game.
Judgment Reasoning: *Being good at a game and being lucky are independent variables, Alonso could lose the game due to unseen external factors like a really brilliant opponent.*

A.1.2 Sample Failure 2

Example 325v12

Text: Sending a green wedge of citrus skittering across the floor, the bartender bumped the cutting board while mixing a classic Margarita.
Query: *Rewrite the original text, based on what would have happened instead had he been mixing a classic Old Fashioned.*
Phi-4 Output: *Sending a splash of ice clinking against the glass, the bartender gently tapped the mixing glass while preparing a classic Old Fashioned. He carefully measured the whiskey, added a dash of bitters, and muddled a sugar cube with a twist of orange peel, ensuring each ingredient blended seamlessly into the timeless cocktail.*
Sample Representative Answer: *Sending an orange wedge of citrus skittering across the floor, the bartender bumped the cutting board while mixing a classic Old Fashioned.*
Evaluation Criteria: *Should NOT change: The bartender bumped the cutting board.*
Should change: Sending a green wedge of citrus skittering across the floor.
Judgment Reasoning: *The downstream variable should change to an orange wedge as an equivalent but different ingredient used in Old Fashioned over the green wedge used in Margarita due to the intervention.*

A.1.3 Sample Failure 3

Example 153v1

Text: The crowd cheered wildly as the gymnast nailed the dismount. She held the strenuous iron cross on the rings for 4 seconds and wore a sparkling red leotard.
Query: *Rewrite the original text, based on what would have happened instead had the gymnast held the strenuous iron cross on the rings for 3 minutes.*
Gemini 3.1 Pro Output: *What started as standard cheering dissolved into a stunned, breathless silence as the gymnast defied the limits of human physiology. She held the strenuous iron cross on the rings for an agonizing, impossible three minutes, her muscles trembling violently beneath her sparkling red leotard as the judges stared in absolute disbelief. When she finally broke the hold and miraculously still managed to nail the dismount, the arena didn't just cheer—they erupted into sheer pandemonium, completely bewildered by the superhuman feat they had just witnessed.*
Sample Representative Answer: *The crowd gasped in horror as the gymnast tore a muscle and collapsed. She held the strenuous iron cross on the rings for 3 minutes and wore a sparkling red leotard.*
Evaluation Criteria: *Should NOT change: The gymnast wore a sparkling red leotard.*
Should change: The crowd cheered wildly as the gymnast nailed the dismount.
Judgment Reasoning: *The world record for the longest duration for iron cross on the rings is around 40 seconds, so 3 minutes is beyond possible for any human, thus nailing the dismount is an impossible outcome given the physical limits of the human body.*

A.1.4 Sample Failure 4

Example 126v7

Text: Walking east across the Brooklyn Bridge at 8 AM, Sato faced into the morning sun, with the Manhattan skyline at her back.
Query: *Rewrite the original text, based on what would have happened instead had Sato been walking east across the Brooklyn Bridge at 6 PM instead of 8 AM.*
Llama3-8B Output: *Walking east across the Brooklyn Bridge at 6 PM, Sato faced into the fading light of the setting sun, with the Manhattan skyline at her back.*
Sample Representative Answer: *Walking east across the Brooklyn Bridge at 6 PM, Sato had the late-afternoon sun behind her, with the Manhattan skyline at her back.*
Evaluation Criteria: *Should NOT change: Walking east across the Brooklyn Bridge, Sato had the Manhattan skyline at her back.*
Should change: It was 8 AM. The morning sun was in front of Sato.
Judgment Reasoning: *The evening sun should be in the west, which is opposite the direction it was in the morning, so it should be behind Sato's back and not in front.*

A.1.5 Sample Failure 5

Example 114v4

Text: Due to the record-breaking heatwave, the server room temperatures spiked. The website went offline for several hours because the automated cooling systems failed to handle the thermal load.
Query: *Rewrite the original text, based on what would have happened instead had the server room temperatures not spiked.*
Qwen-7B Output: *Due to the record-breaking heatwave, the server room temperatures remained stable thanks to the reliable performance of the automated cooling systems. As a result, the website continued to operate without any interruptions or downtime.*
Sample Representative Answer: *Despite the record-breaking heatwave, the server room temperatures did not spike. The website stayed online because the cooling systems were not overtaxed.*
Evaluation Criteria: *Should NOT change: There was a record-breaking heatwave.*
Should change: The server room temperatures spiked. The website went offline for several hours.
Judgment Reasoning: *The correct conjunction to use in this context is something like despite, nevertheless etc. and not due to or because as is the case in the output.*

A.1.6 Sample Failure 6

Example 403v20

I went to the temple for a family ceremony in the morning. Someone mentioned that parking nearby filled up early. I left the house before sunrise.

Query

Edit the text so it reflects abundant parking all morning.

GPT 5.4 Pro Output

I went to the temple for a family ceremony in the morning. Someone mentioned that there was plenty of parking nearby all morning, so I left the house after sunrise.

Sample Representative Answer

I went to the temple for a family ceremony in the morning. Someone mentioned that parking nearby stayed open all morning. I left the house before sunrise.

Evaluation Criteria

Should NOT change: I went to the temple for a family ceremony in the morning. I left the house before sunrise.
Should change: Someone mentioned that parking nearby filled up early.

Judgment Reasoning

As the ceremony is in the morning, irrespective of the availability of the parking, they have to leave before sunrise.

A.1.7 Sample Failure 7

Example 116v1

Because the carbon content was exceptionally high, a brittle microstructure formed within the steel. Consequently, the alloy fractured during the high-pressure forging process.

Query

Rewrite the original text, based on what would have happened instead had the alloy had not fractured.

Claude Opus 4.7 Output

Because the carbon content was within optimal limits, a ductile microstructure formed within the steel. Consequently, the alloy withstood the high-pressure forging process.

Sample Representative Answer

Because the carbon content was exceptionally high, a brittle microstructure formed within the steel. Despite this, the alloy did not fracture during the high-pressure forging process.

Evaluation Criteria

Should NOT change: The carbon content was exceptionally high. A brittle microstructure formed within the steel.
Should change: The alloy fractured during the forging process.

Judgment Reasoning

The carbon content is an upstream variable which should not be affected by an intervention to the the fracturing.

A.1.8 Sample Failure 8

Example 307v8

Faulty wiring led to distorted audio and combined with uncomfortable seating, caused a terrible audience experience.

Query

Rewrite the original text, based on what would have happened instead had there been an excellent audience experience.

Claude Opus 4.7 Output

Proper wiring delivered crystal-clear audio and, combined with comfortable seating, created a wonderful audience experience.

Sample Representative Answer

Despite the faulty wiring that led to distorted audio and the uncomfortable seating, there was an excellent audience experience.

Evaluation Criteria

Should NOT change: Faulty wiring led to distorted audio and there was uncomfortable seating.
Should change: There was a terrible audience experience.

Judgment Reasoning

The wiring and seating are upstream variables which should not be affected by an intervention to the experience.

A.1.9 Sample Failure 9

Example 315v14

The largest illegal piracy site in the world was suddenly shut down by federal authorities. Meanwhile, the studio elected for an exclusive theatrical release window, which forced all eager audiences to buy physical tickets. Opening weekend box office revenues skyrocketed to an incredible $150 million.

Query

Rewrite the original text, based on what would have happened instead had the studio had elected for a day-and-date streaming release.

Mistral-7B Output

In an unprecedented move, the largest illegal streaming site in the world was unexpectedly taken down by federal authorities. Simultaneously, the studio opted for a day-and-date streaming release, making the highly anticipated film available for digital purchase or rental. As a result, audiences worldwide were able to enjoy the movie from the comfort of their homes, leading to a significant surge in digital sales and rentals. The opening weekend's digital revenue reached an impressive $100 million, demonstrating the growing demand for flexible and convenient viewing options.

Sample Representative Answer

The largest illegal piracy site in the world was suddenly shut down by federal authorities. Meanwhile, the studio elected for a day-and-date streaming release, which allowed audiences to easily watch from home instead of buying physical tickets. Opening weekend box office revenues limped to a disappointing $35 million.

Evaluation Criteria

Should NOT change: The largest illegal piracy site in the world was suddenly shut down by federal authorities.
Should change: Meanwhile, the studio elected for an exclusive theatrical release window, which forced all eager audiences to buy physical tickets. Opening weekend box office revenues skyrocketed to an incredible $150 million.
Numerical change expected: Opening weekend box office revenues were lower than $150 million.

Judgement Reasoning

Instead of talking about the decreased revenue from the box office, the response talks about the digital revenue and how it was impressive. The expectation in this counterfactual world is that the digital release should hamper the box office revenue which is not reflected appropriately in the response.

A.1.10 Sample Success

Example 217v10

Using a paper bookmark to keep her place, Lisa reads a thick novel at the terminal. Her connecting flight has been delayed by three hours.

Query

Rewrite the original text, based on what would have happened instead had Lisa watched a movie on her laptop.

Gemini 3.1 Pro Output: Evaluated as *correct*, despite multiple answers

Here are a few ways to rewrite the text, depending on what specific detail you want to focus on:

Option 1 (Directly mirrors the structure of the original): Using a pair of headphones to block out the noise, Lisa watches a movie on her laptop at the terminal. Her connecting flight has been delayed by three hours.

Option 2 (Mirrors the "keeping her place" aspect of the bookmark): Hitting the spacebar to pause her screen whenever announcements chime, Lisa watches a movie on her laptop at the terminal. Her connecting flight has been delayed by three hours.

Option 3 (Focuses on the reality of using electronics at an airport): Sitting close to a wall outlet to keep her battery charged, Lisa watches a movie on her laptop at the terminal. Her connecting flight has been delayed by three hours.

Evaluation Criteria

Should NOT change: Her connecting flight has been delayed by three hours.
Should change: Using a paper bookmark to keep her place, Lisa reads a thick novel at the terminal.

Judgment Reasoning

All of the options correctly incorporate the fact that Lisa shouldn't be using a bookmark as she is not reading a novel anymore, and correctly reflects the effects of the intervention. Furthermore, the independent variable of her flight being delayed is correctly maintained consistently by all of the outcomes, thus no non-descendant variables are affected by the intervention. The filler text before the actual responses is also ignored in the evaluation criteria.

A.2 Sample JSON

Sample JSON Entry


```
"Core Set ID": 307,
"Variation ID": 8,
"Domain": "Music",
"Variation text": "Faulty wiring led to distorted audio and combined
                   with uncomfortable seating, caused a terrible
                   audience experience.",
"Query": [
    "Rewrite the original text, based on what would have happened
    instead had there been an excellent audience experience.",
    "Based on the preceding text, rewrite the scenario to reflect what
    happens if there had been an excellent audience experience.",
    "Rewrite the above passage to illustrate what would have occurred
    had there been an excellent audience experience."
],
"Representative answer": "Faulty wiring led to distorted audio and
                          combined with uncomfortable seating, still
                          the audience experience was excellent.",
"Evaluation criteria": [
    "Faulty wiring led to distorted audio and
    there was uncomfortable seating.",
    "There was a terrible audience experience."
]
```

B Dataset Details

In this section, we discuss failure modes that represents semantic categorization of failure cases (App. B.1).

B.1 Failure Modes

We expand on the error modes discussed in Sec. 3.2, to take a closer look at the causes of failures in the generation of valid counterfactual edits. The expanded failure modes are useful to determine the characteristics of model performance across these different axes, that encode unique blind-spots in the counterfactual reasoning abilities of models. These failure modes are manually-crafted and the failures mapped according to the semantically closest failure mode.

1. **Mistaking causal connection for logical dependency**: Variables with a causal connection are incorrectly deemed to have a collapsed relationship such that a particular value determines the value of the other completely (e.g., 116v1).
2. **Mistaking correlation for causal connection**: Variables that are independent being incorrectly encoded as having a cause-and-effect relationship, or scenarios where the cause and effect are flipped (e.g., 311v26).
3. **Ignoring causal connections**: Not propagating the effect of the interventions appropriately to downstream variables (e.g., 325v12).
4. **Not understanding physics/chemistry/biology**: The counterfactual story does not adhere to the scientific laws, which need to be implicitly encoded to when considering the effects of the intervention (e.g., 153v1).
5. **Not inferring/fixing latent conditions**: The unobserved variable values inferred from the environment are not held consistent in the counterfactual world (e.g., 126v7).
6. **Wrong causal implication**: The grammatical components of the sentence, particularly the connectors like *so, and, but, therefore* which preserve or negate the flow of logic from one clause to the other are not appropriate for the counterfactual sentence (e.g., 114v4).
7. **Not relying on the appropriate variables as a causal reason**: An independent variable, say C, or just a single parent, say D is used to determine the value of a variable B, when only its parent A should (also) have been used (e.g., 403v20).

C Evaluation Details

In this section, we first describe the infrastructural details on the evaluation in App. C.1, and then provide the prompts used for assessing each each evaluation criteria in App. C.2. We also provide the details of human agreement in App. C.3, and then discuss additional results in App. C.4.

C.1 Infrastructure

For the open-source models the responses were generated using one or two NVIDIA H100. For the frontier models, we generated the responses using API keys. Wherever *thinking-mode* or *reasoning-effort* or an equivalent parameter is available we activate it and set it to *medium* or the *default* applicable.

C.2 LLM-Judge prompts

Because the LLM-Judge evaluation pipeline is based on finding invalid counterfactual edits as opposed to matching to a ground truth sample, our evaluation is robust to varied responses generated by the LLMs and focuses on the relevant content. The LLM-Judge is robust to models outputting unnecessary additional text, outputting multiple possible scenarios and generating long complicated responses. See example 217v10 for reference. This is one of the strengths of our setup, where LLMs aren't wrongly penalized for straying from the exact presentation of the response expected, but only penalized when the content does not align with the expectation.

The prompts provided to the LLM-Judge for evaluating the responses are provided below:

Q0: Wrong connectors

```
Criterion: "Every causal connector in the rewrite accurately
reflects the causal relationship described."

Mentally list each causal connector (so, thus, therefore, but,
however, consequently, as a result, etc.). T only if every
connector signals the correct causal direction; F if any connector
implies a wrong or reversed relationship. If there are no causal
connectors, reply T.

Your entire reply must be exactly one character: T if the criterion
passes for that question type, F if it fails.

Rewrite:
$response_text
```

Q1: Wrongly editing non-descendant variables

```
Short examples (your entire output is exactly one letter):

Facts that should be UNCHANGED - criterion says pipe
leaking and floor wet; rewrite keeps leak but floor is dry → F

Criterion (facts that should be UNCHANGED):
"$eq1"

For each claim: T only if still present and correct in the
rewrite, F if missing or changed.

Your entire reply must be exactly one character: T if the criterion
passes for that question type, F if it fails.

Rewrite:
$response_text
```

Q2: Not editing downstream variables

```
Short examples (your entire output is exactly one letter):

Effects that should be CHANGED - criterion says brakes failed
and car hit wall; rewrite has brakes engaged and car safe → T

Criterion (effects that should be CHANGED):
"$eq2"

For each claim: T only if changed or removed in the rewrite, F
if it still holds as in the original.

Your entire reply must be exactly one character: T if the criterion
passes for that question type, F if it fails.

Rewrite:
$response_text
```

C.3 Human Agreement

C.3.1 Setup

To validate the use of an LLM as a scalable evaluator for our dataset, we first established a robust human baseline. We randomly sampled 100 instances from the counterfactuals generated by one of the models in the model suit for a particular variation and query. This is exactly the same information that we provide to our LLM judge. We then assigned three independent expert human annotators to rate each instance as *Yes* or *No* based on whether they thought the LLM Response to the counterfactual query was in accordance with the rubric. The inter-rater reliability among the human annotators yielded a Fleiss' Kappa of 0.5, indicating moderate agreement. This reflects the inherent ambiguity and difficulty of the evaluation task. A definitive ground truth for these 100 instances was then established via a strict majority vote. There was unanimous agreement on 65% of samples, highlighting that certain counterfactuals are easy to judge, but a significant number of them still are ambiguous.

The human rating was also done in a few-shot manner to appropriately mimic the LLM-as-a-judge pipeline. The prompt and rubric provided to the human raters is as follows.

Rubric

```
Hello!

Thanks for agreeing to be one of the human raters for helping us
understand the counterfactual reasoning capabilities of modern large
language models. The goal of this short (~ 15-20 mins) activity is to
get human opinions on outputs produced by querying LLMs with certain
interesting counterfactual queries.

You would be provided with 25 short conversations with an LLM, and
you are to judge whether the LLM output is correct. In each
conversation you will be provided with the following:

Original Context (T = t): a paragraph describing a scenario with some
events and outcomes.

Query (T_x = t' | T = t, X = x'): A natural language question that
asks how the original story (T) would have changed if one of the
original events/conditions (X = x') was instead fixed by external
intervention to be different (X = x).

LLM Output (T = t''): the response from a model to the original
context and query.

For each conversation, your task is to judge whether the LLM Output
(T=t'') is a valid answer to the Query (T\_x = t' | T = t, X = x').
In other words, if the original story is altered by external
intervention, would the LLM Output be a valid counterfactual
scenario?

Note: while the Query has a fixed technical interpretation, there
could be multiple valid answers to this Query. You are to evaluate
whether the LLM Output is reasonable given the original context.
```

Rubric

```
Please mark "Yes", if you think the LLM Output is
(a) factually consistent with the Original Context; (b) incorporates
the  external intervention mentioned in the Query; (c) appropriately
reflects downstream changes following from the intervention,
without affecting causal non-descendants; and (d) does not change any
causal implications from the Original Context, such as by using
wrong connectors (e.g. ‘‘due to X’’ instead of ‘‘despite X’’).

Please mark "No", if you think the LLM Output is (a) factually
inconsistent with the Original Context; or (b) does not appropriately
incorporate the external intervention from the Query, or reflect this
edit in causally downstream events; or (c) changes events that are
not causal descendants of the  intervention; or (d) changes causal
implications from the Original Context, such as by using wrong
connectors (e.g. ‘‘due to X’’ instead of ‘‘despite X’’); or (e)
hallucinates or produces incoherent output.

Please use the "Other" option sparingly and add comments and
justification, whenever a clear Yes/No answer is not provided.

EXAMPLE 1

Original Context: Due to the storm, the school declared a snow
day. Because they were not in class during the day, the kids were
able to go outside and go sledding while the sun was up.

Query: Rewrite the original text, based on what would have happened
instead had they not been granted the snow day.

LLM Output: Despite the storm, the school did not declare a snow day.
Because they were in class during the day, the kids were unable to go
outside and go sledding while the sun was up.

Human Evaluation: YES - the LLM Output is correct. It incorporates
the intervention and downstream changes, does not change
non-descendants,  and correctly edits ‘‘due to’’ to ‘‘despite’’.

EXAMPLE 2

Original Context: Due to the clear skies, the hawk was able to see
and catch the mouse from a distance. As a result, the wolf  was
unable to find something to eat that night.

Query: Rewrite the original text, based on what would have happened
instead had the wolf been able to find food.

LLM Output: Due to the cloudy skies, the hawk was unable to see and
catch the mouse from a distance. As a result, the wolf was able to
find something to eat that night.

Human Evaluation: NO - the LLM Output is wrong. It changes
prior events when incorporating the intervention.
```

C.3.2 LLM Judge Alignment and Statistical Significance

We then processed the same 100 sampled instances through our LLM-as-a-judge pipeline using a few-shot prompt.

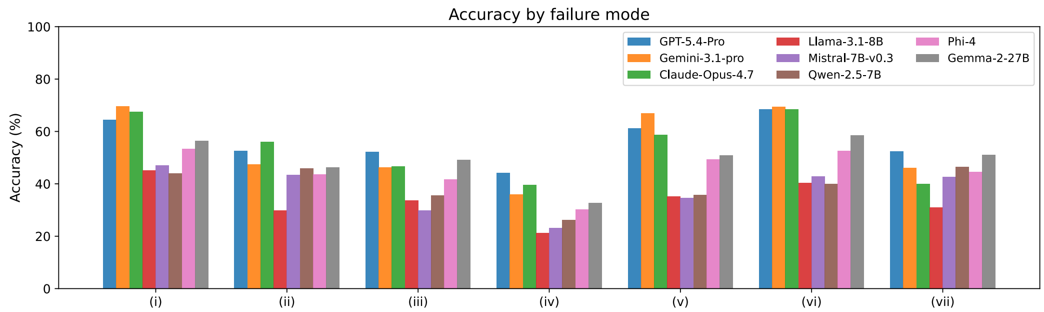

Figure 9: Accuracy across different failure modes.

The LLM achieved an *Accuracy* of 77%, *Precision* of 95.8%, *Recall* of 68.7% and *F1 Score* of 0.8 against the human majority. More importantly, to account for random chance in a binary classification task, we calculated Cohen's κ between the LLM and the human majority, resulting in $\kappa = 0.55$. This demonstrates that the LLM achieves a level of inter-rater reliability with the consensus that outperforms the human-to-human reliability (0.55 vs 0.50). Given the sample size ($N = 100$), this agreement provides sufficient statistical confidence to scale the evaluator across the broader benchmark.

Comparing the LLM to each individual human provides insight into whose *style* of grading the LLM is most closely mimicking. LLM vs. Rater 1: 0.419 (Moderate Agreement) and LLM vs. Rater 3: 0.393 (Fair Agreement) are lower than LLM vs. Rater 2: 0.566 (Moderate Agreement), which further highlights the subjective nature of evaluation in this task. Despite the subjectivity, there is moderate agreement which re-inforces the reliability of the system.

C.3.3 Inter-Human Agreement

Breaking down the agreement between individual raters highlights the variance in human judgment for this specific task. Rater 1 vs. Rater 2: 0.509 (Moderate Agreement) and Rater 1 vs. Rater 3: 0.419 (Fair Agreement) agreements are lower than Rater 2 vs. Rater 3: 0.637 (Substantial Agreement). This highlights that Rater 2 and Rater 3 are highly aligned with each other, but Rater 1 is a minor outlier. Rater 1's disagreement is pulling down the overall Fleiss' Kappa.

C.3.4 Qualitative analysis

A detailed analysis of the confusion matrix reveals a highly asymmetric error profile. Of the 23 total disagreements with the human majority, 21 were False Negatives (the LLM predicted *Wrong* while humans predicted *Correct*), and only 2 were False Positives. This indicates that the LLM judge operates with a highly conservative, strict evaluation. Further qualitative investigation revealed that the disagreement was on samples which required nuanced subject knowledge (eg. chemical reactions with and without a catalyst) where accurate judgment is unreasonable to expect from a human across all domains but reasonable to expect of an LLM due to its knowledge base. The other source of disagreement were counterfactual generations with grammatical errors (eg. placement of connectors before the appropriate clauses) or ambiguity in meaning where human raters tended to mark the generation as correct presumably due to lack of knowledge of strict grammatical rules and constructs. An LLM-Judge is expected, by design, to be aware of all grammatical rules and harsh on marking ambiguous edits as wrong.

C.4 Additional results

Performance across different failure modes (Fig. 9) Investigating the performance of various models across the fine-grained failure modes defined in Sec. B.1 reveals interesting patterns. LLMs struggle in maintaining invariant the scientific laws established either explicitly or implicitly in the scenario while making counterfactual edits.

As expected the LLMs tend to mistake correlational relationships for cause-and-effect relationships and showcase a significant failure rate in scenarios where there is a correlational relationship with the potential to be misinterpreted.

Likewise, there are several instances of failures in scenarios where the effect of the intervention has to be propagated to the downstream variables but the LLMs fail to do so appropriately, highlighting that they still struggle from effectively updating the scenario to map to a valid counterfactual world resulting from the intervention. This observation also corroborates the higher than average failure rate for the Diamond- and Fork-like graphs as noted in Fig. 6.

D LLM Usage

LLMs were used in the project for assistance in code writing and debugging, for paper writing tasks like phrasing. All LLM generated content was reviewed and validated by the authors. The authors take full responsibility for all contents of the paper.