The Illusion of *What If*: Evaluating the Breakdown of Counterfactual Reasoning in LLMs

Yucheng Wang[♠*], Yuetian Du[♠*], Zhengyi Liu[♠*†], Rongyu Zhang[♠], Bing Zhao[♢]
Boyu Yang[♢], Ming Kong[♠], Lin Qu[♢], Hu Wei[♢‡], Jie Liu[♣‡], Qiang Zhu[♠‡]

[♠] Zhejiang University [♢] Alibaba Group [♣] City University of Hong Kong
12621204@zju.edu.cn zhuq@zju.edu.cn

Abstract

Counterfactual reasoning requires models to reason beyond the observed world and explain how altered conditions propagate through downstream consequences. Existing benchmarks largely target bounded settings with fixed variables or single gold outcomes, overlooking open-domain scenarios requiring causal-process evaluation. To this end, we present **WhatIfBench**, a diagnostic benchmark for open-domain, open-form, long-horizon counterfactual causal reasoning, containing 220 what-if questions across STEM, HSS, and Hybrid scenarios. To evaluate free-form responses, we further propose **PRISM**, which first converts each natural-language explanation into a Response-Derived Semantic Causal Graph of events, states, and mechanisms. On top of this graph, PRISM then jointly applies a Process Metric assessing graph-level causal validity and a Rubric Metric assessing answer-level explanatory adequacy. Evaluating six frontier LLMs with this framework, we find that WhatIfBench remains far from saturated: even the strongest model reaches only a 64.62% final score. Further analysis reveals persistent causal gaps, premise drift, and topology fragmentation, suggesting that fluent counterfactual narratives often mask fragile causal processes. The benchmark, code, and evaluation scripts are available at https://github.com/zju-gt/WhatIfBench.

1 Introduction

> *"We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it."*
>
> — David Lewis, *Causation*

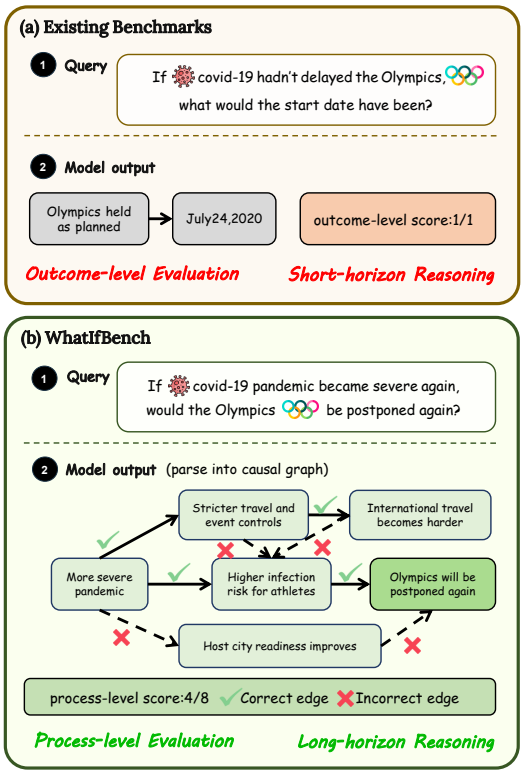

Figure 1: **Comparison of what-if reasoning evaluation paradigms.** (a) Existing benchmarks focus on outcome-level evaluation for short-horizon reasoning. (b) WhatIfBench shifts the focus to process-level evaluation for long-horizon reasoning by parsing model outputs into causal graphs and checking causal edges.

Large language models (LLMs) are increasingly expected to reason not only about what is, but about what would happen under different conditions. Such *what-if* questions arise in policy (He, 2026), science (Zheng et al., 2025), history (Nguyen, 2025), and technology (Vashishtha et al., 2025), and require more than factual recall (Bondarenko et al., 2022; Ho et al., 2022). They require a model to depart from the observed timeline, identify which factors make a difference to an outcome and propagate their consequences through interacting systems (Zhang et al., 2023). They thus

[*]Equal contribution.
[†]Work done during an internship at Zhejiang University.
[‡]Corresponding Authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 17106–17129
October 24-29, 2026 ©2026 Association for Computational Linguistics

serve as a stress test of whether a model's explanations reflect coherent causal processes rather than plausible-sounding narratives.

Existing counterfactual benchmarks test useful but bounded forms of hypothetical reasoning. CRASS converts counterfactual conditionals into QA instances (Frohberg and Binder, 2022), IfQA studies open-domain QA with counterfactual presuppositions (Yu et al., 2023), and recent formal benchmarks evaluate intervention-style reasoning over specified variables, causal structures, and target outcomes (Jin et al., 2023; Wang, 2024; Chen et al., 2025). These benchmarks test whether models can recognize a counterfactual premise, apply it to a local question, or reason within an explicitly defined causal setup, as shown in Figure 1. However, many real-world *what-if* questions are less bounded. The mechanisms that matter are often only implicit, consequences may unfold over long horizons, and a single intervention can propagate through social, technical, political, or scientific systems in multiple plausible ways (Tetlock and Belkin, 1997). This motivates an open-domain, open-form benchmark defined by broad premises rather than predefined variables or single gold outcomes.

This setting raises a central evaluation problem: *when no unique gold outcome exists, what should be judged?* Under the premise "What if smartphones had never been widely adopted?", for example, valid answers may emphasize slower mobile Internet diffusion, alternative portable devices, or different patterns of social coordination. The key issue is not whether these answers converge to the same outcome, but whether each remains faithful to the premise and supports its consequences through a coherent causal structure. A plausible conclusion can still be unreliable if it skips mechanisms, reverses causal direction, or reverts to factual-world assumptions. Following counterfactual accounts of possible alternatives (Lewis, 1973) and interventionist views of causal explanation (Woodward and Woodward, 2005), we argue that evaluation should target the causal explanation rather than the final outcome alone. A benchmark should therefore make free-form answers inspectable, and evaluate premise fidelity, mechanism coverage, domain consistency, and uncertainty handling.

Based on this principle, we introduce **WhatIfBench**, a diagnostic benchmark for counterfactual causal reasoning in open-domain, long-horizon scenarios. It contains 220 open-form what-if questions across STEM (science, technology, engineering, and mathematics), HSS (humanities and social sciences), and Hybrid scenarios. STEM questions focus on natural and technical constraints, HSS questions focus on historical, institutional, and social dynamics, and Hybrid questions involve causal propagation across domains. For each query, WhatIfBench collects model responses as open-form causal explanations, preserving the natural form of counterfactual reasoning while setting up structured evaluation.

To evaluate these explanations, we further introduce **PRISM** (**P**rocess-and-**R**ubric **I**ntegrated **S**coring **M**echanism), an evaluation framework for structured assessment of these explanations. Specifically, PRISM first converts each natural-language response into a **Response-Derived Semantic Causal Graph** by extracting events, states, and mechanisms and organizing them into a causal DAG (directed acyclic graph). On top of this graph, **Process Metric** (PM) evaluates graph-level causal validity, including edge validity, causal direction, mechanism support, and topology continuity. In parallel, **Rubric Metric** (RM) evaluates answer-level explanatory adequacy with question-specific rubrics, including premise fidelity, key mechanism coverage, domain constraints, and uncertainty handling. Together, PM and RM provide a fine-grained diagnostic view of open-domain counterfactual reasoning beyond fixed-answer matching.

Our contributions are summarized as follows:

- **WhatIfBench.** We introduce a diagnostic benchmark for counterfactual causal reasoning in open-domain, open-form, long-horizon scenarios, containing 220 open-form what-if questions across STEM, HSS, and Hybrid scenarios.
- **Response-Derived Semantic Causal Graphs.** We convert natural-language explanations into response-derived semantic causal graphs, making the causal structure expressed in model responses inspectable without requiring graph-form outputs.
- **PRISM Evaluation.** We develop **PRISM**, a process-and-rubric integrated evaluation framework that jointly assesses graph-level causal validity and fine-grained answer-level explanatory adequacy.

2 Related Work

Causal reasoning in LLMs. Causal reasoning is a key capability for evaluating whether LLMs

can go beyond surface associations and produce robust explanations of how events depend on one another (Kiciman et al., 2023; Yu et al., 2025; Wu et al., 2024; Jin et al., 2024; Miliani et al., 2025). Although LLMs often perform well on causal tasks, it remains unclear whether this reflects genuine causal understanding or parametric associations from pretraining (Chi et al., 2024; Yang et al., 2024; Yamin et al., 2024; Wang et al., 2026). Counterfactual reasoning sharpens this issue: models must depart from the factual world, reason under a hypothetical intervention, and propagate its consequences, yet they often revert to factual-world assumptions when counterfactual premises conflict with parametric knowledge (Yamin et al., 2025; Balappanawar et al., 2025). We study this capability in open-domain, long-horizon scenarios, where success requires coherent causal processes rather than plausible short answers.

Benchmarks for counterfactual reasoning. Existing counterfactual reasoning benchmarks have begun to test LLMs under hypothetical assumptions. CRASS (Frohberg and Binder, 2022) evaluates questionized counterfactual conditionals, IfQA (Yu et al., 2023) introduces counterfactual presuppositions into open-domain question answering, and CounterBench (Chen et al., 2025) studies formally specified counterfactual questions with controlled causal structures and difficulty levels. These benchmarks provide valuable diagnostics, but they are often built around short conditionals, answerable QA instances, or formal settings with relatively constrained variables and outcomes (Jin et al., 2023; Wang, 2024; Yang et al., 2024). As a result, open-form counterfactual reasoning questions remain underexplored, especially when they are underspecified, span interacting systems, and admit multiple plausible trajectories.

Evaluation of counterfactual reasoning. The evaluation of counterfactual reasoning remains largely answer-centric (He et al., 2026). IfQA (Yu et al., 2023) uses Exact Match and token-level F1 against reference answers, CRASS (Frohberg and Binder, 2022) compares predictions with human-validated answers, and CounterBench (Chen et al., 2025) evaluates whether models derive the correct final value or binary outcome under formal counterfactual settings. Such metrics are suitable when a task has a short reasoning path or a verifiable target answer, but they are insufficient when no unique gold outcome exists (Yang et al., 2024; Yu et al., 2025). In open-domain long-horizon counterfactual reasoning, a response may sound plausible while drifting from the premise, skipping mechanisms, reversing causal direction, or introducing unsupported causal links. This calls for evaluation that exposes the causal process expressed in a response and assesses explanatory adequacy beyond final-answer matching.

3 WhatIfBench

3.1 Task Definition

We define the task as *open-domain, open-form, long-horizon counterfactual causal reasoning*. Let $x_i = (q_i, I_i)$ denote a what-if instance, where q_i specifies the question context and I_i specifies an explicit counterfactual premise or intervention. Given x_i, a model M generates a free-form natural-language explanation:

$$x_i \xrightarrow{M} y_i. \tag{1}$$

The generated explanation y_i is expected to remain faithful to the premise I_i, instantiate relevant intermediate mechanisms, and propagate downstream consequences from the counterfactual condition. The task has three defining properties.

1) Open-domain indicates that instances span broad domains with underspecified causal spaces.

2) Open-form indicates that outputs are natural-language explanations rather than predefined choices or constrained spans.

3) Long-horizon indicates that reasoning should cover multi-step causal propagation across nontrivial causal structures.

3.2 Benchmark Overview

WhatIfBench contains **220 open-form what-if questions** across **113 STEM**, **79 HSS**, and **28 Hybrid** scenarios. Each instance consists of a normalized query q_i, an explicit counterfactual premise I_i, a domain category d_i, and a question-level rubric $\mathcal{R}_i$. When background materials or reference discussions are available, we use them to identify plausible constraints, mechanisms, and explanatory requirements, without assigning a unique gold conclusion.

3.3 Taxonomy of Counterfactual Scenarios

We categorize each question into one of three scenario types:

STEM covers scientific and technical mechanisms governed by relatively stable constraints, such as physical laws, biological mechanisms, engineering

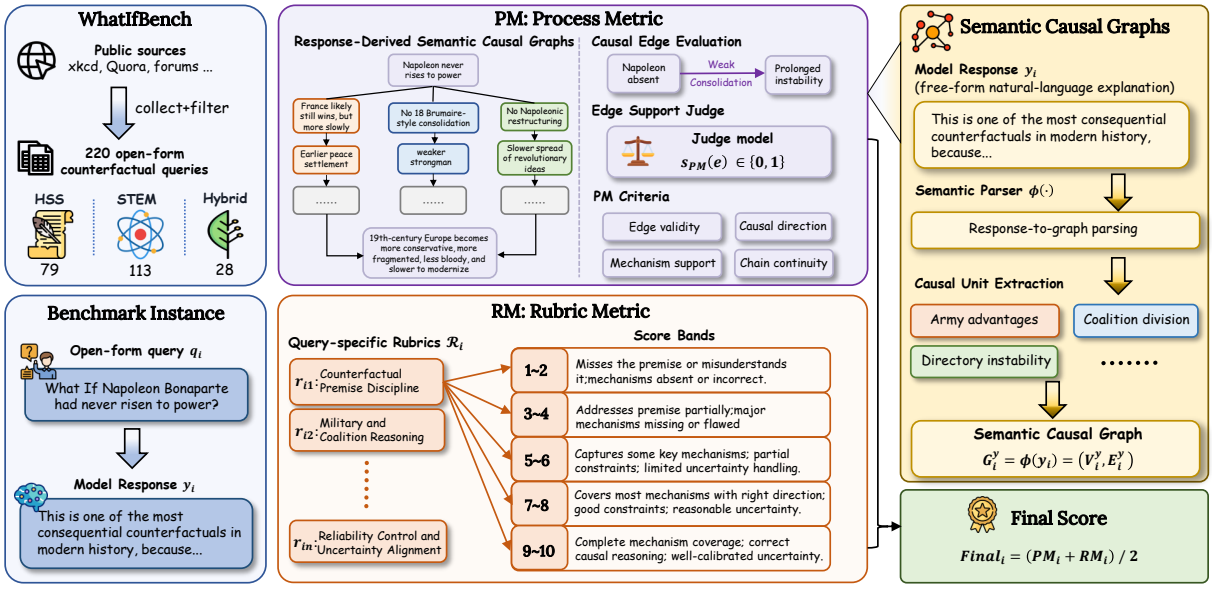

Figure 2: **Overview of WhatIfBench and PRISM.** WhatIfBench contains open-form counterfactual queries spanning STEM, HSS, and Hybrid scenarios. For each query, a model generates a free-form response, which PRISM evaluates through two complementary components: the **Process Metric** (PM), which parses the response into a response-derived semantic causal graph and validates its causal edges, and the **Rubric Metric** (RM), which scores answer-level explanatory adequacy using query-specific rubrics. The final score combines PM and RM, enabling evaluation of both causal-process validity and answer-level explanatory adequacy.

systems, and technological dependencies, testing whether models can propagate counterfactual consequences without violating domain constraints.

HSS covers historical, institutional, social, cultural, and geopolitical dynamics, emphasizing path dependence, multi-agent behavior, social feedback, and long-term institutional consequences.

Hybrid covers cross-domain scenarios where natural, technological, environmental, or resource changes propagate into economic, social, or institutional systems, testing whether models can maintain coherent causal topology across heterogeneous domains.

This taxonomy supports both overall model comparison and category-level diagnosis, allowing us to distinguish failures caused by scientific constraint violations, weak socio-historical reasoning, or cross-domain propagation errors.

3.4 Data Collection and Annotation

We construct WhatIfBench through candidate collection, filtering, annotation, and review. Candidate questions are gathered from public what-if QA sites, online discussion forums, and explanatory materials, including *xkcd What If*[1], *Worldbuilding Stack Exchange*[2], *AlternateHistory*[3], and *Quora*[4]. We rewrite retained questions into normalized queries and keep only those with an explicit counterfactual premise, sufficient contextual grounding, and nontrivial causal propagation requirements. For each retained question, we annotate the domain category d_i, summarize the plausible reasoning space, and construct a query-dependent rubric $\mathcal{R}_i$.

3.5 Quality Control

We conduct quality control at both the question and rubric levels. Question-level review checks whether each example has a clear counterfactual premise, sufficient contextual grounding, nontrivial causal propagation requirements, and an evaluable explanatory structure. Rubric-level review checks whether the criteria cover the core reasoning space while allowing multiple plausible downstream trajectories. As shown in Table 1, 36.0% of 611 candidate questions are retained after direct pass or revision, yielding 220 final questions. All corresponding rubric sets pass after direct review or revision.

[1] https://what-if.xkcd.com/

[2] https://worldbuilding.stackexchange.com/

[3] https://www.alternatehistory.com/

[4] https://www.quora.com/

Target	#	Pass	Revised	Rejected	Valid
Questions	611	14.9	21.1	64.0	36.0
Rubrics	220	92.3	7.7	0.0	100.0

Table 1: **Quality control results for WhatIfBench.** All values except # are percentages at the audit-target level. Valid denotes Pass + Revised.

4 Evaluation Framework: PRISM

We introduce **PRISM**, a **P**rocess-and-**R**ubric **I**ntegrated **S**coring **M**echanism for evaluating open-form counterfactual explanations. As shown in Figure 2, PRISM produces two scores for each response: **Process Metric (PM)**, which evaluates causal-edge validity in a response-derived graph, and **Rubric Metric (RM)**, which evaluates answer-level explanatory adequacy with frozen question-specific rubric set.

4.1 Response-Derived Semantic Causal Graph

Given a question q_i, counterfactual premise I_i, and model response y_i, PRISM parses the response into a response-derived semantic causal graph:

$$G_i^y = \phi(y_i; q_i, I_i) = (V_i^y, E_i^y). \tag{2}$$

Here, V_i^y denotes elementary discourse units or response-level claims segmented from y_i, and E_i^y denotes directed semantic relations among them. Each edge is represented as

$$e = (v_a \rightarrow v_b, r_e), \tag{3}$$

where v_a and v_b are source and target units, and r_e is a relation label.

The graph is not intended to be a gold structural causal model of the real world. Instead, it is an evaluation-oriented intermediate representation that makes the causal structure expressed in the response inspectable while allowing models to answer naturally in free text. PRISM constructs G_i^y with a Rhetorical Structure Theory (RST)-inspired parsing scheme (Mann and Thompson, 1987). It segments the response into discourse units, identifies semantic relations, and projects them into a directed graph. Relation labels include discourse relations such as ELABORATION, CONTRAST, TEMPORAL, and BACKGROUND, and causal-relevant relations such as CAUSE, CONDITION, RESULT, and CONSEQUENCE.

For process scoring, PRISM uses the causal-relevant edge subset:

$$E_i^c = \{e \in E_i^y \mid r_e \in \mathcal{C}\}, \tag{4}$$

where $\mathcal{C}$ includes CAUSE, CONDITION, RESULT, and CONSEQUENCE. Non-causal discourse edges are retained for inspection but excluded from PM.

4.2 Process Metric

PM verifies the local validity of each causal-relevant transition by checking textual support, relation type, causal direction, and mechanism-level plausibility. For each edge $e = (v_a \rightarrow v_b, r_e) \in E_i^c$, PRISM uses a strict binary judge:

$$s_{\mathrm{PM}}(e) = \begin{cases} 1, & \text{if } e \text{ is explicitly supported by } y_i, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

An edge is supported only when both endpoint claims are stated or unambiguously paraphrased in the response, the source-to-target relation is locally recoverable from the text, the relation label matches the expressed dependency, and the direction is not ambiguous, reversed, circular, or contradicted. For causal-relevant labels, the response must express an actual causal, conditional, resultative, or consequential dependency rather than a mere temporal sequence, topical association, or relation inferred from external knowledge.

The PM score is the mean support score over all causal-relevant edges:

$$\mathrm{PM}(y_i) = \begin{cases} \frac{1}{|E_i^c|} \sum_{e \in E_i^c} s_{\mathrm{PM}}(e), & |E_i^c| > 0, \\ 0, & |E_i^c| = 0. \end{cases} \tag{6}$$

Thus, PM should be interpreted as a strict causal-edge validity score over the response-derived semantic causal graph, rather than as a full judgment of real-world causal correctness.

4.3 Rubric Metric

PRISM further employs a frozen question-level rubric set $\mathcal{R}_i$ to compute the Rubric Metric (RM), which measures whether the response covers the required mechanisms, constraints, and downstream consequences for the given question.

For each question q_i, its rubric set is defined as

$$\mathcal{R}_i = \{r_{i1}, r_{i2}, \ldots, r_{ij}, \ldots, r_{in}\}, \tag{7}$$

where n is the number of rubric criteria for question q_i, and each r_{ij} specifies one question-dependent

Model	STEM		HSS		Hybrid		Overall		Final
	PM	RM	PM	RM	PM	RM	PM	RM	
GLM-5.1	57.72	52.19	49.63	46.93	52.83	49.92	54.07	49.92	51.99
DeepSeek-V4-Pro	64.43	55.81	49.53	52.74	39.42	54.33	55.78	54.50	55.12
Qwen3-Max	64.92	48.32	49.42	43.17	55.45	48.72	58.07	46.45	52.26
Gemini-3.1-Pro-Preview	68.37	53.75	52.28	50.61	51.24	50.72	60.43	52.24	56.33
Claude-Opus-4.7	69.13	56.04	**57.72**	52.02	56.08	58.48	63.39	54.82	59.11
GPT-5.5	**72.77**	**64.00**	57.38	**61.98**	**65.32**	**64.80**	**65.94**	**63.31**	**64.62**

Table 2: **Main results on WhatIfBench.** We report the Process Metric (PM) and Rubric Metric (RM) across STEM, HSS, Hybrid, and overall categories, together with the final aggregated score. All scores are percentages on a 0–100 scale. The best result in each column is shown in **bold**, and the second-best result is underlined.

requirement. These criteria cover premise fidelity, mechanism coverage, downstream consequences, domain constraints, uncertainty handling, and explanatory completeness. For each criterion r_{ij}, the evaluator assigns an integer score

$$S_{\mathrm{RM}}(y_i, r_{ij}) \in \{1, \ldots, 10\}, \tag{8}$$

according to the criterion-specific score-band descriptions associated with r_{ij}. The normalized RM score is:

$$\mathrm{RM}(y_i) = \frac{1}{10n} \sum_{j=1}^{n} S_{\mathrm{RM}}(y_i, r_{ij}). \tag{9}$$

The evaluator rewards only evidence explicitly present or unambiguously paraphrased in the answer and penalizes premise drift, missing mechanisms, ignored constraints, unsupported causal leaps, and overconfident speculation. The rubric set does not prescribe a single gold trajectory or conclusion; instead, it specifies question-dependent requirements that any plausible counterfactual explanation should satisfy.

4.4 Evaluator Standardization

We use *GPT-5.4* as the unified parser and judge for PRISM. All parsing and judging prompts are fixed across evaluated models, and the same evaluator is applied to every response under identical decoding settings. PM uses binary edge-level judgments, while RM uses criterion-specific score bands. Complete parsing schemas, judging prompts, rubric templates, and decoding settings are provided in Appendices C and D.

4.5 Final Score and Reporting

After evaluating all responses, we report model-level All PM and All RM by averaging the corresponding PM and RM scores over all examples. PM is local and process-oriented, while RM is answer-level and rubric-oriented. We use equal weighting as a default diagnostic aggregation and report PM/RM separately to avoid hiding trade-offs between process-level causal validity and answer-level explanatory adequacy. The final score is:

$$\text{Final Score} = \frac{\text{All PM} + \text{All RM}}{2}. \tag{10}$$

We report the same metrics for STEM, HSS, and Hybrid subsets to support category-level diagnosis.

5 Experiments

5.1 Experimental Setup

Models. We evaluate six frontier LLMs on WhatIfBench: GLM-5.1 (GLM-5-Team et al., 2026), DeepSeek-V4-Pro (DeepSeek-AI, 2026), Qwen3-Max (Yang et al., 2025), Gemini-3.1-Pro-Preview, Claude-Opus-4.7, and GPT-5.5. These models cover recent high-performing systems with strong general reasoning and open-form generation capabilities.

Benchmark and protocol. We run each model on the fixed WhatIfBench test set using the same response-generation template. Responses are generated with temperature $\tau = 0.6$ and a 4,096-token budget, and are submitted verbatim to the standardized PRISM pipeline, with evaluator-side decoding set to temperature $\tau = 0.0$. We do not apply model-specific prompt variants, answer normalization, or post-hoc correction. Results are reported over the STEM, HSS, and Hybrid splits, together with All PM, All RM, and the equal-weighted final score. Detailed prompts and decoding budgets are provided in Appendices C and D.

5.2 Main Results

Table 2 presents the main results on WhatIfBench. We summarize three key findings below.

Model Names	Premise Drift (%)	Causal Gap (%)	Topology Fragmentation (%)	Mechanism Missing (%)	Rubric Misalignment (%)
GLM-5.1	38.6	66.5	48.7	48.1	46.8
DeepSeek-V4-Pro	35.5	70.3	37.0	29.7	41.3
Qwen3-Max	38.8	59.9	55.3	57.2	44.7
Gemini-3.1-Pro-Preview	38.8	61.9	47.0	47.8	54.5
Claude-Opus-4.7	30.3	60.6	41.3	34.9	27.5
GPT-5.5	14.5	49.3	42.0	18.8	15.9

Table 3: Failure mode distribution among weak WhatIfBench responses. Percentages are computed over responses with final PRISM score below 60. Labels are not mutually exclusive.

Finding 1: WhatIfBench remains challenging for frontier models. Even the strongest model, GPT-5.5, achieves a final score of only 64.62, followed by Claude-Opus-4.7 with 59.11. All evaluated models remain far from perfect performance, indicating that open-domain, open-form, long-horizon counterfactual causal reasoning is still challenging for current frontier LLMs. This suggests that fluent generation does not necessarily imply reliable reasoning under counterfactual premises, especially when models must propagate consequences across interacting mechanisms.

Finding 2: Domain shifts reveal different causal weaknesses. Models perform consistently better on STEM scenarios, where counterfactual consequences are often governed by more stable scientific or technical mechanisms. HSS scenarios are generally more challenging, reflecting difficulties in modeling path dependence, institutional dynamics, and multi-agent social consequences. Hybrid scenarios are not uniformly the hardest, but they expose stronger model-specific variation and PM–RM trade-offs, suggesting that cross-domain reasoning may cover relevant outcomes without maintaining coherent causal propagation.

Finding 3: Process validity and explanatory adequacy capture complementary abilities. Overall PM is higher than overall RM for every model, but the gap varies substantially across systems. For example, Qwen3-Max achieves a relatively strong overall PM of 58.07 but a much lower overall RM of 46.45, suggesting that locally plausible causal transitions do not necessarily yield a complete or well-grounded explanation. The strongest models, GPT-5.5 and Claude-Opus-4.7, rank first and second on both PM and RM, indicating that stronger counterfactual reasoning requires both coherent causal structure and sufficient answer-level coverage. This supports the design of PRISM as a joint evaluation of process-level causal validity and rubric-level explanatory adequacy.

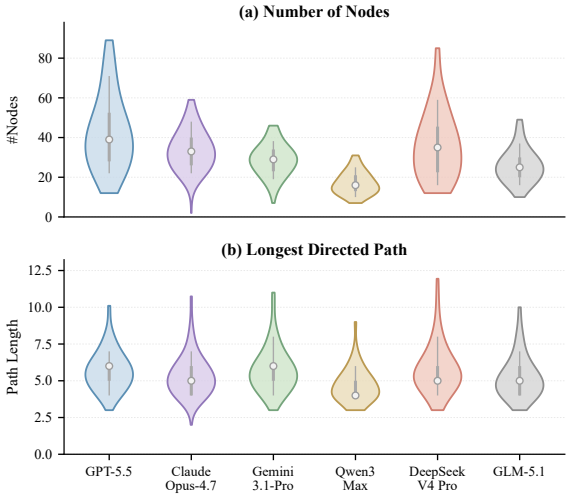

Figure 3: **Topology of response-derived causal graphs across models.** We visualize the distributions of two graph-level structural statistics extracted from parsed model responses: (a) the number of nodes, measuring the scale of instantiated causal factors, and (b) the longest directed path length, measuring the depth of causal propagation.

5.3 Failure Mode Analysis

Aggregate scores identify performance gaps, but they do not reveal which part of counterfactual reasoning breaks. We therefore analyze weak responses with final PRISM score below 60 and assign five non-mutually-exclusive labels using PRISM scores and judge rationales: **Premise Drift** for deviations from the counterfactual premise, **Causal Gap** for unsupported causal edges or missing bridges, **Topology Fragmentation** for broken long-horizon continuity, **Mechanism Missing** for insufficient process-level detail, and **Rubric Misalignment** for failures to satisfy question-specific explanatory requirements.

Finding 4: Counterfactual failures are struc-

Table 4: **Human audit of response-to-graph parsing quality.** All values are percentages. Node, Edge, Graph, and Causal Verification denote the validity of extracted nodes, directed edges, graph structure, and causal relations, respectively.

Audit Set	Node Validity	Edge Validity	Graph Validity	Causal Verification
Human1	99.6	88.7	98.3	94.2
Human2	100.0	92.3	98.3	93.4
Human3	98.7	89.5	100.0	94.7

Table 5: **Correlation with human judgments.** H–H denotes Human–Human agreement, and P–H denotes PRISM–Human agreement.

Score	Pair	Pearson r	Spearman ρ	MAE
PM	H–H	0.918	0.910	0.075
	P–H	0.812	0.846	0.101
RM	H–H	0.921	0.888	0.029
	P–H	0.875	0.726	0.034
Final	H–H	0.937	0.903	0.037
	P–H	0.833	0.815	0.059

tural rather than merely factual. Table 3 shows that weak responses are dominated by causal-process failures. **Causal Gap** is the most frequent failure mode for every model, indicating that models often mention plausible consequences without establishing the intermediate causal bridges from the counterfactual premise. Together with Topology Fragmentation and Mechanism Missing errors, this suggests that many failures arise not because responses are entirely off-topic, but because they fail to sustain coherent, mechanism-grounded causal propagation. This supports the motivation of WhatIfBench and PRISM: open-form counterfactual evaluation must inspect the structure of reasoning, not only the surface adequacy of final answers.

5.4 Visualization

Finding 5: Graph scale and causal depth capture different structural behaviors. Figure 3 shows that broader explanations are not necessarily deeper causal explanations. GPT-5.5 and DeepSeek-V4-Pro tend to produce larger response-derived causal graphs, suggesting broader causal coverage, whereas Gemini-3.1-Pro achieves comparable longest-path depth with fewer nodes. This indicates that models differ not only in how many causal elements they mention, but also in how efficiently they organize them within the response-derived causal graph. The result reinforces the need for structure-aware evaluation: surface-level explanatory richness should not be conflated with coherent causal propagation.

5.5 Correlation with Human Judgments

Because PRISM relies on automatic response-to-graph parsing and evaluator judgments, we examine its reliability from two complementary perspectives. We first audit whether the parser can recover the causal process expressed in model responses. We then compare PRISM scores with human judgments under a Human–Human agreement ceiling.

Finding 6: Response-derived causal graphs provide a reliable basis for process evaluation. Table 4 shows that response-to-graph parsing is stable across three human audit sets. Node Validity remains above 98.7%, Graph Validity above 98.3%, and Causal Verification above 93.4%, indicating that the parser generally recovers valid response-level units, coherent graph structures, and plausible causal relations. Edge Validity is lower, ranging from 88.7% to 92.3%, suggesting that causal-edge extraction is the most challenging part of the parsing pipeline. Overall, these results support the use of response-derived semantic causal graphs as a reliable basis for downstream PM evaluation.

Finding 7: PRISM captures most human evaluation signal under a human-agreement ceiling. Table 5 shows that PRISM aligns strongly with human judgments while remaining below the Human–Human reference. For PM, PRISM achieves 0.812 Pearson and 0.846 Spearman correlation, close to the Human–Human agreement of 0.918 and 0.910. For RM, PRISM obtains a high Pearson correlation of 0.875 and an MAE of 0.034, near the Human–Human MAE of 0.029, although its lower Spearman correlation indicates remaining rank-order differences. These results suggest that PRISM captures the majority of the human evaluation signal under a realistic human-agreement ceiling, supporting its use as a scalable diagnostic evaluator rather than a substitute for human assessment.

6 Conclusion

We presented **WhatIfBench**, a benchmark for open-domain, open-form, long-horizon counterfactual causal reasoning, and **PRISM**, a process-and-rubric framework for evaluating counterfactual explanations. By converting model responses into response-derived semantic causal graphs and combining process- and rubric-level assessment,

PRISM enables diagnosis beyond final-answer matching. Experiments on six frontier LLMs show that WhatIfBench remains far from saturated: models often produce fluent narratives while exhibiting causal gaps, premise drift, and topology fragmentation. These findings suggest that counterfactual reasoning should be evaluated as a causal explanation process, not merely as outcome prediction.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 42394060 and 42394064, and Alibaba Group through Alibaba Research Intern Program.

Limitations

Although WhatIfBench supports controlled evaluation of open-domain, open-form, long-horizon counterfactual reasoning, its coverage remains necessarily limited. The benchmark contains 220 questions across STEM, HSS, and Hybrid scenarios, which enables systematic comparison but cannot fully represent the diversity of counterfactual settings, specialized domains, or culturally grounded assumptions. Since the questions are collected and normalized from public English-centric sources, the benchmark may also inherit topical and stylistic biases from these sources. We will extend WhatIfBench with more domains, more questions, and broader linguistic and cultural coverage.

PRISM should also be understood as an automatic diagnostic evaluator rather than a gold-standard causal truth oracle. It evaluates response-derived semantic causal graphs rather than gold structural causal models of the world. Therefore, the Process Metric measures causal coherence, textual support, and structural consistency within a model response, but does not provide a definitive judgment of real-world causal truth. PRISM may also be affected by evaluator bias, parsing errors, and rubric design choices. We will further extend the evaluation setting beyond single-turn text-only frontier LLMs to retrieval-augmented, interactive, multimodal, and human-in-the-loop counterfactual reasoning.

Ethical Considerations

Counterfactual causal reasoning is a dual-use capability. Evaluating whether a model can reason under an explicitly stated what-if premise is, on balance, beneficial: it helps reveal unsupported causal jumps, hidden assumptions, and brittle explanations in open-domain reasoning. The same capability, however, can be misused. Counterfactual answers about public policy, historical conflicts, social institutions, disasters, or technological risks may be mistaken for factual predictions, policy advice, or authoritative alternate histories. WhatIfBench is therefore intended only as a diagnostic benchmark, not as a decision-making system. Its questions are collected from publicly accessible sources and normalized for evaluation; we filter out cases that require private or personally identifying information, lack a clear counterfactual premise, or cannot be assessed through an explanatory rubric. Because open-domain sources may still reflect topical, cultural, and English-centric biases, the benchmark should not be treated as exhaustive or culturally neutral. PRISM further relies on automatic parsing and judging of response-derived causal graphs, so its scores measure causal coherence, textual support, and rubric-level explanatory adequacy rather than ground-truth real-world causality. We release WhatIfBench and PRISM for research on counterfactual reasoning evaluation, and recommend that any high-stakes use be paired with human review, inspection of model responses and causal graphs, explicit documentation of source coverage, and safeguards against harmful speculation or unsupported policy conclusions.

References

Ishwar B Balappanawar, Vamshi Krishna Bonagiri, Anish R Joishy, Manas Gaur, Krishnaprasad Thirunarayan, and Ponnurangam Kumaraguru. 2025. If pigs could fly... can llms logically reason through counterfactuals? *arXiv preprint arXiv:2505.22318*.

Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. 2022. Causalqa: A benchmark for causal question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3296–3308.

Yuefei Chen, Vivek K Singh, Jing Ma, and Ruxiang Tang. 2025. Counterbench: A benchmark for counterfactuals reasoning in large language models. *arXiv preprint arXiv:2502.11008*.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2024. Unveiling causal reasoning in large language models: Reality or mirage? *Advances in Neural Information Processing Systems*, 37:96640–96670.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Jörg Frohberg and Frank Binder. 2022. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, and 168 others. 2026. Glm-5: from vibe coding to agentic engineering. *Preprint*, arXiv:2602.15763.

Paul He, Yinya Huang, Mrinmaya Sachan, and Zhijing Jin. 2026. Uncovering hidden correctness in llm causal reasoning via symbolic verification. *arXiv preprint arXiv:2601.21210*.

Yanjie He. 2026. Thinking fast, thinking wrong: Intuitiveness modulates llm counterfactual reasoning in policy evaluation. *arXiv preprint arXiv:2604.10511*.

Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. 2022. Wikiwhy: Answering and explaining cause-and-effect questions. *arXiv preprint arXiv:2210.12152*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and 1 others. 2023. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, David Ha, and 1 others. 2024. Can large language models infer causation from correlation? In *International Conference on Learning Representations*, volume 2024, pages 28663–28679.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*.

David Lewis. 1973. Causation. *The journal of philosophy*, 70(17):556–567.

William C Mann and Sandra A Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical report, University of Southern California, Information Sciences Institute Los Angeles.

Martina Miliani, Serena Auriemma, Alessandro Bondielli, Emmanuele Chersoni, Lucia Passaro, Irene Sucameli, and Alessandro Lenci. 2025. Explica: Evaluating explicit causal reasoning in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17335–17355.

Anh Quang Nguyen. 2025. Counterfactual history: Simulating chinese imperial decisions with ai. *ACTA ILOCANDIA: The DMMMSU International Science and Innovative Technology Journal (Formerly: DMMMSU Research and Extension Journal)*, 9(1):117–138.

Philip E Tetlock and Aaron Belkin. 1997. *Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives*. Princeton University Press.

Aniket Vashishtha, Qirun Dai, Hongyuan Mei, Amit Sharma, Chenhao Tan, and Hao Peng. 2025. Executable counterfactuals: Improving llms' causal reasoning through code. *arXiv preprint arXiv:2510.01539*.

Yuzhe Wang, Yaochen Zhu, and Jundong Li. 2026. Causalflip: A benchmark for llm causal judgment beyond semantic matching. *arXiv preprint arXiv:2602.20094*.

Zeyu Wang. 2024. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151.

James Woodward and James Francis Woodward. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.

Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. 2024. Causality for large language models. *arXiv preprint arXiv:2410.15319*.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. 2025. Cofca: A step-wise counterfactual multi-hop qa benchmark. In *International Conference on Learning Representations*, volume 2025, pages 14631–14649.

Khurram Yamin, Gaurav Ghosal, and Bryan Wilder. 2025. Can llms reconcile knowledge conflicts in counterfactual reasoning. *arXiv preprint arXiv:2506.15732*.

Khurram Yamin, Shantanu Gupta, Gaurav R Ghosal, Zachary C Lipton, and Bryan Wilder. 2024. Failure modes of llms for causal reasoning on narratives. *arXiv preprint arXiv:2410.23884*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Linying Yang, Vik Shirvaikar, Oscar Clivio, and Fabian Falck. 2024. A critical review of causal reasoning benchmarks for large language models. *arXiv preprint arXiv:2407.08029*.

Longxuan Yu, Delin Chen, Siheng Xiong, Qingyang Wu, Dawei Li, Zhikai Chen, Xiaoze Liu, and Liangming Pan. 2025. Causaleval: Towards better causal reasoning in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12512–12540.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8276–8288.

Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, and Chris Callison-Burch. 2023. Causal reasoning of entities and events in procedural texts. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 415–431.

Tianshi Zheng, Kelvin Kiu-Wai Tam, Newt Hue-Nam K Nguyen, Baixuan Xu, Zhaowei Wang, Jiayang Cheng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, and 1 others. 2025. Newtonbench: Benchmarking generalizable scientific law discovery in llm agents. *arXiv preprint arXiv:2510.07172*.

Appendix

In this appendix, we provide supplementary materials for **WhatIfBench** and **PRISM** to improve the transparency and reproducibility of our benchmark and evaluation framework. We first present additional dataset statistics in Appendix A, including the overall scenario taxonomy, fine-grained category distributions, and lexical characteristics of WhatIfBench. We then describe the data collection, annotation, and quality-control process in Appendix B, covering candidate sources, filtering criteria, rubric construction guidelines, and review protocols. Appendices C and D provides the complete prompt templates used for model response generation, response-to-graph parsing, Process Metric evaluation, and Rubric Metric evaluation, together with decoding and evaluator settings. Appendix E further details the response-to-graph parsing procedure and presents the parser pseudocode. Appendix F reports additional experiments on benchmark positioning, model scale, response length, paraphrase robustness, prompting strategies, and evaluator-family bias. Finally, Appendix G gives a complete worked example of PRISM evaluation, including a what-if question with its query-dependent rubrics, a model response, the derived semantic causal graph, and RM judgment.

A Dataset Details and Statistics

A.1 Scenario Taxonomy

WhatIfBench organizes open-form counterfactual questions into three scenario types: STEM, HSS, and Hybrid. STEM scenarios focus on scientific, technical, physical, biological, and engineering mechanisms, where counterfactual consequences are often constrained by relatively stable domain principles. HSS scenarios cover historical, political, social, cultural, and institutional dynamics, where reasoning must account for path dependence, multi-agent behavior, and long-term social feedback. Hybrid scenarios involve cross-domain propagation, where changes in natural, technical, or material conditions further affect social, economic, or institutional systems. As shown in Figure 4, STEM constitutes the largest portion of WhatIfBench, followed by HSS and Hybrid, reflecting our emphasis on both mechanistic reasoning and socio-historical causal propagation.

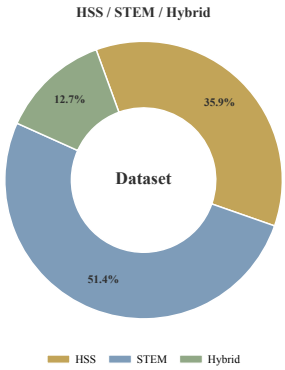

Figure 4: **Scenario distribution of WhatIfBench.** The benchmark contains 220 open-form what-if questions across STEM, HSS, and Hybrid scenarios. STEM accounts for 51.4% of the dataset, followed by HSS at 35.9% and Hybrid at 12.7%.

A.2 Fine-grained Category Distribution

We further group STEM and HSS questions into fine-grained subcategories to characterize the topical coverage of WhatIfBench. For HSS, the largest group is State, Regime, and Macrohistory, followed by War and Geopolitics, while the remaining questions cover dynastic leadership, domestic politics and institutions, and culture, technology, and society. For STEM, Physical and Planetary Systems is the largest group, followed by Engineering, Transport, and Design, Society, Computation, and Logistics, and Life Sciences and Human Body. These subcategories are not intended as exhaustive domain taxonomies; rather, they provide diagnostic slices for analyzing whether models fail more often on physical mechanisms, biological constraints, institutional dynamics, geopolitical reasoning, or cross-system dependencies.

A.3 Lexical Overview

To provide an intuitive view of the topical coverage of WhatIfBench, we visualize the lexical distribution of normalized questions in Figure 7. The most frequent terms reflect the benchmark's mixture of scientific, historical, geopolitical, and technological scenarios, including planetary systems, empires, wars, societies, and alternative historical trajectories. This lexical overview is intended as a qualitative summary rather than a formal diversity measure; the structured scenario taxonomy and fine-grained category distributions provide the primary basis for dataset characterization.

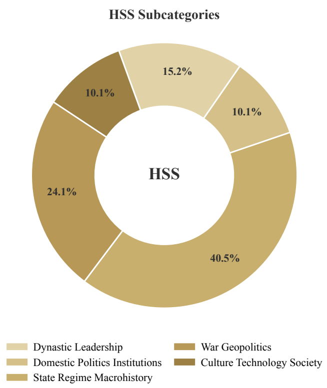

Figure 5: **Fine-grained distribution of HSS questions.** HSS scenarios are divided into State, Regime, and Macrohistory; War and Geopolitics; Dynastic Leadership; Domestic Politics and Institutions; and Culture, Technology, and Society. State, Regime, and Macrohistory is the largest HSS subcategory at 40.5%.

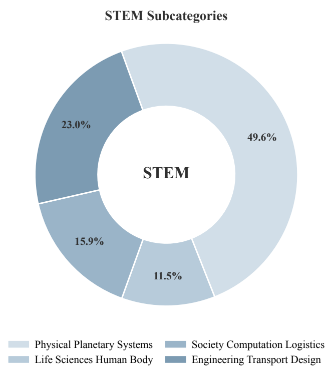

Figure 6: **Fine-grained distribution of STEM questions.** STEM scenarios are divided into Physical and Planetary Systems, Life Sciences and Human Body, Society, Computation, and Logistics, and Engineering, Transport, and Design. Physical and Planetary Systems forms the largest STEM subcategory at 49.6%.

Figure 7: **Lexical overview of WhatIfBench questions.** The word cloud visualizes frequent terms in normalized what-if questions, offering a qualitative view of the benchmark's topical coverage across scientific, historical, geopolitical, technological, and cross-domain scenarios.

B Data Collection, Annotation, and Quality Control

B.1 Candidate Collection

We construct WhatIfBench from publicly accessible what-if questions and explanatory discussions. Candidate questions are collected from public what-if QA sites, online discussion forums, and explanatory materials, including xkcd What If, Worldbuilding Stack Exchange, AlternateHistory, and Quora. These sources contain naturally occurring counterfactual questions that often require open-form explanations rather than short factual answers. During collection, we retain the original counterfactual intent of each question while normalizing its wording into a clear query format suitable for model evaluation.

B.2 Question Filtering Criteria

We filter candidate questions according to four criteria. First, each question must contain an explicit counterfactual premise or intervention. Second, the question should require nontrivial causal propagation rather than simple factual lookup or direct answer retrieval. Third, the question should admit open-form explanation, allowing multiple plausible downstream trajectories rather than a single fixed outcome. Fourth, the question should provide sufficient contextual grounding for constructing a meaningful rubric and evaluating the response. We reject questions that lack a clear counterfactual premise, depend primarily on private or personally identifying information, are too underspecified for rubric-based evaluation, or cannot be assessed through explanatory criteria.

B.3 Rubric Construction Guidelines

For each retained question, we construct a query-specific rubric to specify the reasoning requirements expected of a high-quality counterfactual explanation. Each rubric is designed to cover the key dimensions used by PRISM, including fidelity to the counterfactual premise, coverage of central mechanisms, treatment of downstream consequences, respect for domain constraints, uncertainty handling, and overall explanatory completeness. The rubric does not prescribe a unique conclusion. Instead, it defines question-dependent requirements that different plausible answers should satisfy, allowing evaluation to focus on whether the response supports its trajectory with coherent causal reasoning.

Benchmark	Open-domain	Open-form	Long-horizon	Process-level Evaluation	Question-specific Rubrics
CRASS (Frohberg and Binder, 2022)	×	×	×	×	×
IfQA (Yu et al., 2023)	✓	×	×	×	×
CLadder (Jin et al., 2023)	×	×	×	×	×
CofCA (Wu et al., 2025)	×	×	×	✓	×
CounterBench (Chen et al., 2025)	×	×	✓	×	×
WhatIfBench	✓	✓	✓	✓	✓

Table 6: **Comparison of counterfactual reasoning benchmark characteristics.**

Model	WhatIfBench		IfQA		CounterBench	
	Final ↑	Rank	EM ↑	Rank	Acc. ↑	Rank
GLM-5.1	51.99	6	89.71	6	80.00	3
DeepSeek-V4-Pro	55.12	4	90.43	4	85.71	1
Qwen3-Max	52.26	5	93.00	1	76.74	6
Gemini-3.1-Pro-Preview	56.33	3	90.00	5	78.33	5
Claude-Opus-4.7	59.11	2	91.86	3	79.97	4
GPT-5.5	64.62	1	92.57	2	83.57	2

Table 7: **Performance and rankings of frontier LLMs across counterfactual reasoning benchmarks.**

B.4 Quality Control Protocol

We conduct quality control at both the question and rubric levels. For questions, review checks whether the normalized query has a clear counterfactual premise, sufficient contextual grounding, nontrivial causal propagation requirements, and an evaluable explanatory structure. For rubrics, review checks whether the criteria cover the core reasoning space without forcing a single gold trajectory. Each item is assigned one of three outcomes: PASS, if it can be used directly; REVISED, if it becomes valid after normalization or rubric refinement; and REJECTED, if it does not meet the benchmark requirements. As reported in the main text, 36.0% of 611 candidate questions are retained after direct pass or revision, yielding 220 final questions, and all corresponding rubric sets pass after direct review or revision.

C Decoding and Evaluator Settings

To ensure consistent text generation and rigorous evaluation within the **PRISM** framework, we establish a standardized set of decoding hyperparameters and structural parsing rules across our entire pipeline.

For the initial *Model Response Generation* phase on **WhatIfBench**, models generate counterfactual answers using a sampling temperature of $\tau = 0.6$ and a maximum budget of $4,096$ tokens to encourage natural, comprehensive, and descriptive long-range causal explanations. Conversely, for all downstream parsing and judgment tasks, we enforce a strictly deterministic decoding policy by setting the temperature to $\tau = 0.0$ across all evaluation API calls. The maximum generation budget varies depending on the structural density required by each core component: the *Response-to-Graph Parser* is allocated up to $16,384$ tokens to accommodate construction of a full *Response-Derived Semantic Causal Graph*, whereas the *Process Metric (PM) Edge Evaluator* and *Rubric Metric (RM) Evaluator* are capped at $1,024$ and $4,096$ tokens, respectively.

D Prompt Templates

For complete reproducibility, we detail the prompt templates (figs. 8 to 11) used for model response generation, graph parsing, and dual-metric evaluation within the **PRISM** framework.

E Response-to-Graph Parsing Algorithm

Algorithm 1 summarizes how we convert each model response y into a response-derived semantic causal graph $\mathcal{G}_y = (\mathcal{V}_y, \mathcal{E}_y)$. The parser first extracts elementary discourse units (EDUs) and semantic relation records with labels from the predefined set $\mathcal{L}$. These records are normalized and organized into a compact support–core structure. We then project the structure onto EDU-level graph nodes, where each resolved relation becomes a directed edge after duplicate removal. If parsing fails or no valid EDU is extracted, we use a conservative heuristic fallback that splits the response into

Algorithm 1 Response-to-Graph Parsing

Require: Model response y, parser model p_θ, relation labels $\mathcal{L}$, budget B
Ensure: Response-Derived Semantic Causal Graph $\mathcal{G}_y = (\mathcal{V}_y, \mathcal{E}_y)$

Construct prompt $m(y, \mathcal{L})$ for extracting EDUs, semantic relations, and support–core structure.
Obtain parser output $o \leftarrow p_\theta(m(y, \mathcal{L}); \tau = 0, B)$.
if o is empty **then**
    return HEURISTICGRAPH(y)
end if
Parse EDUs $\mathcal{U} = \{u_i\}_{i=1}^n$ from o.
if $\mathcal{U} = \emptyset$ **then**
    return HEURISTICGRAPH(y)
end if
Parse semantic relations $\mathcal{R} = \{(s_j, t_j, \ell_j, \nu_j)\}_{j=1}^m$, where $\ell_j \in \mathcal{L}$ and $\nu_j \in \{\text{SN}, \text{NS}, \text{NN}\}$.
if $\mathcal{R} \neq \emptyset$ **then**
    Build structure nodes $\mathcal{T} \leftarrow$ BUILDSTRUCTURE($\mathcal{R}$).
else if $|\mathcal{U}| > 1$ **then**
    Build a flat fallback structure $\mathcal{T} \leftarrow$ BUILDFLAT($\mathcal{U}$, SEQUENCE).
else
    $\mathcal{T} \leftarrow \emptyset$.
end if
Initialize graph nodes $\mathcal{V}_y \leftarrow \{(u_i, \text{text}(u_i)) \mid u_i \in \mathcal{U}\}$ and edges $\mathcal{E}_y \leftarrow \emptyset$.
Build reference map $\mathcal{M}$ from EDU spans and structure nodes to their resolved EDUs.
for all structure node $z \in \mathcal{T}$ **do**
    Normalize relation label $\ell \leftarrow$ NORMALIZE(z.relation).
    Resolve source EDUs $\mathcal{S}$ and target EDUs $\mathcal{D}$ from z using $\mathcal{M}$.
    for all $u_s \in \mathcal{S}, u_t \in \mathcal{D}$ **do**
        if $u_s \neq u_t$ and (u_s, u_t, ℓ) is not duplicated **then**
            Add edge (u_s, u_t, ℓ) to $\mathcal{E}_y$.
        end if
    end for
end for
return $\mathcal{G}_y = (\mathcal{V}_y, \mathcal{E}_y)$
function HEURISTICGRAPH(y)
    Split y into sentence-level EDUs and connect adjacent EDUs with SEQUENCE edges.
    return the resulting graph.
end function

sentence-level EDUs and connects adjacent units with SEQUENCE edges. This keeps the graph construction grounded in the response text while ensuring that all responses can be evaluated under a consistent graph format.

F Additional Experiments

We conduct additional experiments to examine the positioning and robustness of WhatIfBench and PRISM, including comparisons with existing benchmarks, smaller-model evaluation, response-length control, paraphrase stress testing, counterfactual-specific prompting, and cross-family evaluation.

Comparison with existing counterfactual benchmarks. Table 6 compares WhatIfBench with representative counterfactual reasoning benchmarks. Existing benchmarks mainly use constrained answer spaces or fixed outcomes, whereas WhatIfBench jointly targets open-domain, open-form, long-horizon reasoning with process-level evaluation and question-specific rubrics.

We further evaluate the same six frontier models on IfQA and CounterBench. As shown in Table 7, performance is substantially higher and more compressed on the closed-form benchmarks, while model rankings differ considerably across tasks.

Qwen3-Max ranks first on IfQA but fifth on WhatIfBench, while DeepSeek-V4-Pro ranks first on CounterBench but fourth on WhatIfBench. As an external comparison, WhatIfBench also shows stronger rank agreement with the Artificial Analysis Intelligence Index[5] ($\rho = 0.94$) than IfQA ($\rho = 0.09$) or CounterBench ($\rho = 0.37$), suggesting that these benchmarks capture complementary capabilities.

Evaluation on smaller models. To examine discriminability beyond frontier models, we evaluate four smaller models spanning dense and MoE architectures. Their final scores range from 40.80 to 44.36, below the frontier-model range of 51.99–64.62, showing that WhatIfBench remains discriminative across model scales and architectures.

Effect of response length. PRISM does not explicitly reward response length, but longer outputs may introduce more causal content. We therefore evaluate GPT-5.5 under Vanilla, Long Output, and Short Output settings. The final-score standard deviation is only 0.48, indicating limited sensitivity to moderate changes in response length.

Paraphrase-only stress test. We test whether simply restating the counterfactual question can receive substantial PRISM credit. Paraphrase-only responses obtain 2.05 PM and 6.04 Final. The RM value near 10 corresponds to the minimum score of 1 on the 1–10 rubric scale, confirming that surface reformulation alone receives negligible credit.

Counterfactual-specific prompting. The main experiments use a standardized prompt for controlled comparison. We additionally evaluate GPT-5.5 with two counterfactual-specific strategies, CausalCoT and CoIn. CausalCoT and CoIn improve the final score by 2.63 and 1.35 points, respectively. The gains are moderate and affect PM and RM dif-

[5] https://artificialanalysis.ai/evaluations/artificial-analysis-intelligence-index

Model	Architecture	STEM		HSS		Hybrid		Overall		Final
		PM	RM	PM	RM	PM	RM	PM	RM	
Qwen3-4B	Dense	48.87	36.18	45.72	32.88	39.16	36.80	46.59	35.02	40.80
Qwen3-8B	Dense	53.08	34.48	42.49	32.54	55.10	37.20	49.36	34.06	41.71
Gemma-3-12B-it	Dense	49.45	41.73	43.92	41.61	44.72	43.28	46.85	41.86	44.36
GPT-oss-20B	MoE	52.47	43.37	42.76	36.54	46.26	38.64	48.20	40.41	44.26

Table 8: **Performance of smaller models on WhatIfBench across model architectures.**

Method	STEM		HSS		Hybrid		Overall		Final
	PM	RM	PM	RM	PM	RM	PM	RM	
Vanilla	72.77	64.00	57.38	61.98	65.32	64.80	65.94	63.31	64.62
Long Output	73.17	64.42	57.13	61.87	67.49	65.60	66.69	63.73	65.21
Short Output	73.31	64.30	58.35	62.46	69.65	65.20	67.47	63.69	65.58
SD	0.28	0.22	0.64	0.31	2.17	0.40	0.77	0.23	0.48

Table 9: **Effect of response length on PRISM scores.** SD is computed across the three response-length settings.

ferently, indicating that prompting helps but does not eliminate the remaining difficulty. Under the prompting strategies evaluated here, WhatIfBench remains challenging for frontier models.

Robustness to evaluator family. To test potential same-family evaluator bias, we re-evaluate all six frontier models using Claude-Sonnet-5 under the same PRISM protocol. The two evaluators produce identical model rankings (Spearman $\rho = 1.00$). The mean absolute difference in final score is 0.63 points, with a maximum of 1.57 points, suggesting that the main comparative conclusions are robust to the evaluator family.

G Worked Example of PRISM Evaluation

For case-level transparency, we provide a worked example of **PRISM** evaluation on one WhatIfBench question. Figure 12 presents the counterfactual question and its query-dependent rubric, while fig. 13 shows the generated model response. PRISM then parses the response into a response-derived semantic causal graph, visualized in fig. 14, where related EDUs are grouped into causal modules for readability. Figure 15 reports the RM judgment with criterion-level scores and rationales. The example shows how PRISM jointly exposes the response's expressed reasoning structure and its answer-level omissions: although the response contains plausible local causal links, it receives a moderate RM score because the point of divergence, institutional integration pathway, and uncertainty control remain underdeveloped.

Method	STEM		HSS		Hybrid		Overall		Final
	PM	RM	PM	RM	PM	RM	PM	RM	
Paraphrase	2.65	10.00	1.83	10.10	0.00	10.00	2.05	10.04	6.04

Table 10: **PRISM stress test with paraphrase-only responses.**

Method	STEM		HSS		Hybrid		Overall		Final
	PM	RM	PM	RM	PM	RM	PM	RM	
Vanilla	72.77	64.00	57.38	61.98	65.32	64.80	65.94	63.31	64.62
CausalCoT	63.98	68.34	68.32	67.88	68.78	70.00	66.12	68.37	67.25
CoIn	69.16	62.63	69.60	61.15	73.65	64.84	69.61	62.33	65.97

Table 11: **Effect of counterfactual-specific prompting strategies on GPT-5.5.**

Model	STEM		HSS		Hybrid		Overall		Final	Rank
	PM	RM	PM	RM	PM	RM	PM	RM		
GLM-5.1	59.59	52.19	48.64	47.00	53.95	50.32	54.81	49.99	52.40	6
DeepSeek-V4-Pro	65.75	55.50	50.91	52.94	42.05	54.75	57.51	54.42	55.97	4
Qwen3-Max	67.75	48.27	53.99	43.00	56.85	48.56	61.33	46.34	53.83	5
Gemini-3.1-Pro-Preview	67.89	54.18	51.57	50.32	50.89	50.24	59.83	52.29	56.06	3
Claude-Opus-4.7	68.88	56.11	56.89	52.37	55.57	58.88	62.87	54.94	58.90	2
GPT-5.5	73.51	64.10	57.85	62.02	63.20	65.68	66.57	63.55	65.06	1

Table 12: **PRISM scores and model rankings using Claude-Sonnet-5 as the evaluator.**

Model Response Generation Prompt

System Prompt:

You answer counterfactual benchmark questions. Give a direct, coherent, non-bulleted answer unless bullets are clearly helpful.

User Prompt:

Question domain: {domain}
Question type: {answer_type}

Question:
{question}

Answer the question carefully and directly.

Figure 8: Model response generation prompt used for WhatIfBench evaluation.

Response-to-Graph Parser Prompt

System Prompt:

Return only the requested sections. Do not judge the answer or add unstated causal claims.

User Prompt:

You are a discourse parser for response-derived semantic causal graph construction.

Analyze the model response under the given question and counterfactual premise. Use an RST-inspired procedure: first segment the text into meaningful elementary discourse units (EDUs), then identify core vs supporting structure, assign semantic discourse relation labels, and finally write a compact structure that preserves the response-expressed reasoning process.

The goal is not to judge whether the response is factually correct, but to expose the events, states, mechanisms, constraints, and causal dependencies expressed in the response for downstream evaluation. Do not add causal steps, mechanisms, or consequences that are not stated or unambiguously paraphrased in the text.

Use only these relation labels when possible:
{relation_list}

Prioritize causal-relevant labels such as CAUSE, RESULT, and CONSEQUENCE when the text explicitly expresses causal, conditional, resultative, or downstream dependencies. Use non-causal discourse labels such as ELABORATION, CONTRAST, TEMPORAL, and BACKGROUND when the relation is rhetorical, temporal, contrastive, or contextual rather than causal.

Output exactly these sections and no extra commentary:
EDUs:
ANALYSIS:
STRUCTURE:

Formatting rules:
- In EDUs, use one line per EDU in the form 'EDU1: ...'.
- Each EDU should be an atomic claim, preferably an event, state, mechanism, constraint, assumption, or downstream consequence relevant to the counterfactual explanation.
- In ANALYSIS, use one line per relation in the form 'RELATION(EDU_source, EDU_target): RELATION_TYPE [NUCLEARITY]'.
- For causal-relevant relations, EDU_source should be the causal or conditional antecedent, and EDU_target should be the effect, result, or downstream consequence whenever the text makes this direction clear.
- For asymmetric non-causal relations, EDU_source must be the supporting/satellite unit and EDU_target must be the core/nucleus unit.
- Use nuclearity tags 'SN', 'NS', or 'NN'.
- In STRUCTURE, write a compact textual structure using ROOT / NUCLEUS / SATELLITE / MULTINUCLEAR labels, while preserving the main counterfactual reasoning trajectory where possible.

Question:
{question}

Text:
{answer}

Figure 9: Response-to-Graph Parser Prompt used to construct response-derived semantic causal graphs.

Process Metric Edge Evaluation Prompt

You are an adversarially strict judge of one extracted discourse edge in a counterfactual answer.
Answer with only one token: yes or no.

Judge only whether the full answer explicitly supports this exact directed relation:
source claim → target claim with relation type = {relation_type}.

Default to no. Say yes only when the answer contains enough local textual evidence for this exact edge.

Strict decision rules:

- Say yes only if both claims are explicitly stated or unambiguously paraphrased in nearby parts of the answer.
- Say yes only if the answer itself makes the source-to-target link clear; the link must not require outside knowledge, reconstruction, or charitable interpretation.
- Say yes only if the relation type is appropriate at the requested strength. For CAUSE / RESULT / CONSEQUENCE, the answer must express an actual causal or consequential dependency, not just sequence or topic association.
- Say no if the answer merely makes the edge plausible, mentions both claims separately, or leaves the relation implicit.
- Say no if the edge is supported only by the question, historical/common knowledge, or what would normally be true in the domain.
- Say no if directionality is ambiguous, reversed, circular, too broad, too specific, or partially contradicted.
- Say no if either claim is vague in the extracted edge and the answer does not pin it down clearly.
- Do not reward fluent prose, answer length, topical relevance, or broad scenario coherence.

Inputs for Evaluation:

Question:
{question}

Source claim:
{source}

Target claim:
{target}

Relation type:
{relation_type}

Full answer to inspect:
{answer}

Figure 10: Unified judge prompt template used for Process Metric (PM) edge evaluation within the PRISM framework.

Rubric Metric Evaluation Prompt

System Prompt:

Return JSON only.

User Prompt:

You are a strict rubric judge for counterfactual long-range reasoning.
Return JSON only with keys criterion_scores and rationale.

Score each rubric criterion independently from 1 to 10. The list criterion_scores must contain one integer score for each rubric criterion, in the same order as the rubric. Judge only the answer against the given question and rubric.

Use this strict grading scale for each criterion:
- 9–10: very strong satisfaction; detailed and criterion-specific, with only minor omissions.
- 7–8: good but incomplete; solid reasoning with noticeable missing mechanisms, constraints, or downstream effects.
- 5–6: middling; broadly relevant but shallow, generic, or uneven.
- 3–4: weak; thin, mostly descriptive, under-justified, or loosely tied to the premise.
- 1–2: poor; wrong, contradictory, mostly unsupported, or fails the criterion.

Rules:
- Grade evidence actually present in the answer, not what a knowledgeable reader could infer.
- Penalize generic alternate-history summaries, unsupported leaps, missing long-range consequences, and lack of mechanism detail.
- Penalize confident claims that ignore uncertainty or historical/domain constraints.
- Do not give high criterion scores for fluency, length, or topical relevance alone.
- If uncertain between two scores, choose the lower one.
- A merely plausible answer is not a strong answer; reward explicit reasoning, not plausibility.

Hard caps:
- If the answer does not directly address the exact counterfactual premise, the average criterion score must be at most 4.
- If the answer is mostly generic background or a high-level summary, the average criterion score must be at most 5.
- If the answer gives conclusions without explaining mechanisms, the average criterion score must be at most 6.
- If it misses one major rubric dimension, the average criterion score must be at most 7.
- If it contains a serious contradiction, premise drift, or major factual/domain error, the average criterion score must be at most 6.
- If it is overconfident about speculative branches without qualification, the average criterion score must be at most 8.
- Give an average score of 9 or above only when the answer has concrete, question-specific mechanisms, multiple downstream consequences, constraint awareness, and no major omissions.

Inputs for Evaluation:

Question:
{question}

Rubric:
{rubric}

Answer:
{answer}

Figure 11: Unified judge prompt template used for Rubric Metric (RM) evaluation within the PRISM framework.

Question and Query-Dependent Rubric

Question:

What If Female US Combat Pilots in WWII?

Rubric Criteria

1. Plausible Point of Divergence for Women in US Combat Aviation

Description. Assesses whether the answer identifies a credible historical change that could allow female US combat pilots in WWII, especially by addressing 19th- and early-20th-century gender norms, aviation culture, and military policy barriers.

1–2: Provides no clear point of divergence or relies on an implausible sudden policy reversal with no background change.

3–4: Names a divergence but treats acceptance of female combat pilots as easy or inevitable despite known cultural resistance.

5–6: Offers a plausible divergence, such as earlier normalization of women in aviation or combat roles, but only lightly explains why it changes WWII policy.

7–8: Connects a credible divergence to changing gender norms, military recruitment standards, and prewar aviation institutions.

9–10: Builds a highly plausible divergence rooted in US social history, showing how altered gender ideology, aviation precedent, and institutional incentives make female combat pilots possible by WWII.

2. Military Institutional Mechanisms and Integration Pathway

Description. Assesses whether the answer explains how the US Army Air Forces, Navy, training pipeline, command structure, and combat units would actually incorporate female pilots rather than merely asserting their presence.

1–2: Simply states that women fly in combat with no explanation of recruitment, training, assignment, or command acceptance.

3–4: Mentions training or units in passing but ignores major institutional obstacles such as policy, logistics, or officer resistance.

5–6: Describes a basic pathway, such as expanding WASP-like roles into combat, but leaves gaps in rank, deployment, and unit organization.

7–8: Explains recruitment, screening, training, rank/status, unit placement, and likely compromises such as elite exceptions or all-female squadrons.

9–10: Gives a detailed, historically grounded integration pathway, including bureaucratic opposition, manpower pressures, combat validation, casualty handling, and differences across theaters or services.

3. Operational and Strategic Effects in WWII

Description. Assesses whether the answer plausibly evaluates what female US combat pilots would change militarily, including likely scale, mission types, performance, propaganda value, and limits on battlefield impact.

1–2: Claims sweeping victory changes or total irrelevance without reasoning.

3–4: Gives vague effects, such as "more pilots," without considering scale, mission suitability, or wartime constraints.

5–6: Identifies some likely roles, such as ferrying, transport, reconnaissance, fighters, or bombers, but weakly estimates their significance.

7–8: Plausibly distinguishes limited versus large-scale integration and explains effects on pilot shortages, replacement rates, morale, and public perception.

9–10: Carefully models operational consequences by theater and timeframe, balancing added pilot capacity and symbolic impact against training bottlenecks, aircraft supply, combat attrition, and command prejudice.

4. Domestic Political, Cultural, and Postwar Consequences

Description. Assesses whether the answer covers downstream effects on US gender norms, veterans' rights, labor, feminism, media narratives, and backlash after WWII.

1–2: Ignores domestic consequences beyond the war.

3–4: Mentions that society would become more equal or more resistant, but without causal detail.

5–6: Discusses some effects on gender roles, employment, or veterans' status, but treats them as linear and uncontested.

7–8: Explains both empowerment and backlash, including effects on women veterans, postwar employment, family ideology, and public memory.

9–10: Develops a nuanced long-range chain from wartime service to postwar policy, civil rights comparisons, feminist movements, military integration debates, and contested cultural narratives.

5. Counterfactual Discipline, Uncertainty, and Contradiction Control

Description. Assesses whether the answer maintains internal consistency with the premise, avoids anachronism, calibrates uncertainty, and compares relevant analogues such as Soviet women pilots, WASP, and other contested US military integrations.

1–2: Contains major contradictions, anachronisms, or ignores the counterfactual premise.

3–4: Has some plausible claims but also unsupported leaps, present-day assumptions, or inconsistent treatment of US WWII institutions.

5–6: Mostly coherent, with some uncertainty language, but misses key analogues or leaves causal jumps underexplained.

7–8: Maintains a coherent causal chain, uses relevant historical comparisons, and distinguishes likely, possible, and unlikely outcomes.

9–10: Shows rigorous counterfactual control: no major contradictions, strong use of analogues, explicit uncertainty bounds, and careful separation of premise-driven changes from unchanged historical constraints.

Figure 12: Example WhatIfBench question and its query-dependent rubric.

Model Response

Question:

What If Female US Combat Pilots in WWII?

Model Response:

If the United States had allowed women to serve as combat pilots in World War II, the most likely result would have been a modest military benefit and a much larger social and political impact.

Militarily, women pilots could have helped ease pilot shortages, especially in 1942–44 when the U.S. military was expanding rapidly. The U.S. already had skilled women aviators in the Women Airforce Service Pilots program, or WASP, who ferried aircraft, towed targets, tested planes, and trained male pilots, though they were barred from combat and were not fully militarized during the war. If even a portion of these women had been trained and assigned to combat roles, they might have flown transport, reconnaissance, antisubmarine patrols, and eventually fighters or bombers.

The immediate effect on the war's outcome would probably not have been decisive. The U.S. won through overwhelming industrial production, logistics, naval power, air superiority, and the combined effort of the Allies. Adding women combat pilots would not by itself have shortened the war dramatically. But it could have expanded the pool of available pilots, freed more men for other combat roles, and improved flexibility in theaters where pilot fatigue and replacement needs were serious issues.

The largest obstacles would have been institutional and cultural rather than technical. Many women were fully capable of flying military aircraft, as the WASP program demonstrated. The resistance would have come from military leadership, Congress, parts of the public, and male servicemen who believed combat was inherently male. If women pilots had flown in combat and performed well, those assumptions would have been challenged much earlier.

There would also have been controversy if women pilots were killed, captured, or publicly celebrated as combat heroes. Female casualties in combat would have forced Americans to confront questions about women's citizenship, sacrifice, and equality decades earlier than they actually did. Some families and politicians would likely have opposed the policy fiercely, while others would have used women's service as proof that women deserved full military status, veterans' benefits, and broader civil rights.

After the war, the consequences could have been significant. Women veterans with combat records would have made it harder to dismantle women's military roles, as happened with the WASP program in 1944. The campaign for women's integration into the armed forces, equal pay, veterans' recognition, and later access to military academies and combat arms might have advanced earlier. It could also have strengthened the postwar women's rights movement by giving it a powerful patriotic argument: women had fought and died for the country on equal terms.

So, in short: female U.S. combat pilots in World War II probably would not have changed the final Allied victory, but they could have improved wartime manpower flexibility and would likely have transformed the debate over women in the military. Their success would have made it much harder after 1945 to argue that women were unfit for combat or full military citizenship.

Figure 13: Model response generated for the selected WhatIfBench case.

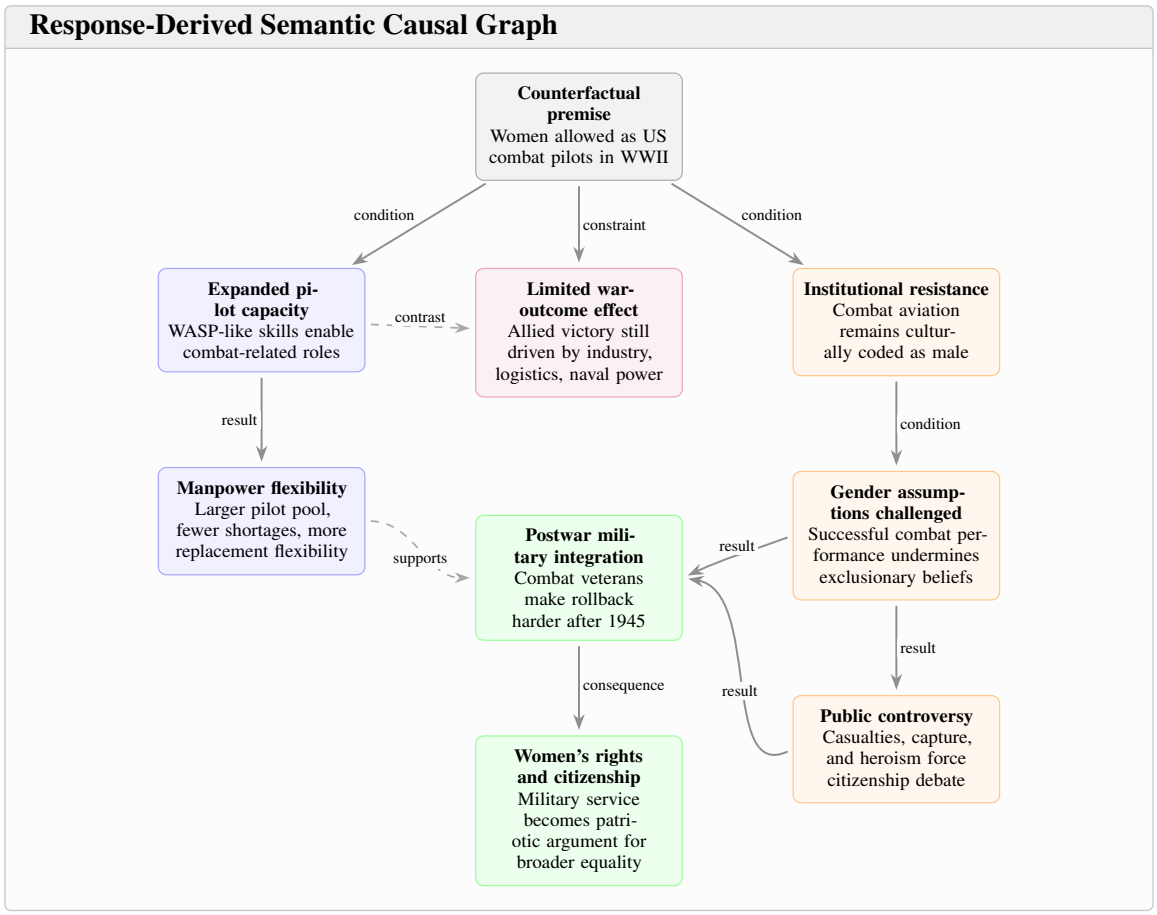

Figure 14: Abstracted response-derived semantic causal graph for the selected case. Nodes represent interpretable causal modules, while arrows show the main causal and discourse relations expressed in the response.

Rubric Metric Judgment

Overall: RM $= 0.50$; raw rubric score $= 5.0/10$; criterion scores $= [3, 5, 6, 6, 5]$.

Criterion 1: Plausible Point of Divergence for Women in US Combat Aviation. Score: 3. The answer is relevant to women as WWII combat pilots, but it does not actually supply a point of divergence. It mostly discusses likely effects if women had been allowed into combat, not the prior historical change that would make that policy credible. It mentions institutional and cultural resistance and cites WASP as evidence of capability, but it does not trace a concrete prewar shift in gender norms, aviation culture, or military policy that overcomes those barriers. Because the criterion is specifically about identifying a plausible divergence and explaining the mechanism, this remains thin and generic rather than a credible, historically rooted divergence.

Criterion 2: Military Institutional Mechanisms and Integration Pathway. Score: 5. The answer engages the criterion at a basic level by citing WASP as a potential feeder, noting pilot shortages, and identifying institutional/cultural resistance from military leadership, Congress, and male servicemen. It also gestures toward possible combat assignments and the issue of women being barred from combat and not fully militarized. However, it does not explain a concrete integration pathway through the Army Air Forces or Navy: no recruitment/screening mechanism, no training modifications, no rank/status solution, no unit organization, no command structure, no deployment process, and no explanation of how women would be assigned into existing combat units versus separate squadrons. It is therefore relevant but shallow, fitting the 5–6 band, and the omissions of central elements named in the criterion keep it at the low end.

Criterion 3: Operational and Strategic Effects in WWII. Score: 6. The answer gives a plausible, relevant military assessment: it identifies likely roles, such as transport, reconnaissance, antisubmarine patrols, and possibly fighters or bombers; argues the effect would be modest rather than war-winning; and notes constraints from institutional resistance. It also mentions pilot shortages, replacement/flexibility, and propaganda/public perception effects. However, it does not model operational consequences in much detail by theater or timeframe, does not estimate likely scale beyond vague phrases like "even a portion," and only lightly addresses limits such as training bottlenecks, aircraft availability, combat attrition, or command prejudice in operational terms. Because the military effects are treated competently but still generically and unevenly, this fits the 5–6 band, at the top of that range rather than 7–8.

Criterion 4: Domestic Political, Cultural, and Postwar Consequences. Score: 6. The answer directly engages domestic/postwar consequences and gives some causal mechanisms: women's combat service would challenge gender assumptions, create controversy over casualties and heroism, strengthen claims for full military status and veterans' benefits, and potentially advance later integration and women's rights arguments. That is more than a generic equality claim. However, it remains fairly linear and incomplete. It does not explicitly develop both empowerment and backlash after the war in any depth, says little about postwar employment, family ideology, media/public memory beyond brief mention of celebration, labor effects, or contested cultural narratives, and it lacks the longer-range chain to civil rights comparisons, feminist movements, and military integration debates required for 9–10. Because it misses several central elements named in the criterion and does not fully elaborate downstream domestic consequences, it fits 5–6, at the top of that band rather than 7–8.

Criterion 5: Counterfactual Discipline, Uncertainty, and Contradiction Control. Score: 5. The answer is coherent and does use some uncertainty language, such as "most likely," "probably," and "might," and it avoids obvious contradictions or anachronisms. It also grounds the premise somewhat with WASP and distinguishes military from social effects. However, for this criterion it remains generic and undercontrolled: it does not compare key analogues explicitly named in the criterion such as Soviet women pilots or other contested U.S. military integrations, and its causal chain is only lightly developed. Several claims are plausible but broad and insufficiently bounded, especially the downstream effects on postwar integration and the women's rights movement. It also blurs role categories by grouping transport and antisubmarine patrols with "combat roles" without carefully separating premise-driven changes from existing institutional constraints. Because it is mostly coherent but misses central analogues and leaves important causal jumps underexplained, it fits 5–6, and the lower score is appropriate.

Figure 15: Rubric-based judgment for the selected model response under the RM component of PRISM.