CoBe: A Benchmark For Conversational Counterfactual Text Editing

Anonymous authors
Paper under double-blind review

Abstract

Counterfactual text editing is the ability to revise a given text in light of a hypothetical change while preserving facts not causally affected by that change. When asked what would have happened if a doctor had not recommended surgery, the model output should retain the patient's medical history and modify only the consequences of the decision. This requires causal understanding: a model must determine, without being explicitly instructed to, what events are affected by the hypothetical change and what should remain fixed. This skill is central to human reasoning, especially in safety-critical domains like healthcare or law. By contrast, the task of *associational text generation* merely generates text statistically associated with the altered condition. In the medical example, an associational response might make the patient's symptoms less severe because milder cases correlate with no surgery. Such responses produce a plausible scenario, but not the counterfactual version of the given one. Prior counterfactual-text datasets do not clearly separate these modes of reasoning. We introduce the **Co**unterfactual **Be**nchmark (`CoBe`), a testbed for models' zero-shot ability to counterfactually edit a given textual scenario when prompted to do so in a *conversationally natural* way. `CoBe` contains 2000 questions spanning healthcare, science and engineering, business, and everyday scenarios. We test a suite of frontier and smaller models, with frontier systems achieving an average score of 58.05%. Our findings indicate that models frequently default to associational generation rather than preserving the causal structure of the original scenario. Moreover, performance does not scale consistently with either model size or output token length, suggesting that the failure may reflect model training objectives rather than insufficient model capacity.[1]

1 Introduction

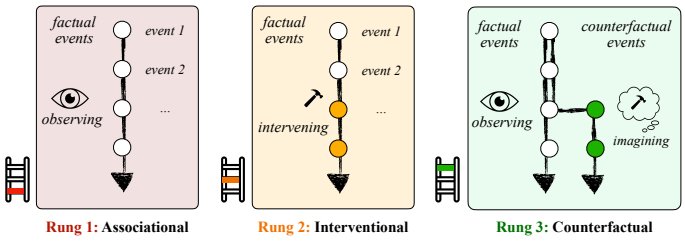

Figure 1: The three rungs of Pearl's causal ladder are (i) associational; (ii) interventional; and (iii) counterfactual.

Conversational counterfactuals are a vital tool for understanding the world and the consequences of choosing different actions in a given scenario. Consider a classic example in the philosophy of language: the statements (1) *If Oswald did not shoot Kennedy, someone else did*; and (2) *Had Oswald not shot Kennedy, someone else would have*. Most readers would agree with the first statement while disagreeing with the second, intuitively grasping that the two make fundamentally different claims. The first is a factual claim, whereas the second is a *counterfactual* claim about what *would have happened* had we intervened in the given scenario with a different condition.

Counterfactual reasoning belongs to the third rung of Judea Pearl's celebrated *ladder of causation*, a.k.a the Pearl Causal Hierarchy, or PCH (Pearl & Mackenzie, 2018)(Bareinboim, 2025, Ch. 5). Rung 1 involves observing statistical *associations* among events in the factual world. Rung 2 involves *intervening* in the factual world and observing outcomes. Rung 3 involves observing factual events and reasoning about what would *counterfactually* have happened had we intervened in the given

[1]We provide an anonymized link to sample data representing approximately 10% of our dataset.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

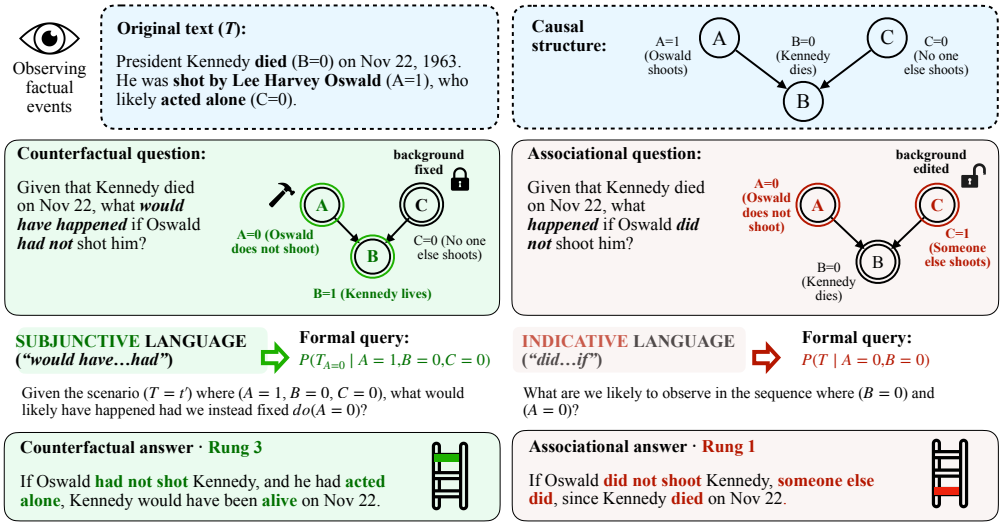

Figure 2: Counterfactual text editing, belonging to Rung 3 of Pearl's *causal ladder*, keeps background factors constant, and reasons about what would have happened as the result of a hypothetical intervention. By contrast, associational text generation belongs to Rung 1, and generates alternative scenarios correlated with the altered condition (possibly changing background factors). **Note**: counterfactual scenarios need not be realistic or likely to occur, ordinarily.

scenario. Notably, a counterfactual only alters events causally downstream of the hypothetical intervention, and keeps fixed background factors in the original scenario that are causally prior to the intervention (Fig. 1). Refer to App. A.1 for a primer on causal models. Counterfactual reasoning is well-established to be central to personalized decision-making (Bareinboim et al., 2015; Mueller & Pearl, 2023; Raghavan & Bareinboim, 2025), mediation analysis (Avin et al., 2005; Pearl, 2001; Rubin, 2004), fairness analysis (Plecko & Bareinboim, 2024; Zhang & Bareinboim, 2018), and explainability (Halpern & Pearl, 1999; Lee et al., 2025).

In this work we focus on *conversational counterfactuals*: Rung 3 questions articulated in everyday colloquial language, using the subjunctive "would have...had" grammatical construction. Would the patient have succumbed to an infection had she not been moved to the crowded ward? Would oil prices be lower today if the UAE had quit OPEC earlier? Conversational counterfactuals are integral to how humans think, communicate, and construct explanations for why events occurred; they are foundational to our intuitions about ethics, regret, credit, and responsibility (Bareinboim, 2025; Gopnik et al., 2004; Hume, 1748; Lewis, 1973; Pearl, 2000; Pearl & Mackenzie, 2018). We next formally define this important capability that we wish to test on language models.

Conversational counterfactual text editing. A language model receives two inputs: (i) a textual story T describing a set of events $(A, B...)$ where some events causally affect other events in the story; and (ii) a conversational counterfactual question involving a hypothetical intervention in this particular storyline, phrased in colloquial subjunctive "would have...had" grammar. From this prompt, the model is expected to infer and execute three steps: *(a) keep fixed any background factors/events in the story that are not causally affected by the intervened-variable in the original timeline; (b) enact the requested intervention by imagining an externally enforced edit to the storyline; and (c) reason about the commonsense downstream effects on subsequent events in the storyline due to enforcing this change, and produce edited text reflecting this counterfactual storyline.*

Example shown in Fig. 2 in green. We also provide a detailed illustration, and additional formal background on conversational counterfactuals and the subjunctive conditional construction in App. A.2, using this Oswald/Kennedy example. Note: in our definition of conversational counterfactuals, we intentionally avoid any technical instructions about causal graphs, counterfactuals, or step-by-step procedures in our prompts. Our motivation for focusing on conversational prompting is twofold. First, this is a natural human capability: when asked a question like *"what would have happened in this scenario if I had done something different?"* we automatically understand that this means the intervention should only affect downstream events holding fixed all background factors and prior events, without requiring causal formalisms or detailed instructions such as "keep prior events

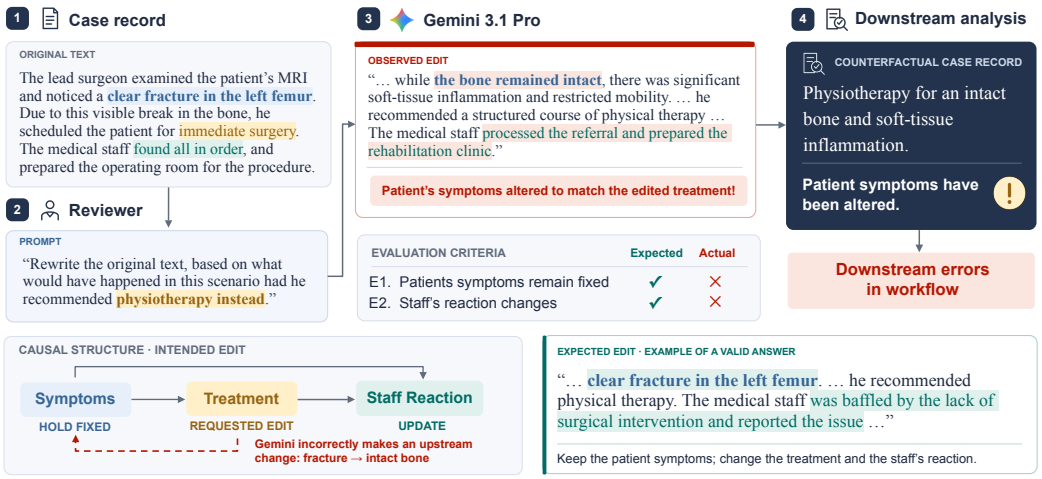

Figure 3: Idealized healthcare workflow depicting a staff reviewing medical records to generate case studies for training. Gemini 3.1 is prompted, in conversational language, to incorporate a counterfactual edit to patient treatment. Gemini incorrectly alters patient symptoms, causing downstream errors.

fixed". Second, AI assistants are increasingly embedded in everyday and professional workflows (Anthropic, 2026b), with personal agents such as Meta's Muse explicitly designed for communication as users would with another person (Meta, 2026). In these settings, users should be able to express a hypothetical change in ordinary interactive language without needing to detail in each case which variables should remain invariant, or requiring expertise in task-specific prompt tuning (Zamfirescu-Pereira et al., 2023). Our task is therefore designed to test whether models can *infer* these edit requirements from the scenario and conversational request, without requiring the user to supply a detailed causal task recipe. App. E examines more explicit prompt instructions to distinguish performance in the conversational setting from performance with additional procedural guidance.

Unfortunately, a growing body of work documents that LLMs often rely on *associational* signals rather than genuine causal structure, learning surface heuristics from their pretraining distribution (Joshi et al., 2024; Wu et al., 2024) and often learning shortcuts to predict "right but for the wrong reasons" (Du et al., 2023; Geirhos et al., 2020; McCoy et al., 2019). In contrast to counterfactual edits, the task of **associational generation** (belonging to Rung 1 of the causal ladder) *generates text by merely conditioning on an alternative event to ask what scenarios are statistically likely to co-occur with the given change*. We illustrate this in Fig. 2 in red, and expand in App. A.2.

Given how embedded language models are in modern workflows, confusing counterfactual and association generation can lead to errors in safety-critical domains like medicine, law etc. We show in Fig. 3 an idealized healthcare workflow where a user is interacting with an AI assistant to generate case studies for the same patient under different treatments. Frontier models incorrectly alter the patient history in the output, introducing downstream errors in the workflow. Such shortcut learning based on spurious correlations in the training data risks systematic, hard-to-predict errors (Kiciman et al., 2024; Kim et al., 2025). For instance, when trying to mitigate bias using counterfactual data augmentation, naive approaches can introduce new biases due to spurious correlations in the data (Jin et al., 2025; Mouli et al., 2022; Tokpo & Calders, 2024).

With this motivation, we introduce **Co**unterfactual **Be**nchmark (`CoBe`): a dataset for testing the native ability of language models to produce causally-valid counterfactual edits to text, when prompted to do so in *conversational* language, without specialized prompts. The key highlights are as follows:

- We curate a set of 2000 **samples** (200 public), each containing, in natural language, (i) a short text describing a scenario, and (ii) a set of single-turn prompts requesting, in colloquial everyday language, a **counterfactual edit** to the paragraph.
- We provide the **evaluation criteria** for each question to determine the validity of model responses, offering a principled and causally-grounded way of evaluating counterfactual edits, verified by human experts.
- Our dataset covers scenarios from **diverse domains**: Science & Research, Engineering, Business & Services, Everyday Scenarios, Healthcare, and Legal, Policy & HR, among others.

- We investigate a comprehensive suite of closed and open-source models, with results summarized in Sec. 4. Notably, **frontier models only score an avg. of 58.05%** as of July 2026.
- We taxonomize the failure modes we observed and characterize the idiosyncratic error profiles of different models. For instance, we observe that Gemini-3.1-Pro outputs appear to strongly favor correlational consistency, and GPT-5.4-Pro struggles to incorporate background information from the scenario in its edits.

To our knowledge, we provide the first testbed for zero-shot conversational counterfactual edits, grounded in formal causal semantics, in the domain of narrative stories. Our evaluation criteria allow for multiple valid responses that respect counterfactual reasoning. We are also the first to characterize the distinct error-profiles of different models on this important task. All of these allow `CoBe` to be seamlessly integrated as a training signal into pipelines designed to improve model reasoning, both fine-tuning and inference-time.

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 (Bareinboim, 2025, Ch. 1). `CoBe` focuses on a very specific sub-task: we test whether models can infer which facts to preserve or revise from the narrative text and a conversational edit request, without explicit instructions specifying causal dependencies, invariant facts, or a counterfactual reasoning procedure..

2 Related Work

Table 1: Comparison to prior benchmarks.

Prior work	Ctf-editing 3 steps	Narrative scenarios	No causal structure	Open-ended generation	Expert/formal verification
TimeTravel	✓	✓	✗	✓	✗
CRASS	✗	✓	✓	✗	✗
CLadder	✓	✗	✗	✗	✓
CounterBench	✓	✗	✗	✗	✓
CoBe (our work)	✓	✓	✓	✓	✓

We group related work under common themes for discussion. In Table 1, we compare notable datasets to `CoBe` across five dimensions: does the benchmark (1) ground itself in the formal 3-step definition of counterfactual text editing (see Sec. 1); (2) test understanding of narrative stories (vs. mathematics or coding questions phrased in text); (3) withhold information about the story's causal structure (i.e., not providing a causal graph, nor pre-specifying which sentences need to remain invariant); (4) evaluate open-ended generation (vs. multiple choice selection); and (5) include evaluation criteria verified by experts or formal solvers (vs. crowd-sourced).

General benchmarks. It is now standard for model cards to describe performance on general benchmarks like MMLU Hendrycks et al. (2020), BigBench Srivastava et al. (2022), GSM8K Cobbe et al. (2021), ARC-AGI-2 Chollet et al. (2025) etc., which test a combination of domain knowledge, and abstract and common-sense reasoning skills. Some of these baselines do occasionally test counterfactual reasoning skills, albeit in an indirect and non-rigorous way, pointing to a gap in the literature of a test for making formally-grounded and open-ended counterfactual edits to a text.

Counterfactual story rewriting. *TimeTravel* Qin et al. (2019) and *PASTA* (Ghosh et al., 2023) test the ability to rewrite a simple story based on a counterfactual edit. Both are built on the dated *ROCstories* dataset (Mostafazadeh et al., 2016): a restrictive corpus of short everyday stories that follow a simple linear timeline, often with a single event/actor, where story structure makes it clear that only sentences following an edit should be changed. Both benchmarks use crowd-sourced judgments and similarity metrics like BLEU for evaluation, which later work notes only correspond weakly with counterfactual validity (Li et al., 2023). `CoBe`'s evaluation criteria are expert-verified, allowing multiple valid rewrites without requiring similarity to a reference answer. Mu & Li (2024) propose a method for counterfactual story rewrites, but they also assume the task pre-specifies which parts of the story need to remain invariant.

Other causal reasoning tasks. *CLadder* (Jin et al., 2023) and *CounterBench* (Chen et al., 2025) evaluate causal queries over explicitly specified causal structures expressed in natural language. *Corr2Cause* (Jin et al., 2024) tests whether causal hypotheses follow from supplied statistical dependence and independence relations. Other benchmarks assess causal understanding in text: *CRASS* (Frohberg & Binder, 2022) tests multiple-choice questions about natural-language counterfactuals; *CRAB* (Romanou et al., 2023) assesses the strength of causal relationships between events in news narratives; *CausalT5k* (Geng et al., 2026) diagnoses causal reasoning failures in natural-language

Table 2: Dataset statistics.

Statistic	Value
No. of scenarios	2000
No. of domains	7
Avg. words/scenario	39
Avg. sentences/scenario	2.35
Avg. clauses/scenario	5.43
Avg. clauses/sentence	2.44

Table 3: Domain types.

Career Domain	%
Science & Research	16.0
Engineering	15.7
Business & Services	14.5
Everyday	14.2
Healthcare	6.0
Legal, Policy & HR	4.7
Others	28.9

Table 4: Graph types.

Graph	%
Chain-like	51.0%
Collider-like	21.9%
Correlated	5.7%
Diamond-like	4.5%
Fork-like	4.8%
Hybrid	11.9%

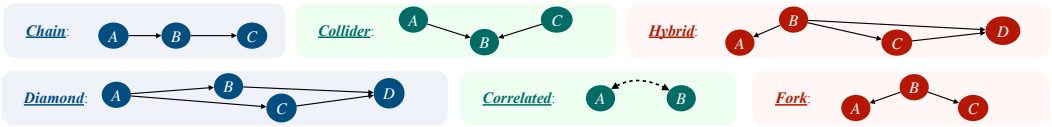

Figure 4: The types of underlying graphs that determine the interaction of variables in a story.

scenarios, including counterfactual validity and appropriate refusal under insufficient evidence. These tasks complement CoBe, which evaluates narrative revision from conversational counterfactual requests, requiring models to infer which facts to preserve and which consequences to update without a supplied causal graph or explicit causal editing instructions.

3 CoBe: A Dataset for Counterfactual Text Editing

In this section, we discuss dataset construction, the dataset schema, key statistics, and our evaluation methodology for model responses to each sample.

Dataset construction. We constructed CoBe in four stages (App. D.3). First, student researchers who had completed a graduate causal inference course contributed 425 candidate examples. Second, the authors reviewed and refined promising submissions into 106 core examples spanning different domains, causal structures, and potential reasoning failures. Third, LLM-assisted augmentation produced a pool of 25 variations per core example, with the goal of adding domain and linguistic variety while preserving the intended causal relationships. Fourth, author review and external expert feedback informed the final filtering, removing 650 of the 2,650 augmented examples and yielding 2,000 retained samples.

Filtering and quality assurance. At stages two and four, each example was reviewed by at least two authors for story coherence, clarity of the intervention, and the soundness of its preservation and update criteria. Each example required both authors to support inclusion in the dataset, with disagreements resolved through discussion. Following initial author filtering, three causal inference experts uninvolved in construction independently assessed a set of 100 random samples for coherence, grammar, and validity of the counterfactual edit. Their feedback identified a systematic inconsistency and prompted further exclusions. App. D.3 details the review procedures and audit results.

Dataset schema. Each sample in the dataset contains the following fields. We provide an instantiated sample in App. B.2.

- `Story`: a natural-language text containing a story describing a series of events.
- `Domain`: a metadata tag used to group scenarios into domains, summarized in Table 3.
- `Query`: a counterfactual edit request, in conversational language, to change some detail in the story and produce the counterfactual text; described in more detail below.
- `Illustrative answer (not scored)`: an author-crafted example of a valid response to the query, for reference (multiple valid answers are possible); this is not scored.
- `Evaluation criteria`: the criteria used to judge whether a model response is a valid counterfactual edit to the text, described in more detail in Sec. 3.1.

Query. Following the task definition in Sec. 1, a counterfactual edit query asks what would the edited text read if we were to hypothetically intervene in the original story described in the text to enforce a change, and allow the change to affect subsequent events in the story. 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 would have happened if ...<intervention had occurred>."*
- (C) *"Rewrite the above passage to illustrate what would have occurred had ...<intervention>."*

These phrasings were crafted to elicit the interpretation of a counterfactual edit, following standard subjunctive "would have" grammar. We provide detailed formal background on how such counterfactual statements are to be interpreted, and the role of subjunctive conditions in App. A.2. By averaging scores for responses to all three queries per scenario, we reduce any prompt-related ambiguity that the question is intended to be interpreted counterfactually, in the sense of the 3-step definition in Sec. 1. We also verify human agreement with our prompt interpretation and evaluation criteria in App. D.4.

We discussed extensively in Sec. 1 our motivation for keeping these prompts conversational, and avoiding any causal terminology or explicit task instructions. We provide prompt ablations studies in App. E, where we test models using increasingly more detailed prompts that provide procedural guidance for this task, while making a clear scope separation from our task setting.

Graph variety. Unlike previous counterfactual text editing datasets, `CoBe` ensures variety across different causal structures connecting events in each story, as listed in Fig. 4 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`.

3.1 Evaluation Methodology

For each story in the dataset, the `Evaluation criteria` are crafted to check whether a model's response to the query is a valid counterfactual edit under the benchmark's intended commonsense interpretation. These criteria are grounded in the three steps of the counterfactual text editing task formally defined in Sec. 1.

The first two strings in `Evaluation criteria` specify story-specific constraints. Together with a separate system-level check of connectors, these identify three error types in the model response. Each error represents a violation of the intended counterfactual edit. Examples of each error type are linked in parentheses:

1. **Wrongly editing unaffected facts**: The first string in `Evaluation criteria` lists facts in the original scenario that should remain unchanged in the model response. These include events that are neither directly intervened on nor causally downstream of the intervention, as well as facts that remain unchanged despite the intervention in the given scenario (e.g., 403v1).
2. **Failing to make required changes**: The second string in `Evaluation criteria` lists original facts that should change to reflect the requested intervention and its consequences. These changes are specified for the particular scenario; being causally downstream does not, by itself, require a variable's value to change (e.g., 309v26).
3. **Wrong connectors**: A separate system-level check assesses whether connector words such as *because of, since, so, but* appropriately reflect the relationships in the rewritten scenario. For instance, retaining *due to* can incorrectly present an unchanged danger as the reason for a newly stipulated decision, where *despite* would express the intended contrast (e.g., 115v3). This check doesn't require any particular replacement words, but rather concerns the relationship expressed.

The `Illustrative answer (not scored)` provides one valid rewrite but is not used for reference matching. Multiple responses may satisfy the evaluation criteria. We discuss more fine-grained failure modes in Sec. C.

Pipeline. We first generate responses for each combination of `Story` and `Query`, as depicted in Fig. 5. We then use Claude-Opus-4.7 as an LLM judge to evaluate each response using three checks. The first two assess the supplied `Evaluation criteria`: facts that should remain true in the rewrite (E1), and original facts that should change (E2). A separate check assesses causal connector consistency (E0). Each check returns `T` if its requirements are satisfied and `F` otherwise; a response is marked correct only if all three checks pass. Thus, `T` denotes satisfaction of a check, rather than the truth of every original statement.

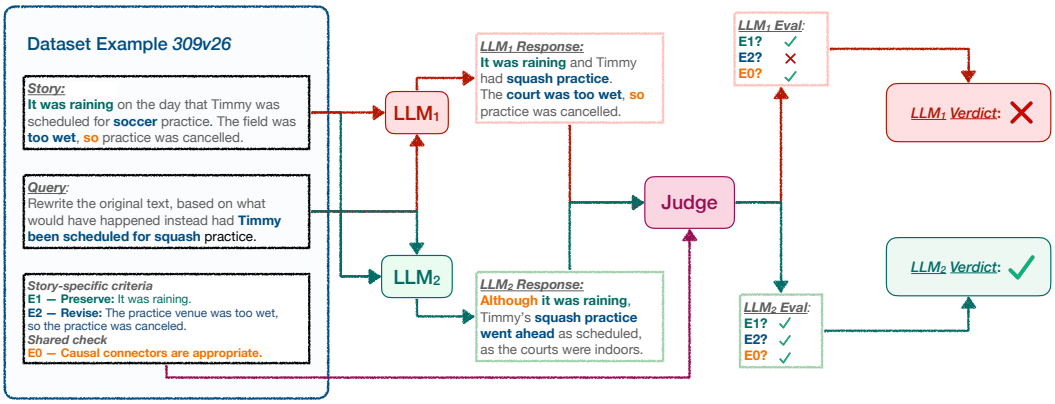

Figure 5: The two respondent LLMs illustrate a failing and a passing rewrite, evaluated independently by the same judge; a response passes only when all three checks pass.

Table 5: Performance of models averaged across three queries.

Metric	GPT	Gemini	Claude	Llama-8B	Mistral	Qwen	Phi	Gemma	Llama-70B
Acc. (%)	61.34±0.45	56.68±0.44	56.14±2.50	33.77±2.50	32.90±3.36	36.83±3.41	44.73±2.02	48.26±1.45	49.64±1.42

Evaluation rationale. Counterfactual editing admits multiple valid rewrites, making agreement with a reference answer an incomplete measure of correctness. Reference-based metrics such as *BLEU* (Papineni et al., 2002), *ROUGE* (Lin, 2004), and *MoverScore* (Zhao et al., 2019) do not directly test which facts should remain invariant or change under the intervention (a major limitation in prior benchmarks). Our evaluation instead checks these scenario-specific constraints and the consistency of causal language, allowing different wordings and outcomes that satisfy the criteria.

Evaluation limitation. The implementation of E2 accepts the removal of a listed fact as well as its modification. Consequently, an incomplete rewrite can pass this check by omitting a consequence that should have been updated. We did not observe systematic use of this shortcut in the responses we manually inspected. Moreover, the expert evaluation of 250 sampled responses in App. D.4 provides evidence of agreement between the judge and human ground truth assessments. This supports the final results as indeed representative of the model's actual performance on this task. We do not claim our evaluation criteria are robust to deliberate omission or other strategies designed to satisfy the checks without completing the editing task.

Human agreement. We conduct human agreement experiments to assess the reliability of our evaluation pipeline, with three independent expert human raters. Our LLM-Judge shows a high agreement (Cohen's $\kappa = 0.812$) with human raters, demonstrating the soundness of our evaluation criteria, and sufficient statistical confidence to scale the evaluator across the broader benchmark. See App. D.4 for details.

4 Results

Model Suite. We benchmark three recent closed-source models, *GPT-5.4-Pro* (OpenAI, 2026), *Claude-Opus-4.7* (Anthropic, 2026a), *Gemini-3.1-Pro* (Google DeepMind, 2026). We 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 *Llama-3.1-8B, 70B* (Dubey et al., 2024), *Mistral-7B* (Jiang et al., 2023), *Qwen-7B* (Bai et al., 2023), *Phi-4* (Abdin et al., 2024), and *Gemma-27B* (Gemma Team et al., 2024). Overall, the total number of models benchmarked is nine, incorporating various scales, training paradigms, and modeling techniques. For brevity, we refer to models by name only hereafter.

Overall results (Tab. 5). We investigate the benchmark performance of the models averaged across three different query phrasings described in Sec. 3. We observe that all models fail to show strong performance, where the highest accuracy is only 61.34%, highlighting the brittle nature of the conversational counterfactual reasoning capabilities of modern LLMs.

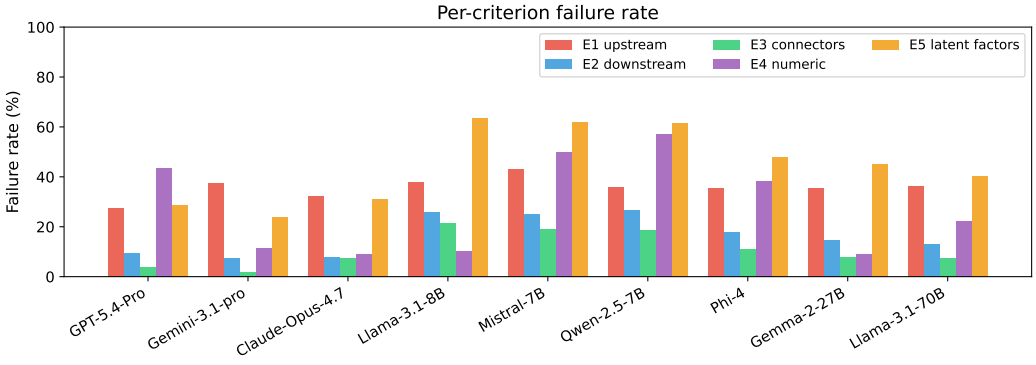

Figure 6: Failure rates across different evaluation criteria. E1-3 are described in Sec. 3.1. E4 and E5 are interesting sub-types where LLMs reason wrongly about the direction of a downstream edit, and fail to correctly infer latent factors implicit in the story.

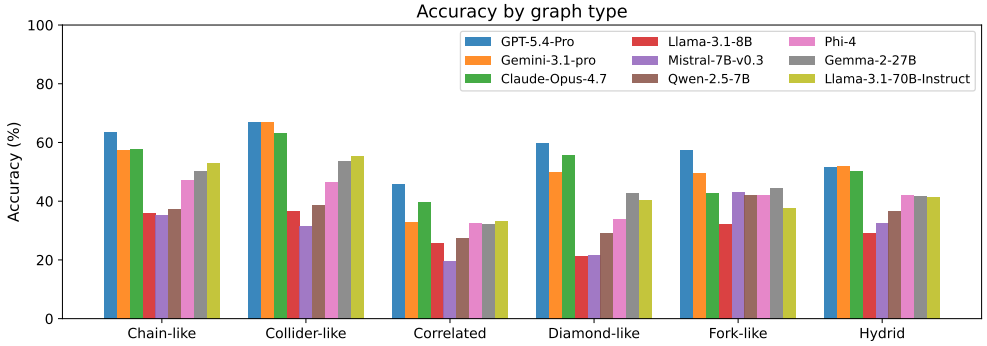

Figure 7: Accuracy across different graph types.

An intriguing observation is that the performance of LLama-70B is close to larger models (e.g., 6.5% gap with Claude), despite it having a 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 to the same extent. This suggests that there needs 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. 6). To analyze different types of errors in model responses, we examine the failure rates per the evaluation criteria described in Sec. 3.1. The general trend of models being prone to making incorrect edits to the non-descendant variables (E1 in Fig. 6), 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, the latter of which does 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 fewer parameters.

Performance across different graph types (Fig. 7). The differences in accuracy of different models across varied graph types show that the models are prone to misinterpreting correlated graphs as cause-and-effect graphs, illustrating the behavior of LLMs that primarily rely on associational knowledge. Furthermore, the performance is also lower for Diamond-like and Fork-like graphs, indicating that models struggle with consistently propagating effects of interventions across different downstream branches.

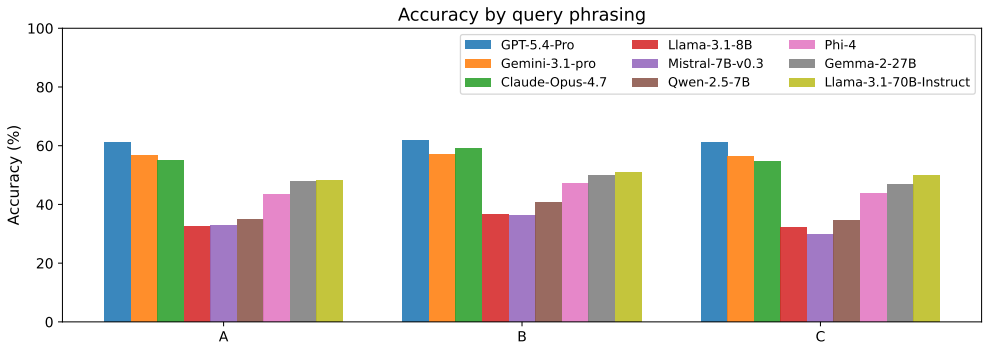

Figure 8: Accuracy across different query phrasings.

Performance across different query phrasings (Fig. 8). We investigate how different query phrasings impact the performance of models. 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 corroborates that the high failure rates of models are likely due to their lack of native, conversational counterfactual reasoning capabilities.

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.

Prompt ablations. We additionally show results from prompting three models with alternative prompt templates designed to provide increasingly more procedural instructions to the model. Results are in App. E, using the same 500 random samples across prompt ablation settings. We find that providing additional conversational prompts that even more strongly suggest the counterfactual interpretation only improve performance by around 5% on average for frontier models, and lower performance for smaller ones. We also show results from adding increasingly more explicit causal instructions to the prompt, while clarifying that this is strictly outside our task scope. We find that adding a simple clarification like *"Keep prior and unaffected facts unchanged, and update the consequences of the requested change."* only improves scores by around 3%. Finally, we show that even a fully detailed step-by-step causal recipe that elaborates this task following structural causal semantics still leaves our benchmark unsaturated, with frontier models scoring 81.3% and 76.1%.

5 Conclusion

We introduce `CoBe`, a principled and causally-grounded approach to benchmarking the ability of language models to perform conversational 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.

Limitations. As clarified in Sec. 1, counterfactual text editing is an important but specific example of counterfactual reasoning; solving our benchmark should not be read as the sign of having achieved general counterfactual reasoning capabilities. A natural future direction is extending the benchmark to different domains, e.g., vision. A further limitation, common to benchmarks, is the possibility of author-bias in the curation of samples, which we address by ensuring the diversity in terms of domains and query phrasing, as validated by independent human quality checks.

AI USE STATEMENT

In this work, we used generative AI tools as part of the dataset generation and evaluation pipeline, as detailed in Section 3 and 4 in the paper. Specifically, LLMs were used to augment an author-curated set of core scenarios by generating variations to increase diversity and LLMs were also used as a part of the benchmark evaluation pipeline to assess whether model-generated counterfactual edits satisfied the evaluation criteria. All dataset examples went through several stages of human review and the reliability of the LLM-evaluation pipeline also underwent human agreement experiments to ensure correctness. We additionally used LLMs for assistance in code writing and debugging, and for paper editing tasks like improving phrasing. All AI-assisted material was reviewed by the authors, and the codebase and experimental results were independently and thoroughly checked by the authors before inclusion. We did not rely on generative AI to draw the insights and conclusions highlighted in our work. We take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

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

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

Anthropic. Making Claude Cowork ready for enterprise, April 2026b. URL https://claude.com/blog/cowork-for-enterprise.

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}*, pp. 357–363, Edinburgh, UK, 2005. Morgan-Kaufmann Publishers.

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

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

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

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*, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 1st edition, 2022.

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.

François Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. ARC-AGI-2: A new challenge for frontier AI reasoning systems, 2025. URL https://arxiv.org/abs/2505.11831.

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.

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.

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

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 (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2126–2140, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.229/.

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.

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

Longling Geng, Andy Ouyang, Theodore Wu, Daphne Barretto, Matthew J. Hayes, Rachael Cooper, Yuqiao Zeng, Sameer Vijay, Gia Ancone, Ankit Rai, Matthew Wolfman, Patrick Flanagan, and Edward Y. Chang. CausalT5k: Diagnosing refusal and failure modes in trustworthy causal reasoning across causal rungs. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*. Association for Computing Machinery, 2026. doi: 10.1145/3770855.3817567. URL https://doi.org/10.1145/3770855.3817567.

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.

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

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.

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

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.

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

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

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

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 (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 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/.

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.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vqIH0ObdqL.

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.

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.

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.

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).

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

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.

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

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 (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Meta. Introducing Muse: The world's first personal AI agent built for everyone, September 2026. URL https://about.fb.com/news/2026/09/introducing-muse-personal-ai-agent/.

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*, pp. 839–849, San Diego, California, June 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N16-1098.

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

Feiteng Mu and Wenjie Li. A causal approach for counterfactual reasoning in narratives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6556–6569. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.354/.

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

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

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, pp. 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

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

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

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

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.

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.

Arvind Raghavan and Elias Bareinboim. Counterfactual realizability. In *The Thirteenth International Conference on Learning Representations*, number R-113, 2025. URL https://arxiv.org/abs/2503.11870.

Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. CRAB: Assessing the strength of causal relationships between real-world events. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15198–15216. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.940. URL https://aclanthology.org/2023.emnlp-main.940/.

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

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.

R C Stalnaker. A Theory of Conditionals. In N. Rescher (ed.), *Studies in Logical Theory*, volume No.\ 2, Am. Blackwell, Oxford, 1968.

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*, pp. 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.

Kai von Fintel. Subjunctive conditionals. In Gillian Russell and Delia Graff Fara (eds.), *The Routledge Companion to Philosophy of Language*, pp. 466–477. Routledge, New York, 2012.

Julius von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 177–196. PMLR, 2023. URL https://proceedings.mlr.press/v213/kugelgen23a.html.

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.

J. D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–21. Association for Computing Machinery, 2023. doi: 10.1145/3544548.3581388. URL https://doi.org/10.1145/3544548.3581388.

Junzhe Zhang and Elias Bareinboim. Fairness in Decision-Making-The Causal Explanation Formula. In *AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1609/aaai.v32i1.11564.

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 (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 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 Background on Counterfactuals **16**

A.1 Structural Causal Models and the Pearl Causal Hierarchy 16

A.2 Subjunctive Conditionals and Conversational Counterfactuals 17

B Dataset Samples **18**

B.1 Examples 18

B.1.1 Sample Failure 1: Spatial consequence 18

B.1.2 Sample Failure 2: Misattributing an unchanged event 18

B.1.3 Sample Failure 3: Retaining an inapplicable consequence 19

B.1.4 Sample Failure 4: Unsupported causal connector 19

B.1.5 Sample Failure 5: Overlooking an unchanged purpose 19

B.1.6 Sample Failure 6: Altering a historical fact 20

B.1.7 Sample Failure 7: Violating a quantitative constraint 20

B.1.8 Sample Success: Marked correct despite multiple rewrites 21

B.2 Sample JSON 21

C Qualitative Failure Modes **23**

D Evaluation Details **23**

D.1 Infrastructure 23

D.2 LLM-Judge prompts 23

D.3 Dataset Construction and Quality Assurance 26

D.3.1 Sample curation and augmentation 26

D.3.2 External Expert Quality Audit 27

D.4 Human Agreement 28

D.4.1 Setup 28

D.4.2 LLM Judge Alignment and Statistical Significance 29

D.5 Additional results 29

E Prompt Ablation Experiments **30**

E.1 More Explicit Conversational Prompts (Tier 2) 30

E.2 Explicit Preservation and Updating Instructions (Tier 3) 31

E.3 A Detailed Causal Recipe (Tier 4) 32

A Background on Counterfactuals

A.1 Structural Causal Models and the Pearl Causal Hierarchy

We provide an optional primer on Pearl's structural account of causality. We denote random variables by capital letters, X, and their values by lowercase letters, x. Bold letters, $\mathbf{X}$ and $\mathbf{x}$, denote sets of variables and their corresponding values. $P(\mathbf{x})$ is shorthand for $P(\mathbf{X} = \mathbf{x})$.

Structural causal models. We use *Structural Causal Models* (SCMs) to describe the generative process for a system of interest (Bareinboim, 2025; Pearl, 2000). An SCM $\mathcal{M}$ is a tuple $\langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$, where $\mathbf{V}$ is the set of endogenous variables, $\mathbf{U}$ is the set of exogenous variables representing background factors outside the model, and $P(\mathbf{u})$ is their joint distribution. $\mathcal{F} = \{f_i\}$ is a set of structural functions, each of which causally generates the value of $V_i \in \mathbf{V}$ as

$$V_i \leftarrow f_i(\mathbf{U}_i, \mathbf{Pa}_i),$$

where $\mathbf{U}_i \subseteq \mathbf{U}$ and $\mathbf{Pa}_i \subseteq \mathbf{V} \setminus \{V_i\}$. We consider recursive SCMs, so these assignments can be evaluated in an acyclic order. A realization $\mathbf{U} = \mathbf{u}$, called a *context*, determines the values of all endogenous variables through these functions.

Causal graphs. Each SCM induces a *causal graph*, or *causal diagram*, $\mathcal{G}$, containing a vertex for each $V_i \in \mathbf{V}$. A directed edge $V_j \to V_i$ is included when $V_j \in \mathbf{Pa}_i$, and a bidirected edge $V_i \leftrightarrow V_j$ represents dependence between their exogenous inputs $\mathbf{U}_i$ and $\mathbf{U}_j$, including shared unobserved causes. Recursiveness implies that $\mathcal{G}$ has no directed cycles. An ancestor of a variable is a vertex from which it can be reached along a directed path; a descendant is a vertex reachable from it along a directed path. Following our convention, these sets include the variable itself. The graph encodes qualitative causal dependencies; it does not, by itself, specify the structural functions or the distribution of background factors.

Interventions and potential responses. The operator $\mathrm{do}(\mathbf{X} = \mathbf{x})$ indexes a submodel $\mathcal{M}_{\mathbf{x}}$ in which the functions generating $\mathbf{X}$ are replaced by the constant assignments $\mathbf{X} \leftarrow \mathbf{x}$. All other structural functions and the exogenous distribution remain unchanged. The value of a variable Y in this submodel, for context $\mathbf{u}$, is its *potential response*, denoted $Y_{\mathbf{x}}(\mathbf{u})$. Unlike conditioning on $\mathbf{X} = \mathbf{x}$, an intervention overrides the mechanisms that naturally determine $\mathbf{X}$.

Layers of the PCH. The *Pearl Causal Hierarchy* (PCH), also known as the *ladder of causation*, distinguishes three progressively richer regimes of reasoning (Bareinboim et al., 2022; Pearl & Mackenzie, 2018):

1. **Association ($\mathcal{L}_1$; seeing).** Observational quantities, such as $P(Y = y \mid X = x)$, describe the system under its natural mechanisms.
2. **Intervention ($\mathcal{L}_2$; doing).** Interventional quantities, such as $P(Y = y; \mathrm{do}(X = x)) = P(Y_x = y)$, describe the system when specified mechanisms are overridden.
3. **Counterfactuals ($\mathcal{L}_3$; imagining).** Counterfactual quantities relate potential responses under possibly different regimes for the same unit. For example, $P(Y_x = y \mid X = x', Y = y')$ asks what Y would have been had X been fixed to x, given that $X = x'$ and $Y = y'$ were actually observed.

Higher layers subsume lower ones, but lower-layer distributions do not, in general, determine higher-layer quantities without additional causal assumptions.

Connection to counterfactual text editing. Given factual evidence $\mathbf{E} = \mathbf{e}$, counterfactual inference proceeds through *abduction*, updating the distribution of background factors to $P(\mathbf{u} \mid \mathbf{e})$; *intervention*, constructing $\mathcal{M}_{\mathbf{x}}$; and *prediction*, evaluating the resulting potential responses. For discrete exogenous variables and evidence of positive probability,

$$P(Y_{\mathbf{x}} = y \mid \mathbf{e}) = \sum_{\mathbf{u}:\, Y_{\mathbf{x}}(\mathbf{u}) = y} P(\mathbf{u} \mid \mathbf{e}).$$

The factual and counterfactual scenarios thus share the same background context. Non-intervened variables that are not descendants of the intervention remain unchanged, while changes to descendants depend on the structural mechanisms and context. This supplies the formal basis for CoBe's intended editing task: preserve causally unaffected facts and revise the consequences of the requested change within the given scenario.

A.2 Subjunctive Conditionals and Conversational Counterfactuals

Next, we illustrate the task of conversational counterfactual text editing using the Oswald/Kennedy example from Sec. 1 and Fig. 2. Consider a text $T = t$ describing the shooting of President Kennedy by Lee Harvey Oswald. Variables A, B, C annotate events in the story.

Original text (T): President Kennedy died ($B = 0$) on Nov 22, 1963. He was shot by Lee Harvey Oswald ($A = 1$), who likely acted alone ($C = 0$).

Question	Grammar	Formal query	Causal ladder	Answer
1. Given that Kennedy died on Nov 22, what *would have happened* if Oswald *had not* shot him?	Subjunctive *(would have...had)*	$P(T_{A=0} \mid A = 1, B = 0, C = 0)$	Rung 3	Kennedy would have been alive.
2. Given that Kennedy died on Nov 22, what *happened* if Oswald *did not* shoot him?	Indicative *(did...if)*	$P(T \mid A = 0, B = 0)$	Rung 1	Someone else shot Kennedy.

This is a classic example, going back at least to Ernest Adams on the difference between subjunctive and indicative conditionals, and later discussions in the Stalnaker-Lewis tradition (Adams, 1970; Lewis, 1973; Stalnaker, 1968). What separates the two text questions is that the first asks a counterfactual question and uses a *subjunctive* conditional (*"what would have happened...had"*), whereas the second asks an associational question and uses an *indicative* conditional (*"what did happen...if"*).

The use of **subjunctive or indicative language is thus the intuitive linguistic indicator of whether a natural language question is to be interpreted as a Rung 3 or Rung 1 question**, respectively (Ippolito, 2013; von Fintel, 2012). We stress that our scope is *conversational* language. Humans are capable of hearing a question like *"what would have happened in this scenario if I had done something different?"* and understanding that this means background factors and prior events should be kept fixed, and the intervention should only affect downstream events. They require no formal background about causal graphs or counterfactuals, nor a detailed step-by-step procedure such as "hold causal ancestors fixed" in order to answer this question. This is the native skill we evaluate in our benchmark.

Conversational counterfactual text edit: Answering question 1 involves the three steps of counterfactual text editing below. This protocol is inspired by the 3-step process of *Abduction-Intervention-Prediction* in Pearl's structural causal model framework described in the previous Sec. A.1.

(a) We situate ourselves in the given scenario ($A = 1, B = 0, C = 0$) and hold constant the background factors that may have produced this timeline of events, such as facts about Kennedy's public schedule and security detail;

(b) We imagine an external intervention that miraculously stops Oswald from shooting, written symbolically as $do(A = 0)$; and

(c) We produce text $T = t'$ that reflects the likely downstream effects of this edit in the original scenario.

In this example, we are thus sampling from the distribution $P(T_{A=0} \mid A = 1, B = 0, C = 0)$: the distribution of text had we intervened with $do(A = 0)$ in the original scenario having $A = 1, B = 0, C = 0$. Notably, this edit should hold fixed any background factors and other events that are not causally downstream of the edited events, as shown in green in Fig. 2.

Associational text generation: By contrast, answering question 2 involves generating alternative text by sampling from $P(T \mid A = 0, B = 0)$. This is the distribution of text we are likely to see, conditioned on the events $A = 0, B = 0$, possibly updating background conditions and prior events to maintain factual correlation as shown in red in Fig. 2.

B Dataset Samples

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. B.1). Then, we provide a sample JSON format in our dataset (App. B.2).

B.1 Examples

The examples below illustrate preservation of unaffected facts, revision of intervention-dependent facts, and consistency of causal connectors. Blue highlights facts to preserve; purple marks the requested intervention; green marks appropriate revisions; and red identifies the error under discussion. Illustrative answers are not scored and need not be unique. Author-annotated reasoning explains the intended interpretation and is not part of the dataset.

B.1.1 Sample Failure 1: Spatial consequence

Example 126v7

Story 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 in this specific scenario instead had Sato been walking east across the Brooklyn Bridge at 6 PM instead of 8 AM.*

Llama3-8B response 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.

Illustrative answer (not scored) 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 sun was in front of Sato.

Author-annotated reasoning (not part of the dataset) Changing the time leaves Sato's direction of travel and Manhattan's relative position unchanged. A visible setting sun would be in the western sky, behind an eastward walker rather than in front. The illustrative answer shows a daylight realization; the criteria do not require sunlight to remain visible at 6 PM.

B.1.2 Sample Failure 2: Misattributing an unchanged event

Example 311v21

Story Vera is a highly uncharismatic politician, but a major scandal ruined her opponent's reputation, so she won the election.

Query *Rewrite the original text, based on what would have happened in this scenario had Vera been a highly charismatic politician and lost the election instead.*

Claude Opus 4.7 response Vera was a highly charismatic politician, but a major scandal ruined her reputation, so she lost the election.

Illustrative answer (not scored) Vera is a highly charismatic politician, and a major scandal ruined her opponent's reputation, but she lost the election.

Evaluation criteria
Should NOT change: A major scandal ruined her opponent's reputation.
Should change: Vera is a highly uncharismatic politician. She won the election.

Author-annotated reasoning (not part of the dataset) The query stipulates both charisma and electoral defeat; it does not ask the model to predict the election from charisma. The opponent's scandal is unchanged background in the intended scenario. Reassigning the scandal to Vera changes that background to rationalize the requested defeat.

B.1.3 Sample Failure 3: Retaining an inapplicable consequence

Example 309v26

Story It was raining and Timmy had soccer practice. The field was too wet, so practice was canceled.

Query *Rewrite the original text, based on what would have happened in this scenario had Timmy had squash practice instead.*

GPT 5.4 Pro response It was raining and Timmy had squash practice. The court was too wet, so practice was canceled.

Illustrative answer (not scored) It was raining, and the outdoor soccer field was still too wet. Timmy had squash practice instead, and it went ahead on an indoor court.

Evaluation criteria
Should NOT change: It was raining.
Should change: The practice venue was too wet, so the practice was canceled.

Author-annotated reasoning (not part of the dataset) Here, "practice venue" refers to the venue used for the activity in the rewritten scenario. Under the ordinary indoor-squash interpretation, rain making the outdoor soccer field too wet is no longer a reason to cancel that practice. The response transfers the wet-venue explanation to the squash court. The outdoor field need not become dry, and the criterion does not rule out every other possible reason for cancellation.

B.1.4 Sample Failure 4: Unsupported causal connector

Example 115v3

Story The structural engineer observed that the primary support beam had a three-inch fissure across its center. Due to the high risk of immediate collapse, the city council ordered the building to be evacuated immediately.

Query *Rewrite the original text, based on what would have happened instead had the city council decided to keep the building open for business, rather than order an evacuation.*

GPT 5.4 Pro response The structural engineer observed that the primary support beam had a three-inch fissure across its center. Due to the high risk of immediate collapse, the city council decided to keep the building open for business, rather than order an evacuation.

Illustrative answer (not scored) The structural engineer observed that the primary support beam had a three-inch fissure across its center. Despite the high risk of immediate collapse, the city council decided to keep the building open for business, rather than order an evacuation.

Evaluation criteria
Should NOT change: The support beam had a three-inch fissure. There was a high risk of collapse.
Should change: The city council ordered the building to be evacuated immediately.

Author-annotated reasoning (not part of the dataset) The intervention changes the council's decision, not the observed damage or the stated risk. The response incorporates that decision but retains "due to," presenting the danger as the reason to keep the building open. That connection does not follow from the given scenario. "Despite" expresses the intended contrast; neutral wording without a causal connector would also be valid. No collapse or injury is required.

B.1.5 Sample Failure 5: Overlooking an unchanged purpose

Example 403v1

Story I am going to the shop to repair my umbrella. It is raining heavily outside. I took my umbrella when I went out.

Query *Rewrite the original text, based on what would have happened in this scenario had it not been raining at all instead.*

GPT 5.4 Pro response I am going to the shop to repair my umbrella. It is not raining outside at all, so I did not take my umbrella when going out.

Illustrative answer (not scored) I am going to the shop to repair my umbrella. It is not raining outside at all. I still took my umbrella when I went out, since I needed to get it repaired.

Evaluation criteria

Should NOT change: I am going to the shop to repair my umbrella. I took my umbrella when I went out.
Should change: It is raining heavily outside.

Author-annotated reasoning (not part of the dataset) The repair errand remains unchanged and supplies a reason to bring the umbrella even without rain. The response changes the weather correctly but treats rain as the only reason for carrying the umbrella, overlooking the purpose explicitly stated in the story.

B.1.6 Sample Failure 6: Altering a historical fact

Example 419v1

Story Hari's favorite color is red. He discovers a childhood photo of his girlfriend wearing a red dress. He had never seen this photo before.

Query *Rewrite the original text, based on what would have happened instead had Hari's favorite color been blue instead of red.*

Claude Opus 4.7 response Hari's favorite color is blue. He discovers a childhood photo of his girlfriend wearing a blue dress. He had never seen this photo before.

Illustrative answer (not scored) Hari's favorite color is blue. He discovers a childhood photo of his girlfriend wearing a red dress. He had never seen this photo before.

Evaluation criteria
Should NOT change: Hari discovers a previously unseen childhood photo of his girlfriend. She is wearing a red dress in the photo.
Should change: Hari's favorite color is red.

Author-annotated reasoning (not part of the dataset) The requested change concerns Hari's preference, not the historical contents of the photograph. Changing the dress to blue makes it match his new preference but alters an unaffected fact. The criteria make no demand about his emotional reaction to the photo.

B.1.7 Sample Failure 7: Violating a quantitative constraint

Example 317v3

Story The water level in the bathtub is rising rapidly because the main faucet is pouring water in at a rate of 5 gallons per minute.

Query *Rewrite the original text, based on what would have happened instead had the drain been opened, leaking water at 8 gallons per minute.*

Claude Opus 4.7 response The water level in the bathtub is rising rapidly because the main faucet is pouring water in at a rate of 5 gallons per minute. The drain was opened, leaking water at 8 gallons per minute.

Illustrative answer (not scored) The faucet continues to pour water into the bathtub at 5 gallons per minute, while the opened drain removes 8 gallons per minute. The water level falls as the tub loses a net 3 gallons per minute.

Evaluation criteria
Should NOT change: The main faucet is pouring water in at a rate of 5 gallons per minute.
Should change: The water level in the bathtub is rising rapidly.

Author-annotated reasoning (not part of the dataset) The faucet's inflow remains fixed, and the intervention adds an outflow exceeding it. While water remains and the stated rates hold, the net volume change is $5 - 8 = -3$ gallons per minute, so the water level falls. The response includes the requested drain opening but incorrectly retains the original rising level.

B.1.8 Sample Success: Marked correct despite multiple rewrites

Example 217v10

Story 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 in this specific scenario instead had Lisa watched a movie on her laptop.*

Gemini 3.1 Pro response Three alternative rewrites are shown below; introductory text and option headings are omitted.

1. 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.
2. 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.
3. 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.

Illustrative answer (not scored) Lisa watches a movie on her laptop at the terminal. Her connecting flight has been delayed by three hours.

Evaluation criteria
Should NOT change: Lisa is at the terminal. Her connecting flight has been delayed by three hours.
Should change: Lisa is reading a thick novel and using a paper bookmark to keep her place in that reading activity.

Author-annotated reasoning (not part of the dataset) Each alternative replaces the foregrounded reading activity with watching a movie while preserving the location and flight delay. Headphones, pausing, and charging are optional elaborations, not required consequences. The change does not require Lisa to discard the bookmark; it may remain in the closed book.

B.2 Sample JSON

Sample JSON entry: 317v3


```
{
  "Core Set ID": 317,

  "Variation ID": 3,

  "Domain": "Everyday",

  "Story": "The water level in the bathtub is rising rapidly because
   the main faucet is pouring water in at a rate of 5 gallons per
   minute.",

  "Query": [
    "Rewrite the original text, based on what would have happened
   instead had the drain been opened, leaking water at 8 gallons per
   minute.",
    "Based on the preceding text, rewrite the scenario to reflect
   what would have happened if the drain had been opened, leaking
   water at 8 gallons per minute.",
    "Rewrite the above passage to illustrate what would have occurred
   had the drain been opened, leaking water at 8 gallons per minute."
   ],

 "Illustrative answer (not scored)": "The faucet continues to pour
   water into the bathtub at 5 gallons per minute, while the opened
   drain removes 8 gallons per minute. The water level falls as the
   tub loses a net 3 gallons per minute.",

  "Evaluation criteria": [
```

```
    "The main faucet is pouring water in at a rate of 5 gallons per
   minute.",
    "The water level in the bathtub is rising rapidly."
  ]
}
```

The entry above illustrates the dataset schema. The two strings in `Evaluation criteria` list, respectively, facts to preserve and original facts to revise. The `Illustrative answer (not scored)` provides one example of a valid rewrite; multiple valid answers are possible.

Here, the faucet's inflow remains at 5 gallons per minute. The requested intervention introduces an outflow of 8 gallons per minute, so the water level falls rather than rises.

C Qualitative Failure Modes

The following qualitative patterns elaborate the evaluation checks in Sec. 3.1. They describe errors in the rewritten text rather than establishing the internal reasoning process that produced them. The patterns can overlap and do not introduce additional scoring criteria. The linked examples illustrate each pattern.

1. **Overlooking an unchanged reason for an action.** A rewrite removes an action after changing one associated condition, even though the story supplies another reason for the action that remains in place. In 403v1, the umbrella is still needed for the repair errand when the rain is removed.
2. **Changing background facts to rationalize the intervention.** A rewrite alters an unaffected event to make a stipulated outcome appear more expected. In 311v21, the query specifies both greater charisma and electoral defeat, but the response reassigns the opponent's scandal to Vera.
3. **Retaining a consequence whose stated reason no longer applies.** A rewrite incorporates the requested intervention but carries over an explanation tied to the original situation. In 309v26, the response transfers the wet-venue cancellation from outdoor soccer to the counterfactual squash session, under the intended indoor interpretation.
4. **Violating physical or quantitative constraints.** A rewrite retains an outcome incompatible with the conditions in the counterfactual scenario. In 317v3, an inflow of 5 gallons per minute and an outflow of 8 gallons per minute imply a falling water level while the stated flows hold, not a rising one.
5. **Updating spatial consequences inconsistently.** A rewrite changes an event but fails to keep its consequences consistent with the preserved spatial setting. In 126v7, an eastward walker cannot face a setting sun in the western sky. The criterion does not require the sun to remain visible at 6 PM.
6. **Using an unsupported causal connector.** A rewrite changes the requested event but retains wording that assigns it an unsupported cause. In 115v3, "due to the high risk of immediate collapse" is carried over as the explanation for keeping the building open. The issue is the asserted causal relationship, not the absence of a particular replacement word such as "despite."
7. **Changing historical facts to match a preference.** A rewrite makes an unaffected past event agree with a newly stipulated preference. In 419v1, changing Hari's favorite color does not change the red dress recorded in his girlfriend's childhood photograph.

D Evaluation Details

In this section, we first describe the infrastructural details on the evaluation in App. D.1, and then provide the prompts used for assessing each evaluation criterion in App. D.2. We also provide the details of human agreement in App. D.4, and then discuss additional results in App. D.5.

D.1 Infrastructure

For the open-source models the responses were generated using 1-4 NVIDIA H100 GPUs. 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.

D.2 LLM-Judge prompts

Our LLM-Judge evaluation pipeline checks the supplied criteria rather than matching responses to a single reference answer. It is designed to accommodate differences in wording and presentation, including responses with additional text or multiple valid rewrites (e.g., 217v10). Q0 examines connector usage in the rewrite, while Q1 and Q2 assess the rewrite against the first and second evaluation-criterion strings, respectively. Agreement with human judgments is reported in App. D.4.

The prompts provided to the LLM-Judge are shown below. `$response_text` denotes the model's rewrite, and `$eq1` and `$eq2` denote the first and second strings in `Evaluation criteria`.

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.).

Reply F if *any* causal connector misrepresents the relationship
  between the connected words. This includes either: the causal
  relationship is in the wrong or reversed direction, or the connected
   words are not causally related.

Reply T if *every* causal connectors signals the correct causal
  direction. If the causal relationship is reasonably plausible, reply
   T. If there are no causal connectors, reply T.

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

Example 1: rewrite says "Due to the storm, the school did not declare
  a snow day. Because they were in class during the day, the kids were
   not able to go outside and go sledding while the sun was up." → F
Example 2: rewrite says "Because it was a home derby at the Bernabéu,
  Real Madrid wore their purple away kit." → F
Example 3: rewrite says "The weather was hot, so Alex grabbed a hot
  tea" → F
Example 4: rewrite says "The weather was hot, so Alex grabbed an iced
  coffee" → T
Example 5: rewrite says "Because the weather was sunny, Alice brought
  her umbrella" → F
Example 6: rewrite says "Despite the sunny weather, Alice brought her
  umbrella" → T

Rewrite:
$response_text

Your entire reply must be exactly one character: T or F.
```

Q1: Preservation of invariant facts

```
Criterion: "Facts that should be UNCHANGED"

Check whether every claim in the criterion is preserved in the rewrite
   with the same meaning.

Reply T if *all* criterion claims are still present and factually
  unchanged, even if they are paraphrased or expressed differently.

Reply F if *any* criterion claim is omitted, contradicted, or
  materially changed.

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

Example 1: criterion says pipe leaking and floor wet; rewrite keeps
  leak but floor is dry → F
Example 2: criterion says "There was a defect at the factory"; rewrite
   says "Without the factory defect ..." → T
Example 3: criterion says "Jake scored the goal"; rewrite says "Amy
  scored the goal" → F
Example 4: criterion says "David is driving a car"; rewrite says "A
  car ran out of gas because david drove the car for 10 hours" → T
```

```
Example 5: criterion says "Defenders successfully blocked the attack
  ."; rewrite says "Defenders detected the intrusion too late. Despite
   the defenders' attempt to block the attack, they allowed the
  opponent to score the goal" → F

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

Rewrite:
$response_text

Your entire reply must be exactly one character: T or F.
```

Q2: Required changes

```
Criterion: "Facts that should be CHANGED"

Check whether every claim in the criterion has been changed so that it
   no longer holds in the rewrite with the same substantive meaning.

Reply T if *every* criterion claim is contradicted, removed, or
  otherwise changed so that the original claim no longer remains true.

Reply F if *any* criterion claim is still preserved and continues to
  hold in the rewrite, even if it is paraphrased or embedded in a
  longer statement.

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

Example 1: criterion says "brakes failed. car hit wall."; rewrite says
   "brakes engaged and car was safe" → T
Example 2: criterion says "Jake scored the goal"; rewrite says "Amy
  scored the goal" → T
Example 3: criterion says "David drove the car"; rewrite says "The car
   ran out of gas because David drove it for 10 hours" → F
Example 4: criterion says "The government ordered the street to be
  evacuated"; rewrite says "The government chooses to keep the street
  open instead of evacuation" → T
Example 5: criterion says "Alice pours water on the table. She cleaned
   the water."; rewrite says "Alice was surprised that she poured milk
   on the table. She cleaned the milk" → T
Example 6: criterion says "Defenders successfully blocked the attack
  ."; rewrite says "Defenders detected the intrusion too late,
  allowing the opponent to score the goal" → T
Example 7: criterion says "Drinking a cold water, Jake ate a spicy
  kimchi"; rewrite says "Drinking a cold water, Jake ate a mild tofu"
  → F
Example 8: criterion says "test was hard for Alice"; rewrite says "
  test would not have been hard for Alice if she tried hard" → T

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

Rewrite:
$response_text

Your entire reply must be exactly one character: T or F.
```

D.3 Dataset Construction and Quality Assurance

D.3.1 Sample curation and augmentation

The dataset construction ran in four stages:

Table 6: Dataset construction pipeline.

Stages (1-4)	Output	Steps taken	Quality assurance and filtration
Counterfactual example collection	**425 candidate counterfactual examples**	A cohort of student researchers constructed examples following the dataset schema in Sec. 3.1. Each example contained an original scenario, a counterfactual edit request, and explicit evaluation criteria for LLM answers.	All contributors had completed a graduate causal inference course. Each example was required to specify a coherent counterfactual intervention, the resulting edit, and criteria identifying which facts should remain fixed and which consequences should change.
Core-set selection	**106 diverse core examples crafted**	The authors manually reviewed the 425 candidates and used them as a basis to craft a core set spanning diverse domains, event interactions, and causal structures.	We rejected candidates with ill-defined counterfactuals, ambiguous or unsound causal relationships, incoherent scenarios, or evaluation criteria that did not clearly distinguish valid from invalid edits.
Augmentation of core set	**25 candidate variations per core example**	Each core example was augmented using LLMs to produce variations across diverse themes and domains while preserving the original causal structure and logical relationships among events.	Generation was repeated until 25 satisfactory variations were obtained for each core example. We rejected and regenerated variations that were incoherent or contained unsound evaluation criteria.
Final verification and filtering	**2,650 examples; 2,000 final examples retained after quality checks**	The authors manually reviewed every set of 25 augmented examples and its associated evaluation criteria.	Every augmented set of 25 variations was checked for story coherence and for the correctness and clarity of its evaluation criteria. Only examples passing this review were included in the final dataset. External expert feedback led to additional examples being removed.

Augmentation and development checks. Augmentation used Gemini-3.1-Pro and GPT-5.4-Pro through separate accounts without conversational memory. The authors reviewed generated variations and their evaluation criteria before inclusion. During development, we also spot-checked selected examples on language models. These checks did not determine inclusion: examples were retained even when model responses were correct, or were initially incorrect but later became correct with model updates. Selection prioritized scenario quality, diversity of domains and causal structures, and clarity of the evaluation criteria.

Author review and filtering. At stages 2 and 4 above, each item was reviewed by at least two authors. Review covered scenario coherence, clarity of the requested intervention, and the correctness of the facts designated for preservation or revision. Authors used color-coded markings to indicate approval and disagreement, and discussed disagreements before a final decision. We only retained

examples where both reviewing authors came to a consensus that the example and its criteria were acceptable. We do not report a quantitative inter-author agreement statistic for this stage.

Grounds for rejection included ill-defined counterfactuals, ambiguous or unsound causal relationships, incoherent scenarios, and evaluation criteria that did not clearly distinguish acceptable from unacceptable edits. Initial author filtering was followed by external expert review and additional exclusions informed by that review. Together, these checks removed 650 incorrect or weak examples of the 2,650 augmented examples, leaving 2,000 final scenarios.

D.3.2 External Expert Quality Audit

We recruited three experts in causal inference who had not participated in dataset construction. After initial author filtering, we randomly sampled 100 examples from the remaining candidate pool. All three experts independently reviewed the same 100 examples, assessing coherence, grammatical quality, and validity of the counterfactual request on a five-point scale. They also assessed the diversity of the reviewed collection and provided written feedback. The experts were not shown the preservation and revision criteria. This review assessed scenario and query quality, complementing the authors' separate verification of those criteria. The questions they answered were,

- **Coherence**: Does the story present a clear and logical flow of events? Is its overall structure coherent?
- **Grammar**: Is the text grammatically natural, intelligible, and easy to read?
- **Counterfactual validity**: Does the query express a valid counterfactual intervention on the events described in the story?
- **Dataset diversity**: Taken together, do the reviewed examples represent a diverse range of scenarios, domains, and causal structures?
- **Additional feedback (text)**: Are there any other issues or observations that should be brought to our attention?

Interestingly, expert feedback identified a systematic inconsistency and led to further exclusions of examples exhibiting that issue. Namely, the experts noticed that some examples contained a causal relationship that was contrived. We noticed this was due to a systemic issue from a batch of augmented samples. These exclusions are included in the total of 650 reported above.

Across all 100 audited examples, mean scores were **4.85/5** for coherence, **4.94/5** for grammatical quality, and **4.71/5** for counterfactual-query validity. The reviewed collection received a mean diversity rating of **5/5**. These scores include examples subsequently excluded and therefore describe the audited candidate sample before the final exclusions.

Reviewers further noted that the dataset covers *"social life, physics, biology, chemistry, sports and other domains"* and contains *"a good mix of random daily life scenarios and science ones and a few calculation ones too with a good variety of graph types and intervention locations."*

D.4 HUMAN AGREEMENT

D.4.1 SETUP

To assess agreement between our LLM judge and independent human judgments, we randomly sampled 250 model responses, each paired with its original story and counterfactual edit query, from the outputs of one model in our evaluation suite. Three expert human annotators independently assessed each response using the general rubric below. Annotators were shown the story, query, and model response, but not the story-specific `Evaluation criteria`. They were provided with two worked examples illustrating the intended counterfactual interpretation. The majority human judgment served as the reference label for comparison with the LLM judge.

Human-rating instructions

```
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 activity is to get human opinions
   on outputs produced by querying LLMs with certain interesting
  counterfactual queries.

You would be provided with 250 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.
```

Human-rating rubric and examples

```
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.
```

D.4.2 LLM JUDGE ALIGNMENT AND STATISTICAL SIGNIFICANCE

We then processed the same 250 sampled instances through our LLM-as-a-judge pipeline using a few-shot prompt.

The LLM achieved an approximate *Accuracy* of 92.1%, *Precision* of 82.5%, *Recall* of 91.7% and *F1 Score* of 86.81% (where failure is the positive class) 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 $\kappa = 0.812$. This demonstrates that the LLM achieves a level of inter-rater reliability with the consensus that is *almost-perfect*. Given the sample size ($N = 250$), this agreement provides sufficient statistical confidence to scale the evaluator across the broader benchmark. Further, the high agreement is strong evidence that our evaluation criteria that we use in the dataset are themselves sound.

D.5 ADDITIONAL RESULTS

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

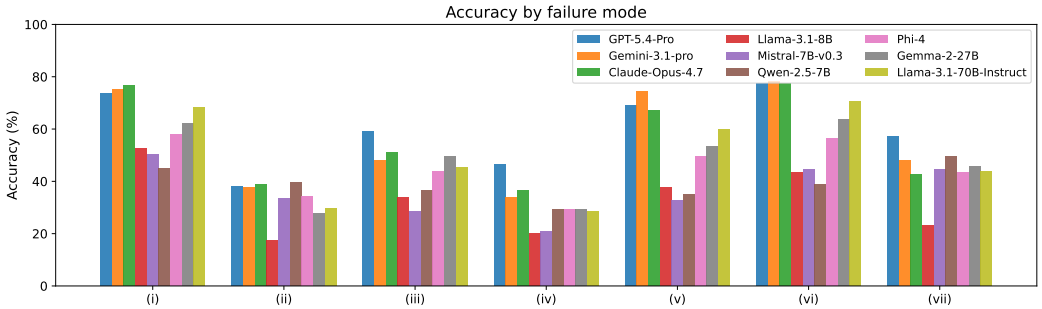

Figure 9: Accuracy across different failure modes.

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 to effectively update 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. 7.

E Prompt Ablation Experiments

We evaluate four tiers of prompting: the original benchmark prompts and three additional ablations that provide progressively more explicit guidance about the intended counterfactual edit.

- **Tier 1 (A–C): Original conversational prompts.** The three query templates in Sec. 3 specify a hypothetical change and ask the model to rewrite the story, leaving the intended editing procedure implicit.
- **Tier 2 (D–G): More explicit conversational prompts.** The templates in App. E.1 more strongly cue an intervention in the given scenario, while leaving the preservation and updating rules implicit. They remain within our conversational task setting.
- **Tier 3 (H1): Explicit preservation and updating instructions.** The template in App. E.2 explicitly instructs the model to preserve prior and unaffected facts and update the consequences of the requested change. This supplies an editing rule that the model must infer in the primary task.
- **Tier 4 (H2): A detailed causal recipe.** The template in App. E.3 provides a step-by-step procedure using causal terminology, including graph construction, intervention, and updating causal descendants.

Our scope distinction concerns whether the prompt supplies the editing rules, rather than simply whether it uses conversational language. Tiers 1 and 2 leave these rules implicit; Tiers 3 and 4 provide explicit guidance and are therefore outside the primary task setting defined in Sec. 1. We nevertheless report them as instruction-assisted controls to examine how performance changes when some of the inference burden is removed.

We evaluated all four tiers on the same 500 randomly sampled examples for GPT-5.4-Pro, Gemini-3.1-Pro, and Qwen-2.5-7B. Table 7 reports their pass rates, with scores averaged over the corresponding templates for Tiers 1 and 2.

E.1 More Explicit Conversational Prompts (Tier 2)

We provide four additional templates that strengthen the intended interventionist interpretation without specifying which facts to preserve or giving a causal procedure. These phrasings are designed to

Table 7: Pass rates across four prompt tiers on the same 500 randomly sampled examples. A–C and D–G are conversational prompts; H1 supplies explicit preservation and updating instructions, and H2 supplies a detailed causal recipe.

Model	Tier 1 A–C	Tier 2 D–G	Tier 3 H1	Tier 4 H2
GPT-5.4-Pro	66.2%	71.7%	73.5%	81.3%
Gemini-3.1-Pro	58.7%	64.4%	67.1%	76.1%
Qwen-2.5-7B	36.2%	33.3%	25.0%	32.0%

discourage alternative readings such as *backtracking counterfactuals*, in which background conditions may be revised to accommodate the hypothesized event (von Kügelgen et al., 2023). In particular, references to an external intervention or to going back in time and altering an event emphasize a change within the given scenario. These templates retain the requirement that the model infer how to carry out the edit from the conversational request.

Additional prompts.

(D) Rewrite the above text to illustrate what would have happened in this specific scenario, had <intervention happened>.

(E) Rewrite the above text to incorporate an imagined external intervention in this scenario, where instead of <original scenario happened>, <intervened scenario happened>.

(F) Rewrite the story described above, had we gone back in time and altered <original scenario happened> to <intervened scenario happened>.

(G) Rewrite the above text based on how events would have unfolded, but where instead of <original scenario happened>, <intervened scenario happened>.

Results. Averaging over D–G, GPT improves from 66.2% to 71.7% (+5.5 percentage points) and Gemini from 58.7% to 64.4% (+5.7 points). Qwen instead declines from 36.2% to 33.3% (−2.9 points). Thus, stronger conversational cues provide moderate gains for the two frontier models but do not consistently improve performance across the models tested. Substantial failure rates persist under these alternative phrasings.

E.2 Explicit Preservation and Updating Instructions (Tier 3)

Template H1 explicitly states the preservation and updating rules that are left implicit in Tiers 1 and 2. Although the instruction uses everyday language, it supplies part of the intended editing policy rather than requiring the model to infer it from the counterfactual request. It is therefore outside our primary task setting, but provides a useful control for the effect of explicit task guidance.

Template H1: Explicit editing instructions

Original story: ...

Rewrite prompt: Rewrite the original text based on the counterfactual scenario in which `<intervention>`.

Keep prior and unaffected facts unchanged, and update the consequences of the requested change.

Your final answer should consist only of the rewritten sentence.

Results. Relative to Tier 1, H1 raises the reported pass rates of GPT and Gemini by 7.3 and 8.4 percentage points, respectively. The additional gains over Tier 2 are smaller: 1.8 and 2.7 points. For Qwen, performance instead falls to 25.0%, below both conversational tiers. Explicitly stating the editing rules therefore has a model-dependent effect and does not uniformly resolve the observed failures.

E.3 A Detailed Causal Recipe (Tier 4)

Template H2 provides a detailed procedure grounded in structural causal models. Unlike the conversational templates, it explicitly specifies how to interpret the requested change and how to update the story. This evaluates performance with a supplied causal procedure; the model must still infer the scenario's causal structure and apply that procedure correctly.

Template H2: Detailed causal recipe

Original story: ...

Perform the following steps:

(1) *Identify the key variables in the original story above, following the semantics of Pearl's structural causal models;*

(2) *Provide the causal graph connecting these variables and add any unobserved confounders that might be present;*

(3) *The rewrite prompt below specifies a change in the story; treat this change as an exogenous intervention to be interpreted in the sense of the do(V=v) operation in causal semantics; update the variables in the graph under this intervention with the causal effects cascading only to causal descendants, leaving graphical non-descendant variables unchanged;*

(4) *Provide the full text of the rewritten counterfactual story under this intervention.*

Rewrite prompt: Rewrite the original text based on the counterfactual scenario in which `<intervention>`.

Your final answer should consist only of the rewritten counterfactual story.

Results. With the detailed causal recipe, GPT reaches 81.3% and Gemini reaches 76.1%, increases of 7.8 and 9.0 percentage points over H1. Qwen improves from 25.0% to 32.0%, but remains below its performance under both conversational tiers. The strongest result, GPT's 81.3% pass rate, still leaves 18.7% of evaluated responses failing the benchmark checks. Performance therefore remains below saturation in this evaluation, even when a detailed causal procedure is provided.

Together, these experiments distinguish robustness to conversational phrasing from performance under explicit task guidance. Supplying the causal procedure can substantially improve performance for some models, but those gains are obtained under a different level of assistance from the primary conversational task.