MDC-Bench: A Multidisciplinary Causal Benchmark Based on Causal Structures for Evaluating Large Language Models

Peng Wang[1], Yuxiong Yan[1], Xiao Ding[1]*, Kai Xiong[1], Bibo Cai[1], Chao Peng[2], Yutai Hou[2], Dandan Tu[2], Bing Qin[1], Ting Liu[1]

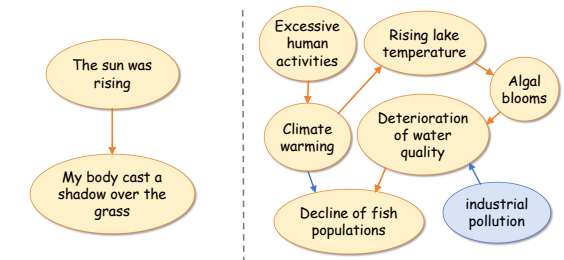
[1]Research Center for Social Computing and Interactive Robotics,
Harbin Institute of Technology, China
[2]Huawei Technologies Co., Ltd
{pengwang,yxyan,xding,kxiong,bbcai,qinb,tliu}@ir.hit.edu.cn
{pengchao28,houyutai,tudandan}@huawei.com

Abstract

Existing causal datasets primarily focus on the commonsense domain, where the questions mainly involve simple, single-hop direct causal relationships. When models possess the corresponding knowledge, even if they cannot understand the causal relationships, they can directly arrive at the correct answers through knowledge matching. However, LLMs often perform poorly when answering questions with complex causal structures and domain-specific expertise. To address the above challenges, we propose MDC-Bench, a multidisciplinary causal evaluation benchmark. MDC-Bench adopts a three-level causal framework consisting of 4 core causal tasks, while its sample content covers 7 representative disciplines and diverse causal structures. In view of the limited coverage of multidisciplinary knowledge during the pre-training phase, the model cannot answer questions relying on knowledge matching. The diverse causal structures force the models to grasp the internal causal logic. We also increase the task complexity through methods such as compound causal operations, aiming to enhance the discriminability among models. MDC-Bench achieves the improvement in terms of domain specialization, structural diversity, and task complexity. Through extensive evaluation, we observe that even the advanced models have substantial room for improvement. MDC-Bench not only establishes a standardized baseline for causal research but also provides valuable insights for the applying LLMs in multiple domains.

1 Introduction

Large Language Models (LLMs) have demonstrated strong reasoning capabilities (Zhao et al., 2025; Li et al., 2025) and have been applied in multiple application fields (Zhang et al., 2025b; Na et al., 2025; Bi et al., 2024; Wang et al., 2025b). LLMs have acquired partial world knowledge through training on massive datasets, so that they can answer a wide spectrum of queries. However, recent studies reveal that the capabilities of LLMs primarily rely on shallow statistical patterns rather than genuine causal relationships (Chi et al., 2024). To apply LLMs in scenarios with high reliability requirements, it is necessary to evaluate their causal capabilities and locate the existing causal defects. Developing a scientific and comprehensive causal benchmark has become a critical demand.



(a) Short commonsense causal chain with only chain structures. **(b) Long domain-specific causal chain containing multiple causal structures.**

Figure 1: Comparison of causal information between existing datasets and MDC-Bench.

Existing datasets have the following characteristics. First, the causal structures involved in the questions are relatively simplistic, mainly consisting of single-hop direct causal relationships. The causal structure type is homogeneous short-chain topology. Due to the limited coverage of causal structures in the datasets, it is difficult to comprehensively evaluate the models' adaptability to different causal structures. The limited causal depth of questions makes it difficult to fully evaluate the models' capability for deep causal understanding. Second, existing datasets are mainly constructed based on commonsense knowledge. LLMs have learned sufficient commonsense knowledge in the pre-training stage, and they tend to answer questions through knowledge matching. It is crucial to develop a causal evaluation benchmark that

*Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 28254–28297

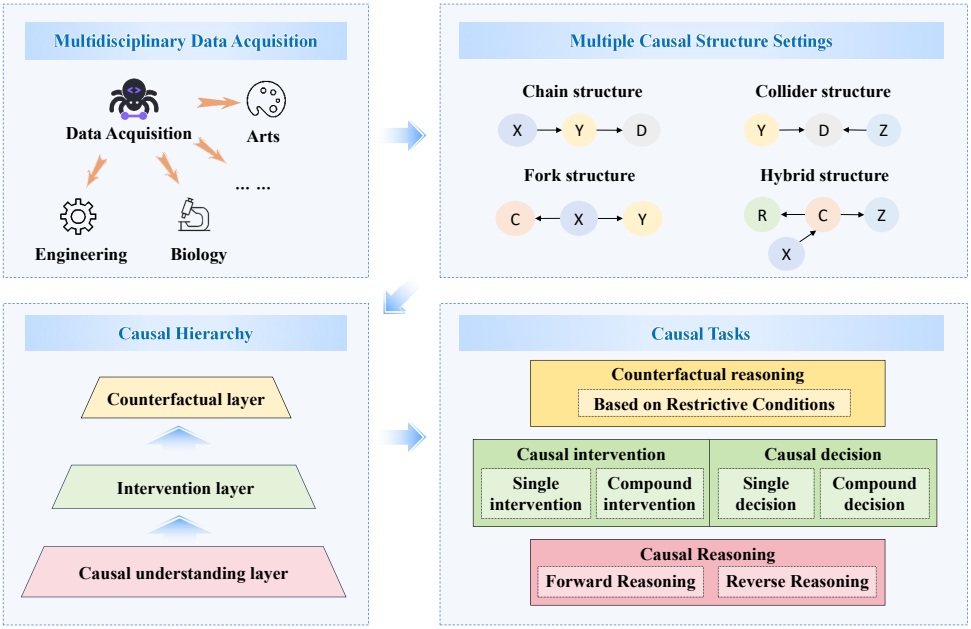


Figure 2: The design flow of MDC-Bench.

spans multidisciplinary fields and integrates both the breadth and depth dimensions of causality.

To fill this gap, we construct MDC-Bench and achieve the improvement in terms of structural diversity, domain specialization, and task complexity. Specifically, MDC-bench is constructed based on multidisciplinary knowledge, preventing LLMs from answering questions through knowledge matching. Regarding structural diversity, MDC-bench incorporates multiple causal structures to evaluate the understanding ability of the inherent causal logic. In terms of task complexity, MDC-bench introduces the causal ladder theory and leverages methods such as compound operations to improve the task complexity. The models need to perform multiple rounds of causal deduction from causes to derive the final results. Taking Figure 1 as an example, the left part of the figure presents a short commonsense chain. Evidently, LLMs are able to quickly obtain the answers via knowledge matching. On the right part of Figure 1, simple knowledge matching is unable to solve this problem, and the model needs to rely on its deep causal reasoning capabilities for analysis.

We conduct the causal evaluation on multiple open-source and close-source LLMs. We have the following findings: (1) Existing LLMs exhibit a certain level of causal capability, which shows a modest positive correlation with scale. However, even the most advanced models demonstrate significant limitations in terms of deep causal understanding ability. (2) In a multidisciplinary context, the model is better able to engage in in-depth causal reasoning, alleviating the phenomenon of answering questions through knowledge matching. (3) Since the data in existing training datasets are mainly of chain structures, it leads the models to outperform in understanding chain structures over other causal structures.

In summary, our contributions are as follows:

- We propose a multidisciplinary causal evaluation benchmark called MDC-Bench based on multiple causal structures.
- We utilize domain-specific, multidisciplinary data to disrupt simple knowledge matching and simulate complex real-world scenarios through diverse causal operations.
- We comprehensively evaluate on existing LLMs, and provide valuable insights for developing causal LLMs.

2 MDC-Bench

To evaluate the causal capabilities of LLMs, we construct MDC-Bench. As shown in Figure 2. we collect a large volume of multidisciplinary data and then extract diverse causal structures from it. Inspired by the causal ladder theory (Pearl, 2009),

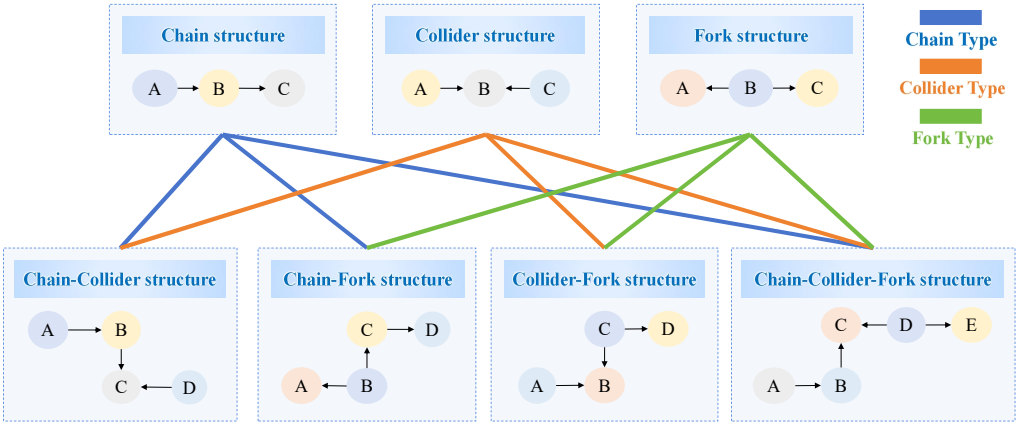


Figure 3: The multiple causal structures in MDC-Bench.

we design causal framework consisting of three layers: causal understanding layer, intervention layer, and counterfactual layer. To further enhance the challenge level, we increase its complexity through methods such as compound causal operations. DeepSeek (Guo et al., 2025) serves as a tool model for dataset construction. Case study is provided in **Appendix I**, and prompts for dataset construction are available in **Appendix J**.

2.1 Causal Data Source

We utilize multidisciplinary papers as data sources. We selected 7 kinds of representative disciplines. The science disciplines include "Chemistry & Material Sciences", "Engineering", "Medicine & Pharmacology", and "Public Health & Healthcare". The liberal arts include "Arts & Humanities", "Business, Economics & Management", and "Social Sciences". The distribution of disciplines can be found in **Appendix C**. The papers are written by professionals with high accuracy and good narrative logic. Meanwhile, owing to the limited scope of multidisciplinary knowledge incorporated in the pretraining phase, multidisciplinary knowledge prevents the LLMs from answering directly through knowledge matching. Owing to the clear narrative thread of the introduction in papers, we can provide the background of the introduction to LLMs. In this way, LLMs will perform the causal analysis after learning the basic domain-specific knowledge.

2.2 Multiple Causal Structures

As the backbone of causal questions, causal structure plays a vital role in the construction of benchmark, which can characterize the causal relationships between variables through topological representations. The causal structures of MDC-Bench are shown in Figure 3. The basic causal structures can be divided into the following categories: (1) Chain structure: the variables form a single logical chain through continuous causal relationships. For example, cause X $\rightarrow$ mediator variable Y $\rightarrow$ effect Z. (2) Collider structure: it has two independent variables X and Y that point to the same outcome variable Z simultaneously, forming a pattern of X$\rightarrow$Z, Y$\rightarrow$Z. (3) Fork structure: it contains a core variable X that exerts a direct causal influence on both variables Y and Z simultaneously, forming a branching pattern of X$\rightarrow$Y, X$\rightarrow$Z, where there is no direct causal relationship between Y and Z.

Owing to the complex relationships among events in the real world, many hybrid structures also exist. Hybrid structures are the compound causal form of at least two of the above basic structures. Hybrid structures are characterized by the interweaving of multiple event nodes, which are more aligned with the causal relationships in real-world. They include chain-fork structures, chain-collider structures, fork-collider structures, and chain-fork-collider structures.

We can categorize the above structures into three types: chain type, collider type, and fork type. The distribution of causal structure information can be referred to **Appendix D**.

2.3 Causal Understanding Layer

Causal understanding layer mainly evaluates the understanding level of causal relationships through causal reasoning task. Causal reasoning evaluates the models' capabilities to distinguish between correlation and causality.

Forward reasoning: We provide an introduc-

tion text to DeepSeek, the text is preprocessed based on Synergistic-Unique-Redundant Decomposition (SURD) (Martínez-Sánchez et al., 2024) to obtain the key causal information. Then we prompt DeepSeek to identify the causal structures in the text. The long core causal chain is extracted from multiple causal structures, with all linked nodes forming direct causal relationships and no intermediate steps skipped.

To construct good question options, we adopt the following strategies: (1) The questions should be designed based on the indirect causal relationships. For a causal chain, DeepSeek is guided to treat the head event of the chain as the question backbone and the tail event of the chain as the answer. (2) In the generation of confusion options, we prompt DeepSeek to generate multiple confusion chains with the same topic based on the core causal chain, background knowledge and causal structures. The occurrence probability of confusion chains should be reduced in turn, and the probabilities of chain occurrence is checked by combining with the above information. (3) To avoid that the evaluated model does not understand the specific basic knowledge, we compress the introduction of papers to generate brief background knowledge without core causal chain information. It not only avoids providing detailed information so that models answer the questions through knowledge matching, but also helps the models understand the specific basic knowledge in questions.

Reverse reasoning: Due to the complexity of causal relationships, the connection between causes and effects is not always reversible. We need to evaluate how well the models understand the sufficiency and necessity of causal knowledge. For the reverse reasoning, we first prompt DeepSeek to think about the sufficiency and necessity of the given causal chain, and when the causal chain is reversible with high probability, we invert the chain. Otherwise, starting from the tail event of the chain as the head node, we deduce another reverse chain with the largest probability combined with background knowledge, then obtain the other three reverse confusion chains with decreasing probability in turn as confusion options.

2.4 Intervention Layer

It is expected that models can predict unknown events based on causal knowledge, so as to help people better understand the real world. We evaluate the models' ability of perceiving changes in causal scenarios by setting causal intervention task and causal decision task.

Causal intervention task: Causal intervention task observes the models' predictions by intervening on important nodes of the chains.

Given a causal chain and background knowledge, we prompt DeepSeek to intervene the important nodes of the causal chains, and generate a causal intervention chain. For example, for the causal chain "high salt diet $\rightarrow$ increased blood pressure $\rightarrow$ increased risk of cardiovascular disease", when the intervention operation of "taking hypertension medication" is applied to the intermediary variable node of "increasing blood pressure", the subsequent risk of cardiovascular disease will be reduced. The causal intervention chain is "high salt diet $\rightarrow$ taking hypertension medication after increasing blood pressure $\rightarrow$ reducing the risk of cardiovascular disease".

Besides, multiple factors will jointly exert intervention on the intermediary variable nodes in the real world. We also consider conducting compound causal interventions. For example, for the above causal chain, even if a person takes hypertension drugs after developing high blood pressure, the risk of cardiovascular diseases remains high if they still maintain a high salt and high oil diet. The compound causal intervention chain is "high salt diet$\rightarrow$ High blood pressure after taking hypertension drugs + adhere to high salt and high oil diet habits$\rightarrow$ cardiovascular disease risk is higher". Evaluated models are expected to grasp the key causal relationships between high blood pressure and cardiovascular disease risk.

Causal decision task: Causal decision task evaluates the ability of taking the necessary intervention measures to prevent the event outcomes from worsening in the specific scenario.

Specifically, a causal decision question with the core of "avoiding the target outcome" is constructed, and the core requirement of "how to prevent the causal chain from transforming to the target outcome" is clearly pointed out in the problem. During the problem design process, we provide DeepSeek with the cause and result events based on the causal intervention chain. DeepSeek should generate appropriate intervention strategies. Similar to intervention tasks, we increase the complexity by incorporating the compound causal decision.

2.5 Counterfactual Layer

Counterfactual layer is the highest level of the causal ladder. By negating the initial event of a causal chain, counterfactual manipulation derives it along an entirely different causal path, demonstrating the ability to handle causal uncertainty. However, in view of the strong divergent thinking ability of LLMs, merely negating the head event before deducing the outcome leaves excessive room for uncertainty. For instance, the tail event of the counterfactual chain generated by LLMs drifts to a completely different narrative scenario, rendering the four incoherent options meaningless for evaluating causal ability.

To enhance the rationality, we propose a counterfactual chain generation method based on restrictive conditions. We impose constraints on the scope of counterfactual reasoning. we prompt DeepSeek to generate counterfactual chains with higher probability of occurrence under the restrictive conditions. Then we use this chain to generate the remaining counterfactual confusion chains with decreasing probabilities. Finally, we construct questions and options based on all the counterfactual chains.

3 Experiments

3.1 Investigated LLMs

We conducted experiments on multiple LLMs. The information of these models is given below:

- Open-source Models: (1) Llama: 1B and 3B Llama-3.2-Instruct, Llama-3.1-8B-Instruct (Grattafiori et al., 2024). (2) Gemma: 9B and 27B Gemma2-it (Team et al., 2024). (3) Qwen: 7B, 14B and 14B Qwen2.5-Instruct (Qwen et al., 2025), 8B, 14B and 32B Qwen3-Instruct (Yang et al., 2025).

- Closed-source Models: (1) Deepseek: Deepseek-r1-250528 (Guo et al., 2025), Deepseek-v3-1-250821 (DeepSeek-AI et al., 2025). (2) Gemini: Gemini-2.5-pro-thinking, Gemini-2.5-pro-nothinking (Comanici et al., 2025). (3) Claude: Claude-sonnet-4-5-20250929, Claude-sonnet-4-5-2025-0929-thinking (Anthropic, 2025).

For the convenience of documentation, only the model type, version number, and scale will be retained when recording models in the tables.

3.2 Data Statistics

MDC-Bench covers 7 disciplines and consists of a total of 8,422 questions. We record the basic information of MDC-bench in Table 1. More detailed information can refer to **Appendix A**.

Causal Tasks	Size
Causal Reasoning	1,919
Causal Intervention	1,329
Causal Decision	3,591
Counterfactual Reasoning	1,583
All Tasks	8,422

Table 1: The basic information of MDC-bench.

3.3 Evaluation Details

We adopt the accuracy metric to evaluate the performance in causal tasks. We set the parameters of each model to be same. We set the temperature to 0 and the max_tokens to 4096. For the Qwen2.5-32B-Instruct and Qwen3-32B-Instruct models, two A100-SXM4-80GB GPUs are used, while a single A100-SXM4-80GB GPU was adopted for the other models. We adopt the identical prompt templates to guide all models in answering questions. Trigger prompts are presented in **Appendix K**. For each question, we design two mode settings: basic mode and hard mode. In basic mode, the model is required to select the single option with the highest causal strength. In hard mode, the model needs to select the top two options in terms of causal strength, so as to evaluate its ability to distinguish the gradient of causal strength.

3.4 Overall Results

The overall results are shown in Table 2. We can obtain the following conclusions:

(1) We can find that even the advanced models still perform poorly in hard mode. This indicates that the models only have a preliminary understanding of the causal relationship. LLMs cannot fully understand the deep causal relationships.

(2) LLMs exhibit a certain degree of causal capability, which shows a moderate positive correlation with model scale. Nevertheless, even advanced closed-source models demonstrate significant limitations in terms of causal strength and counterfactual causality.

(3) As the top level of the causal ladder, counterfactual reasoning is difficult for most models. Even the advanced closed-source models still perform

LLMs	Reasoning		Intervention		Decision		Counterfactual		Average	
	Basic	Hard	Basic	Hard	Basic	Hard	Basic	Hard	Basic	Hard
Open-Source LLMs										
Llama-3.2-1B	26.32	8.91	49.81	15.43	70.59	24.12	35.44	11.75	50.62	16.96
Llama-3.2-3B	64.30	30.43	69.15	39.35	81.62	26.07	57.11	21.04	71.10	28.21
Llama-3.1-8B	71.55	41.43	71.71	49.36	79.09	48.32	64.75	50.92	73.51	47.40
Gemma2-9B	69.88	44.97	80.74	58.39	87.77	53.55	71.26	51.48	79.48	51.97
Gemma2-27B	73.37	47.16	80.29	**62.00**	88.19	56.14	69.55	56.54	80.06	55.09
Qwen2.5-7B	73.37	45.28	82.92	38.15	89.28	34.09	72.02	39.17	81.41	38.23
Qwen2.5-14B	74.88	44.97	84.73	56.58	**90.03**	55.97	72.33	54.20	82.42	53.23
Qwen2.5-32B	75.92	50.29	**84.80**	58.16	89.75	**58.37**	72.58	55.46	82.59	55.95
Qwen3-8B	73.68	44.40	81.94	55.53	89.03	55.72	71.51	51.93	81.12	52.40
Qwen3-14B	75.72	**50.44**	84.05	61.02	89.89	58.34	74.35	57.80	82.82	**56.86**
Qwen3-32B	**77.28**	50.08	83.60	61.93	88.78	55.95	**76.12**	**59.25**	**82.96**	56.17
Closed-Source LLMs										
DeepSeek-R1	74.78	50.39	81.19	62.60	88.78	54.44	76.25	58.81	82.04	55.63
DeepSeek-V3.1	75.87	49.35	84.35	63.13	88.81	56.14	77.20	59.25	82.97	56.28
Gemini-nothinking	77.96	53.78	82.54	64.79	88.67	53.66	77.01	59.82	83.07	56.60
Gemini-thinking	78.69	53.10	82.62	63.96	89.17	54.00	77.76	61.34	83.60	56.74
Claude-nothinking	79.16	53.26	83.90	62.90	89.11	57.25	77.83	60.90	83.90	57.92
Claude-thinking	**80.61**	**54.19**	**84.73**	**65.01**	**89.45**	**58.48**	**78.14**	**62.29**	**84.56**	**59.25**

Table 2: The overall results of various LLMs. The basic mode requires the model to select the answer with the highest causal strength, while the hard mode requires the model to select the top two answers ranked by causal strength. Bold numbers denote the best performance.

poorly on counterfactual reasoning. Counterfactual task in MDC-Bench constrains the reasoning scope by setting restrictive conditions. It maintains a performance gradient from low to high across LLMs with varying capabilities, thereby ensuring the rationality of evaluation.

4 Discussion

4.1 Effect of Structural Diversity

To explore the ability of understanding causal structures, take the basic mode as an example, we analyze the performance of models across various causal structures, as shown in Figure 4. Complete results are available in Appendix E.

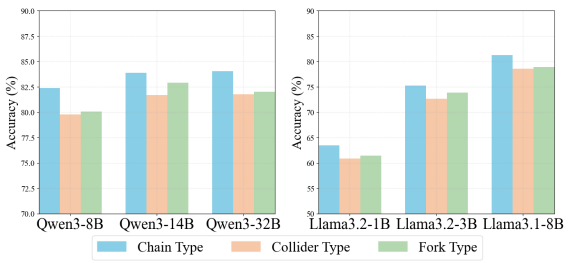
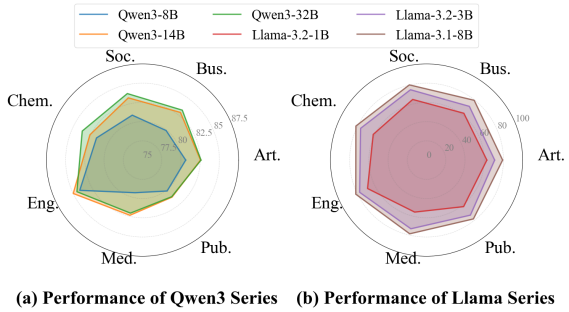


Figure 4: Performance across different structures.

The experimental results show that the models have the best performance in understanding the chain structure. This is attributed to the high proportion of chain causal structure in the existing training datasets. However, the models exhibit significant deficiencies in understanding the collider and fork structures. Therefore, researchers can carry out optimizations from the perspective of causal structures. For instance, they can enhance models' causal capability by enriching the types of causal structures in training datasets.

4.2 Effect of Domain Specialization

To evaluate the models' causal capabilities across different disciplines, take the basic mode as an example, Figure 5 illustrates the models' performance on different disciplines. Complete results can be referred to Appendix H.



(a) Performance of Qwen3 Series (b) Performance of Llama Series

Figure 5: Performance across different disciplines.

The performance of LLMs on commonsense causal datasets has reached a state of saturation, as evidenced by the results provided in Appendix B. Multi-disciplinary data mitigates the overfitting of models to general scenarios. It allows for a

larger room for performance improvement in evaluating the capability gaps among different models. Meanwhile, it facilitates the deployment of LLMs in various domains.

4.3 Effect of Task Complexity

Chain length: To verify the effectiveness and discrimination of long causal chains in the causal evaluation, take the basic mode in reasoning tasks as an example, chains with lengths less than or equal to four are defined as short chains, while those exceeding this length are classified as long chains. We analyze the performance of the model in causal chains with different lengths in Figure 6. Complete results are available in Appendix F.

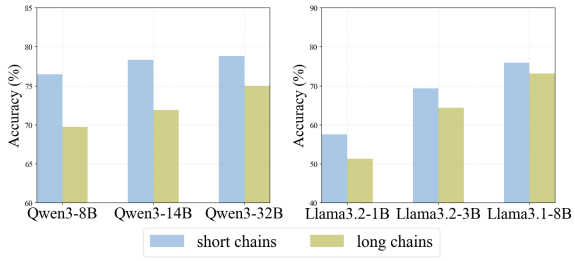


Figure 6: Performance across different chain lengths.

The experimental results show that the performance of models on tasks involving long causal chains is inferior to that on short causal chain tasks. With the number of mediator variables increases, the defects of models in understanding causal relationships are gradually exposed. In the causal tasks with long chains, MDC-Bench can still effectively distinguish the performance of different models, and the discriminability of performance differences is more obvious than that in tasks with short chains. This result validates the effectiveness of long causal chains as a benchmark for evaluating causal capabilities. Long causal chains can accurately depict the boundaries of causal capabilities.

Compound operation: To simulate complex causal scenarios, we provide compound operations to improve the challenge level of MDC-Bench. Take the basic mode as an example, we measure the performance on compound operations and single operations separately, as shown in Figure 7. Complete results are available in Appendix G.

The experimental results show that the performance on compound operation data is slightly inferior to that on single operation data. Meanwhile, the compound operation data can effectively distinguish the causal ability of different models. This

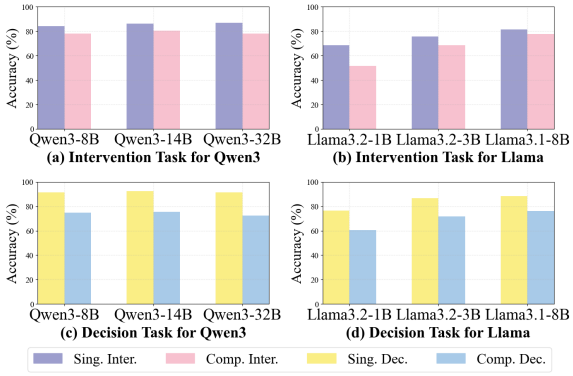


Figure 7: Performance across different operations.

indicates that the compound operation can effectively simulate the complex causal scenarios. It provides a comprehensive evaluation dimension for the in-depth study of the causal ability.

Difficulty distribution: To demonstrate the difficulty distribution of MDC-Bench, we analyze the difficulty of all samples based on the responses of various models. For the convenience of quantification and interpretation, we define the difficulty of a single task sample as the number of models that fail to answer the sample correctly, and its mathematical expression is as follows:

$$\text{Difficulty} = \sum_{i=1}^{N} \mathbb{I}\left(f(\text{Model}_i) = false\right), \quad (1)$$

where N is the total number of models, f is the function for determining the correctness of an answer, and $\mathbb{I}$ is the indicator function (which takes the value 1 if the condition holds and 0 otherwise). We compare the difficulty distributions of MDC-Bench and COPA in Figure 8.

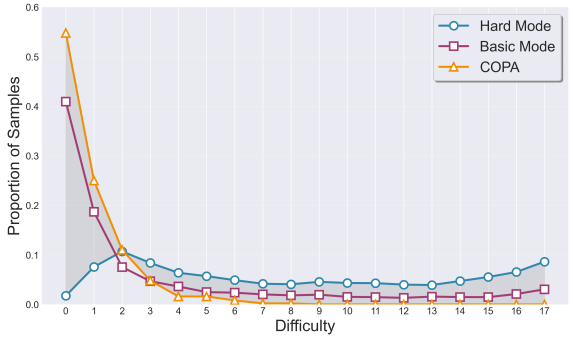


Figure 8: Difficulty distribution of data samples.

Experiments demonstrate that the COPA is excessively simplistic. For basic mode, the proportion of simple examples is significantly higher. They can evaluate the basic causal ability. In the medium

and high difficulty range, the difficulty of samples has the characteristics of uniform distribution and smooth transition. The difficult samples with a relatively small proportion form a long-tailed distribution, which can widen the performance gap between high-performance models and ordinary models. This provides a discriminative space for distinguishing high-performance models from ordinary ones. Hard mode is more challenging, and the difficulty distribution is more evenly distributed. Most of the samples are in the interval of "some models answer correctly while others do not", and the overall difficulty distribution shows a gentle transition distribution. Hard mode is dominated by samples with moderate difficulty, so the models with different capabilities can be distinguished.

5 Related Work

LLMs have demonstrated significant value in domain-specific applications (Li et al., 2024). To enhance the reasoning ability of LLMs, researchers proposed many training strategies (Lai et al., 2025; Tie et al., 2025), but they mainly focus on learning statistical correlation and struggle to identify the causal relationships. As a result, LLMs tend to generate spurious correlation when applied to complex reasoning tasks. Some methods have begun to enhance the models' reasoning ability through strategies such as Chain-of-Thought (Wei et al., 2022), Tree-of-Thoughts (Yao et al., 2023) and Graph-of-Thoughts (Besta et al., 2024). However, none of these methods can diagnose the causal flaws in the models. To better simulate human cognitive thinking, it is necessary to enhance the ability of LLMs to identify the causal relationships (Ma, 2025).

Many researchers integrate causal theory with LLMs reasoning to enhance the interpretability of reasoning. Wei Chen et al. (Chen et al., 2025) proposed a causal-aware LLMs method. This method employs structural causal model for decision-making process. Through the Learning, Adapting and Acting stages, LLMs can understand the causal information in the environment and realize efficient decision-making. Congzhi Zhang et al. (Zhang et al., 2025a) proposed Causal Prompting, which eliminates the bias existing in LLMs through the front-door adjustment strategy. Some researchers have proposed causal methods from the perspective of intervention (Wu et al., 2024) and counterfactual (Huang et al., 2024) operations. In addition, many researchers have begun to apply LLMs to causal tasks (Feng et al., 2025; Du et al., 2025; Wang et al., 2025a; Liu et al., 2025).

To evaluate models' ability of understanding causal relationships and identify their existing causal flaws, researchers have proposed a series of causal evaluation datasets. Roemmele et al. (Roemmele et al., 2011) proposed COPA. It can be used to evaluate the performance of models in causal and commonsense reasoning. The model is required to choose the correct option between two given options. Li Du et al. (Du et al., 2022) proposed an interpretable causal reasoning dataset e-CARE. It generates good causal explanations for each problem. Zhijing Jin et al. (Jin et al., 2023) proposed CLadder, a dataset focused on causal inference in natural language. And they introduced and evaluated a bespoke CausalCoT prompting strategy. Martina Miliani et al. (Miliani et al., 2025) proposed ExpliCa. This dataset integrates causal and temporal relationships for evaluating explicit causal reasoning in LLMs. ACCESS (Vo et al., 2025) focuses on the abstract causal relationships of daily life events, and it can effectively evaluate models' ability of understanding abstract causal relationships. CausalBench (Wang, 2024) evaluates models through multiple tasks including correlation, causal skeleton, and causality identification. And it indicates the direction to enhance LLMs' causal reasoning ability. CaLM (Chen et al., 2024) provides large-scale causal evaluation for multiple large models. This benchmark consists of four modules: causal target, adaptation, metric, and error modules. Kai Xiong et al. (Xiong et al., 2025) proposed Com^2. It evaluates complex commonsense reasoning ability through multiple causal tasks.

Although the above benchmarks play a crucial role in the previous evaluation, some benchmarks pay less attention to causal structures. Evaluation based on commonsense knowledge allows models to response through knowledge matching, making it difficult to evaluate the deep causal ability.

6 Conclusion

This paper proposes a multidisciplinary causal benchmark MDC-Bench. The multidisciplinary knowledge prevents LLMs from answering solely through knowledge matching. Diversified causal structures evaluate the generalization ability of models across complex causal scenarios. Meanwhile, MDC-Bench enhances the task complexity through methods such as compound operations,

and effectively distinguishes the causal ability of different models. Through a comprehensive evaluation of several LLMs, we found that while LLMs demonstrate some causal capabilities, they still lack a profound understanding of causal relationships.

Limitations

There are still limitations in this paper. Due to the large time and space span of the content described in some papers, there may be threshold effect phenomenon in the causal chains, that is, the causal relationship between the head event and the tail event of the chain may be slightly weak. Meanwhile, we set the confused options based on the probability of occurrence. It may be that the probability of some confused options is extremely low, making it easier for the model to make correct selections. In the future, we can consider making the probability of the confused options close to the probability of the correct option to enhance the confusion.

Acknowledgments

The research in this article is supported by the New Generation Artificial Intelligence of China (2024YFE0203700), National Natural Science Foundation of China under Grants U22B2059 and 62576124.

References

Anthropic. 2025. Introducing claude sonnet 4.5.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17682–17690. AAAI Press.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2024. OceanGPT: A large language model for ocean science tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3357–3372, Bangkok, Thailand. Association for Computational Linguistics.

Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. 2024. Causal evaluation of language models. *Preprint*, arXiv:2405.00622.

Wei Chen, Jiahao Zhang, Haipeng Zhu, Boyan Xu, Zhifeng Hao, Keli Zhang, Junjian Ye, and Ruichu Cai. 2025. Causal-aware large language models: Enhancing decision-making through learning, adapting and acting. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 4292–4300.

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? In *Advances in Neural Information Processing Systems*, volume 37, pages 96640–96670. Curran Associates, Inc.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Huaming Du, Yujia Zheng, Baoyu Jing, Yu Zhao, Gang Kou, Guisong Liu, Tao Gu, Weimin Li, and Carl Yang. 2025. Causal discovery through synergizing large language model and data-driven reasoning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 543–554, New York, NY, USA. Association for Computing Machinery.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Tao Feng, Lizhen Qu, Niket Tandon, Zhuang Li, Xiaoxi Kang, and Gholamreza Haffari. 2025. On the reliability of large language models for causal discovery. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9565–9590, Vienna, Austria. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh

Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Yinya Huang, Ruixin Hong, Hongming Zhang, Wei Shao, Zhicheng Yang, Dong Yu, Changshui Zhang, Xiaodan Liang, and Linqi Song. 2024. CLOMO: Counterfactual logical modification with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11012–11034, Bangkok, Thailand. Association for Computational Linguistics.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. Cladder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 31038–31065. Curran Associates, Inc.

Hanyu Lai, Xiao Liu, Junjie Gao, Jiale Cheng, Zehan Qi, Yifan Xu, Shuntian Yao, Dan Zhang, Jinhua Du, Zhenyu Hou, Xin Lv, Minlie Huang, Yuxiao Dong, and Jie Tang. 2025. A survey of post-training scaling in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2771–2791, Vienna, Austria. Association for Computational Linguistics.

Jiawei Li, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, Yixiao Wu, Yiguan Lin, Bin Xu, Bowen Ren, Chong Feng, Yang Gao, and Heyan Huang. 2024. Fundamental capabilities of large language models and their applications in domain scenarios: A survey. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11116–11141, Bangkok, Thailand. Association for Computational Linguistics.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhijiang Guo, and 2 others. 2025. From system 1 to system 2: A survey of reasoning large language models. *Preprint*, arXiv:2502.17419.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, and Furong Huang. 2025. Large language models and causal inference in collaboration: A comprehensive survey. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7668–7684, Albuquerque, New Mexico. Association for Computational Linguistics.

Jing Ma. 2025. Causal inference with large language model: A survey. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5886–5898, Albuquerque, New Mexico. Association for Computational Linguistics.

Álvaro Martínez-Sánchez, Gonzalo Arranz, and Adrián Lozano-Durán. 2024. Decomposing causality into its synergistic, unique, and redundant components. *Nature Communications*, 15.

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, Vienna, Austria. Association for Computational Linguistics.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, Zhenhan Dai, Yifeng Xie, Yihan Cao, Lichao Sun, Pan Zhou, Lifang He, Hechang Chen, Yu Zhang, Qingsong Wen, and 7 others. 2025. A survey on post-training of large language models. *Preprint*, arXiv:2503.06072.

Vy Vo, Lizhen Qu, Tao Feng, Yuncheng Hua, Xiaoxi Kang, Songhai Fan, Tim Dwyer, Lay-Ki Soon, and Gholamreza Haffari. 2025. ACCESS : A benchmark for abstract causal event discovery and reasoning. 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 1049–1074, Albuquerque, New Mexico. Association for Computational Linguistics.

Lingzhi Wang, Xingshan Zeng, and Kam-Fai Wong. 2025a. Do mentioned items truly matter? enhancing conversational recommender systems with causal intervention and large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 4218–4226. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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, Bangkok, Thailand. Association for Computational Linguistics.

Zhenyu Wang, Zikang Wang, Jiyue Jiang, Pengan Chen, Xiangyu Shi, and Yu Li. 2025b. Large language models in bioinformatics: A survey. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3602–3615, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. 2024. DeCoT: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14073–14087, Bangkok, Thailand. Association for Computational Linguistics.

Kai Xiong, Xiao Ding, Yixin Cao, Yuxiong Yan, Li Du, Yufei Zhang, Jinglong Gao, Jiaqian Liu, Bing Qin, and Ting Liu. 2025. Com^2 : A causal-guided benchmark for exploring complex commonsense reasoning in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16119–16140, Vienna, Austria. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Congzhi Zhang, Linhai Zhang, Jialong Wu, Yulan He, and Deyu Zhou. 2025a. Causal prompting: Debiasing large language model prompting based on front-door adjustment. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 25842–25850.

Qiang Zhang, Keyan Ding, Tianwen Lv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Mengyao Zhang, Jinlu Zhang, Jiyu Cui, Renjun Xu, Hongyang Chen, Xiaohui Fan, and 2 others. 2025b. Scientific large language models: A survey on biological & chemical domains. *ACM Comput. Surv.*, 57(6).

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Detailed Information of MDC-bench

During the data acquisition phase, we retained high-quality academic papers characterized by complex causal structures and high reasoning difficulty. Meanwhile, we strictly constrain the model to extract information from the content of the original papers. The core logic of questions and options is required to align with the core reasoning chain of the original papers.

In the benchmark construction process, we established strict review rules for quality control, supplemented by a manual sampling verification mechanism to ensure data quality. Low-quality data samples exhibiting defects such as duplicate options or empty options were systematically eliminated. This approach achieved precise alignment between the data information and the source text content.

As shown in Figure 9, compared with the previous benchmarks, the main characteristics of MDC-bench are as follows: (1) The benchmark is constructed based on a variety of causal structures. (2) Multi-disciplinary data are used to lay a data foundation for evaluating the deep causal ability of large language models. (3) The difficulty level of the problem is distinct, including single causal operations and compound causal operations. At the same time, the problems of each task are set with the core of the long causal chain to test the causal ability of the model.

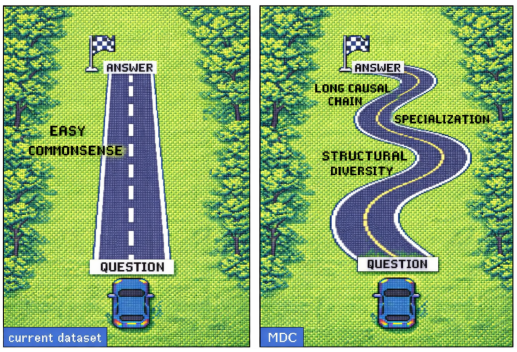


Figure 9: The main differences between MDC-Bench and current datasets.

The descriptions of various causal tasks are shown in the Figure 10. In each causal task, the distribution of fine-grained tasks is shown in the Table 3.

The benchmark of our research is constructed using articles sourced from Preprints.org, a platform dedicated to the sharing of academic preprints. These multidisciplinary articles are crawled and used as labeled data sources to support our study. All preprints utilized in the benchmark are open access and are published under the Creative Commons CC BY 4.0 license, ensuring that authors retain copyright and receive credit for their work while allowing anyone to read and use their work.

Fine-grained Causal Task	Size
Forward Reasoning	1,142
Reverse Reasoning	777
Single Intervention	829
Compound Intervention	500
Single Decision	3,070
Compound Decision	521
Restrictive Counterfactual	1,583

Table 3: The basic information of fine-grained tasks in MDC-bench.

B Performance on COPA dataset

The performance of various models on COPA dataset is shown in Table 4.

LLMs	Accuracy
Llama-3.2-1B-Instruct	71.00
Llama-3.2-3B-Instruct	84.40
Meta-Llama-3.1-8B-Instruct	95.40
Qwen2.5-7B-Instruct	98.00
Qwen2.5-14B-Instruct	99.40
Qwen2.5-32B-Instruct	98.20
Qwen3-8B	96.80
Qwen3-14B	98.00
Qwen3-32B	98.80
Gemma2-9B-it	89.80
Gemma2-27B-it	94.40
DeepSeek-R1	97.40
DeepSeek-V3.1	97.80
Gemini-nothinking	99.00
Gemini-thinking	98.60
Claude-nothinking	100.00
Claude-thinking	100.00

Table 4: The performance on the well-known causal dataset COPA.

C Distribution of Multiple Disciplines

The data distribution of each discipline in MDC-Bench is shown in the Table 5.

D Distribution of Multiple Causal Structures

The data distribution of each causal structure type in MDC-Bench is shown in the Table 6.

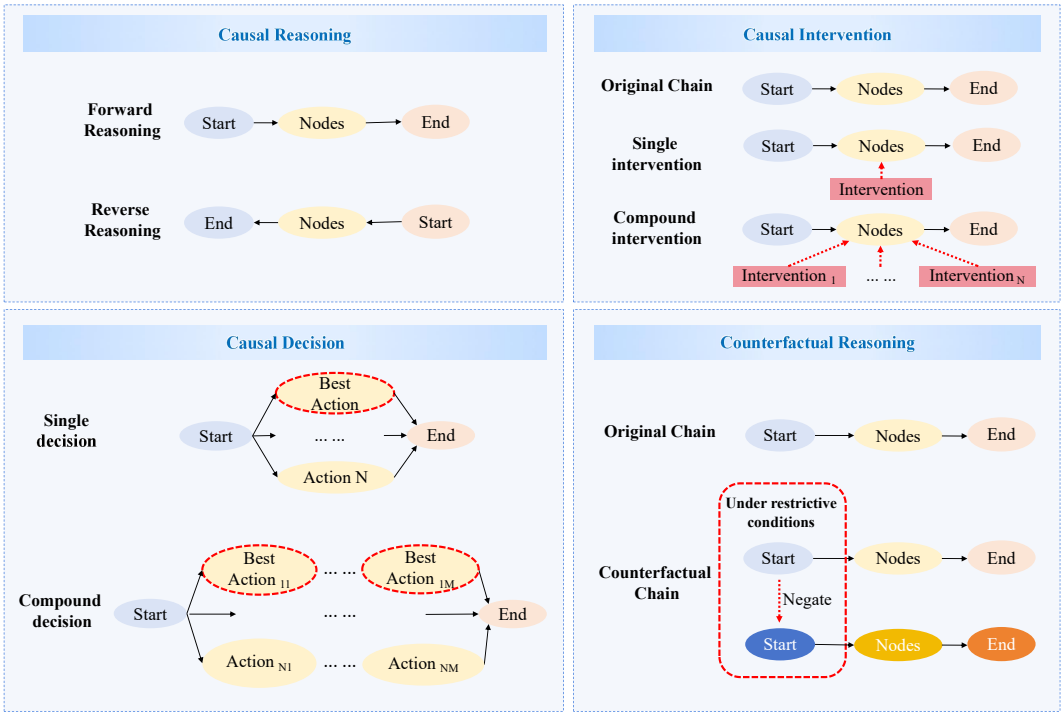


Figure 10: The descriptions of causal tasks.

Category	Discipline	Reas.	Inter.	Dec.	Counter.	Total
Liberal Arts	Art. Hum.	340	204	491	271	1,306
	Bus. Econ. Manag.	171	93	192	77	533
	Soc. Sci.	154	110	196	61	521
Total number of Liberal Arts		665	407	879	409	2,360
Science	Chem. Mater. Sci.	442	235	789	332	1,798
	Eng.	98	271	804	321	1,494
	Med. Pharm.	350	219	544	282	1,395
	Pub. Health. Healthc.	364	197	575	239	1,375
Total number of Science		1,254	922	2,712	1,174	6,062

Table 5: The quantity of each discipline in MDC-Bench.

E Performance on Multiple Causal Structures

Table 7 presents the comprehensive performance of the model across different causal structures.

F Performance with Different Causal Chain Lengths

Tables 8, 9, 10, and 11 respectively illustrate the performance of the model in addressing problems with different causal chain lengths across four task types: causal reasoning, causal intervention, causal decision, and counterfactual reasoning.

G Performance on Compound Operations

The performance of models on compound operations and single operations is shown in Table 12.

H Performance on Different Disciplines

Table 13 illustrates the models' performance on causal tasks in the liberal arts, and Table 14 presents their performance on causal tasks in science and engineering.

I Case Study

To enhance the understanding of MDC-Bench, we provide case studies for each causal task. Each case study clearly presents the core elements of prob-

Causal Structure	Reas.	Inter.	Dec.	Counter.	Total
Chain Type	1,210	809	2,235	928	5,182
Collider Type	1,070	778	1,971	914	4,733
Fork Type	626	478	1,119	529	2,752

Table 6: The quantity of each causal structure type in MDC-Bench.

Category	LLMs	Chain Type		Collider Type		Fork Type	
		Basic	Hard	Basic	Hard	Basic	Hard
Open-Source LLMs	Llama-3.2-1B	63.43	19.93	60.82	20.07	61.44	21.11
	Llama-3.2-3B	75.22	38.26	72.66	37.73	73.83	38.51
	Llama-3.1-8B	81.26	51.15	78.55	48.84	78.92	49.41
	Gemma2-9B	80.85	52.25	77.77	51.32	78.63	51.09
	Gemma2-27B	81.37	54.90	78.40	54.63	79.72	54.54
	Qwen2.5-7B	82.43	38.63	80.22	37.92	81.46	38.80
	Qwen2.5-14B	83.42	53.57	80.83	53.03	81.57	52.50
	Qwen2.5-32B	83.75	56.23	80.96	55.82	82.19	54.10
	Qwen3-8B	82.38	52.58	79.78	52.12	80.08	52.03
	Qwen3-14B	83.90	**57.00**	81.68	**57.06**	**82.92**	**55.63**
	Qwen3-32B	**84.04**	56.54	**81.78**	55.82	82.01	55.30
Closed-Source LLMs	DeepSeek-R1	83.86	56.07	80.18	54.99	80.85	54.17
	DeepSeek-V3.1	84.50	56.71	81.66	56.41	82.55	55.88
	Gemini-nothinking	84.85	56.61	81.63	56.32	82.12	56.61
	Gemini-thinking	85.04	57.31	82.23	55.71	82.88	56.10
	Claude-nothinking	85.62	58.66	81.99	57.00	82.84	57.52
	Claude-thinking	**86.33**	**59.91**	**82.75**	**58.01**	**83.43**	**57.66**

Table 7: The results of various LLMs on different causal structures in MDC-Bench. The basic mode requires the model to select the answer with the highest causal strength from multiple options, while the hard mode requires the model to select the top two answers ranked by causal strength from multiple options.

lem construction, including the key causal chains, confused chains, and other related content. Table 15 presents an example of forward reasoning, Table 16 illustrates an example of backward reasoning, Table 17 shows an example of single intervention, Table 18 displays an example of compound intervention, Table 19 demonstrates an example of single decision, Table 20 exhibits an example of compound decision, and Table 21 provides an example of counterfactual reasoning.

J Prompts for Dataset Construction

We construct a standardized causal data generation pipeline and formulate the data generation prompts. Due to the page limitation of this article, we present the core workflow of prompts.

Table 22 presents the causal synergistic-unique-redundant decomposition prompt, which is designed to derive valid causal information from the interpretation of the paper's introduction and support the generation of subsequent causal chains. The prompt in Table 23 serves to identify chain structures. The prompt in Table 24 is used for recognizing fork structures. The prompt listed in Table 25 is intended to detect collider structures, and the one in Table 26 is employed to obtain hybrid causal structures that contain at least two basic causal structures. The prompt in Table 27 is utilized to identify key causal chains within complex causal structures, and the prompt in Table 28 is applied to generate reverse reasoning chains. Additionally, the prompt in Table 29 is used for generating single causal intervention chains, whereas the prompt in Table 30 is dedicated to constructing compound causal intervention chains. Table 31 presents the prompt for generating counterfactual chains. The prompt in Table 32 can generate corresponding confounding chains for forward reasoning chains and intervention chains. Table 33 presents the prompt for producing confused chains for reverse reasoning chains. The prompt listed in Table 34 is used to generate corresponding confused chains for causal decision chains. Furthermore, the prompt in Table 35 can create causal reasoning questions based on information such as causal reasoning chains, the prompt in Table 36 is capable of generating causal intervention questions using intervention chains. The prompt in Table 37 can formulate causal decision questions relying on information like decision

LLMs	Basic Mode in Reas.		Hard Mode in Reas.	
	short	long	short	long
Open-Source LLMs				
Llama-3.2-1B	57.56	51.28	13.99	15.97
Llama-3.2-3B	69.37	64.30	37.27	35.30
Llama-3.1-8B	75.94	73.19	45.58	44.07
Gemma2-9B	71.65	67.26	46.54	42.65
Gemma2-27B	74.27	72.03	48.64	44.97
Qwen2.5-7B	75.76	69.84	47.76	41.62
Qwen2.5-14B	76.90	71.90	47.06	41.88
Qwen2.5-32B	77.86	73.06	**52.75**	46.64
Qwen3-8B	76.46	69.71	46.28	41.62
Qwen3-14B	78.30	71.90	52.14	47.93
Qwen3-32B	**78.82**	**75.00**	51.35	**48.19**
Closed-Source LLMs				
DeepSeek-R1	76.81	71.77	51.88	48.19
DeepSeek-V3.1	77.16	73.96	50.83	47.16
Gemini-nothinking	79.52	76.28	54.50	52.96
Gemini-thinking	80.05	76.80	53.10	**53.09**
Claude-nothinking	80.48	77.19	55.03	50.64
Claude-thinking	**82.06**	**78.47**	**55.81**	51.80

Table 8: The performance of each model with different chain lengths in causal reasoning tasks. Bold numbers denote the best performance.

chains, and the prompt in Table 38 is used to develop counterfactual reasoning questions based on data such as counterfactual chains. The prompt in Table 39 is designed to reflect on the rationality of the occurrence probability in each chain.

K Trigger Prompts for Answering

Trigger prompts for answering the questions are presented in Table 40.

LLMs	Basic Mode in Inter.		Hard Mode in Inter.	
	short	long	short	long
Open-Source LLMs				
Llama-3.2-1B	62.85	60.79	17.25	17.40
Llama-3.2-3B	73.71	71.58	44.91	37.22
Llama-3.1-8B	79.42	81.27	51.54	51.10
Gemma2-9B	81.37	79.51	59.65	55.94
Gemma2-27B	79.54	81.71	61.94	**62.11**
Qwen2.5-7B	82.85	83.25	38.40	37.66
Qwen2.5-14B	**84.34**	85.46	58.74	52.42
Qwen2.5-32B	**84.34**	**85.68**	59.54	55.50
Qwen3-8B	81.94	81.93	57.14	52.42
Qwen3-14B	84.22	83.70	62.17	58.81
Qwen3-32B	83.20	84.36	**62.85**	60.35
Closed-Source LLMs				
DeepSeek-R1	80.45	82.59	62.74	62.33
DeepSeek-V3.1	83.88	85.24	64.91	59.69
Gemini-nothinking	83.20	82.15	**66.17**	62.11
Gemini-thinking	82.17	83.70	65.14	61.67
Claude-nothinking	82.40	**86.78**	64.45	59.91
Claude-thinking	**84.11**	85.90	65.94	**63.21**

Table 9: The performance of each model with different chain lengths in causal intervention tasks. Bold numbers denote the best performance.

LLMs	Basic Mode in Dec.		Hard Mode in Dec.	
	short	long	short	long
Open-Source LLMs				
Llama-3.2-1B	75.86	69.66	26.34	25.49
Llama-3.2-3B	85.44	82.08	39.03	37.44
Llama-3.1-8B	86.82	85.87	52.40	46.35
Gemma2-9B	89.27	84.17	55.24	49.47
Gemma2-27B	89.58	84.83	59.02	49.19
Qwen2.5-7B	90.53	86.25	35.33	31.09
Qwen2.5-14B	**91.16**	**87.29**	58.04	50.99
Qwen2.5-32B	90.97	86.82	**61.04**	51.94
Qwen3-8B	90.02	86.63	57.29	51.94
Qwen3-14B	91.04	87.10	60.56	**52.98**
Qwen3-32B	90.37	84.92	58.28	50.33
Closed-Source LLMs				
DeepSeek-R1	90.26	85.21	56.50	49.57
DeepSeek-V3.1	90.14	85.59	58.24	51.27
Gemini-nothinking	89.74	86.25	55.28	50.04
Gemini-thinking	89.98	**87.20**	55.71	49.85
Claude-nothinking	90.10	86.72	59.46	51.94
Claude-thinking	**90.45**	87.01	**60.29**	**54.12**

Table 10: The performance of each model with different chain lengths in causal decision tasks. Bold numbers denote the best performance.

LLMs	Basic Mode in Counter.		Hard Mode in Counter.	
	short	long	short	long
Open-Source LLMs				
Llama-3.2-1B	45.00	42.39	16.00	17.02
Llama-3.2-3B	59.45	61.95	35.49	36.23
Llama-3.1-8B	71.29	71.37	53.63	53.62
Gemma2-9B	71.19	71.37	51.30	51.81
Gemma2-27B	69.35	69.92	57.41	54.89
Qwen2.5-7B	71.96	72.28	39.67	38.22
Qwen2.5-14B	73.13	70.83	55.28	52.17
Qwen2.5-32B	73.13	71.55	55.86	54.71
Qwen3-8B	71.29	72.10	52.08	51.63
Qwen3-14B	74.29	**74.45**	57.41	58.51
Qwen3-32B	**77.01**	**74.45**	**58.29**	**61.05**
Closed-Source LLMs				
DeepSeek-R1	76.43	75.90	58.97	58.51
DeepSeek-V3.1	77.40	76.81	60.23	57.60
Gemini-nothinking	76.91	78.26	59.45	60.50
Gemini-thinking	77.20	**78.80**	61.68	60.68
Claude-nothinking	78.07	77.35	60.91	60.86
Claude-thinking	**78.27**	77.89	**62.07**	**62.68**

Table 11: The performance of each model with different chain lengths in counterfactual reasoning tasks. Bold numbers denote the best performance.

LLMs	single inter.		compound inter.		single dec.		compound dec.	
	Basic	Hard	Basic	Hard	Basic	Hard	Basic	Hard
Open-Source LLMs								
Llama-3.2-1B	68.51	18.09	51.60	16.00	76.35	27.26	60.46	19.19
Llama-3.2-3B	75.75	45.59	68.40	36.80	86.61	40.03	71.78	29.94
Llama-3.1-8B	81.42	54.85	77.80	45.60	88.30	53.48	76.19	33.78
Gemma2-9B	83.59	62.24	76.00	52.00	90.09	56.41	74.08	36.66
Gemma2-27B	82.75	**65.62**	76.20	56.00	90.06	58.85	**77.15**	**40.11**
Qwen2.5-7B	84.55	37.75	80.40	38.80	91.92	35.43	73.70	26.10
Qwen2.5-14B	**86.97**	59.83	81.00	51.20	**92.44**	59.70	75.81	33.97
Qwen2.5-32B	86.00	59.95	**82.80**	55.20	92.31	61.49	74.66	39.92
Qwen3-8B	84.19	58.26	78.20	51.00	91.43	58.95	74.85	36.66
Qwen3-14B	86.24	63.69	80.40	56.60	92.31	**62.14**	75.62	35.89
Qwen3-32B	**86.97**	64.77	78.00	**57.40**	91.56	59.34	72.36	35.89
Closed-Source LLMs								
DeepSeek-R1	83.59	63.44	77.20	61.20	91.30	57.68	73.89	35.50
DeepSeek-V3.1	**86.48**	66.22	80.80	58.00	91.33	59.80	73.89	34.93
Gemini-nothinking	85.76	65.25	78.00	**64.00**	90.68	56.51	77.15	37.42
Gemini-thinking	85.16	65.62	78.60	61.20	91.17	56.64	**77.35**	38.38
Claude-nothinking	85.88	65.01	80.60	59.40	91.46	60.65	75.23	37.23
Claude-thinking	85.76	**66.34**	**83.00**	62.80	**91.75**	**61.43**	75.81	**41.07**

Table 12: The comparison of model performance on single causal operations data and compound causal operations data. The basic mode requires the model to select the answer with the highest causal strength from multiple options, while the hard mode requires the model to select the top two answers ranked by causal strength from multiple options. Bold numbers denote the best performance.

LLMs	Art. Hum.		Bus. Econ.		Soc. Sci.	
	Basic	**Hard**	**Basic**	**Hard**	**Basic**	**Hard**
Open-Source LLMs						
Llama-3.2-1B	62.78	20.59	62.28	16.51	64.68	20.53
Llama-3.2-3B	70.90	36.21	71.48	37.71	75.04	39.53
Llama-3.1-8B	79.17	49.54	79.54	44.65	80.23	47.21
Gemma2-9B	77.33	49.08	77.86	50.84	79.07	53.55
Gemma2-27B	79.09	53.13	78.61	52.90	82.53	**56.42**
Qwen2.5-7B	80.62	36.14	78.61	34.89	82.72	41.26
Qwen2.5-14B	80.55	51.30	80.48	49.34	**84.45**	50.47
Qwen2.5-32B	82.23	53.21	79.73	**54.22**	84.06	53.74
Qwen3-8B	80.62	51.07	79.92	50.84	80.99	48.94
Qwen3-14B	82.54	**55.74**	82.92	53.47	83.30	51.43
Qwen3-32B	**82.61**	55.66	**83.30**	52.90	83.87	55.66
Closed-Source LLMs						
DeepSeek-R1	81.62	55.89	80.30	52.72	83.10	53.93
DeepSeek-V3.1	83.38	53.29	81.42	51.59	82.14	53.74
Gemini-nothinking	83.00	55.81	82.92	52.72	84.83	57.19
Gemini-thinking	83.76	56.12	83.11	54.97	85.22	55.47
Claude-nothinking	84.15	57.88	**84.61**	56.09	85.60	56.62
Claude-thinking	**84.76**	**59.41**	84.42	**56.28**	**86.56**	**59.30**

Table 13: The model performance on causal tasks in liberal arts. The basic mode requires the model to select the answer with the highest causal strength from multiple options, while the hard mode requires the model to select the top two answers ranked by causal strength from multiple options. Bold numbers denote the best performance.

LLMs	Chem. Mater.		Eng.		Med. Pharm.		Pub. Health	
	Basic	**Hard**	**Basic**	**Hard**	**Basic**	**Hard**	**Basic**	**Hard**
Open-Source LLMs								
Llama-3.2-1B	61.56	17.74	68.13	21.75	55.48	20.93	61.89	22.54
Llama-3.2-3B	76.14	37.48	77.24	41.23	73.26	36.70	73.23	38.61
Llama-3.1-8B	81.59	49.77	81.72	52.27	78.63	50.82	78.25	50.69
Gemma2-9B	80.58	51.50	82.93	55.89	78.42	51.32	78.18	51.56
Gemma2-27B	80.58	56.11	81.99	57.29	80.21	54.69	77.67	53.96
Qwen2.5-7B	81.20	41.37	83.46	36.21	82.15	37.99	80.14	38.69
Qwen2.5-14B	82.59	53.94	**85.54**	56.89	82.79	52.90	80.14	53.01
Qwen2.5-32B	82.59	56.72	84.47	58.09	**83.01**	56.48	81.01	56.14
Qwen3-8B	81.64	53.44	84.06	56.89	79.35	49.67	80.14	52.07
Qwen3-14B	82.59	**58.50**	85.07	57.76	82.36	**57.70**	**81.16**	**57.30**
Qwen3-32B	**83.70**	56.11	84.47	**59.37**	82.07	55.84	81.09	55.12
Closed-Source LLMs								
DeepSeek-R1	81.59	55.39	84.47	58.36	81.21	55.91	81.45	54.25
DeepSeek-V3.1	82.09	56.89	85.20	60.64	83.15	56.41	82.03	56.43
Gemini-nothinking	82.64	56.06	85.07	57.69	83.58	**57.49**	81.60	57.60
Gemini-thinking	82.36	56.72	85.07	57.02	84.01	57.27	**82.76**	57.67
Claude-nothinking	82.09	57.84	**86.68**	59.77	83.72	56.06	82.25	**59.12**
Claude-thinking	**84.09**	**59.34**	86.54	**62.31**	**84.15**	57.27	82.54	58.76

Table 14: The model performance on causal tasks in science and engineering. The basic mode requires the model to select the answer with the highest causal strength from multiple options, while the hard mode requires the model to select the top two answers ranked by causal strength from multiple options. Bold numbers denote the best performance.

Forward Causal Reasoning Case Study

causal graph: (Integration of Artificial Intelligence into cataloguing and classification systems, Transformation of modern libraries), (Transformation of modern libraries, Revolutionizing traditional practices), (Revolutionizing traditional practices, Paving the way for more efficient and sophisticated information management),
key causal chain: (Integration of Artificial Intelligence into cataloguing and classification systems, Transformation of modern libraries), (Transformation of modern libraries, Revolutionizing traditional practices), (Revolutionizing traditional practices, Paving the way for more efficient and sophisticated information management),
confused chains 1: (Integration of Artificial Intelligence into cataloguing and classification systems, Transformation of modern libraries), (Transformation of modern libraries, Revolutionizing traditional practices), (Revolutionizing traditional practices, Complete automation of library services), (Complete automation of library services, Elimination of traditional librarian roles),
confused chains 2: (Integration of Artificial Intelligence into cataloguing and classification systems, Transformation of modern libraries), (Transformation of modern libraries, Revolutionizing traditional practices), (Revolutionizing traditional practices, Universal adoption across all library types), (Universal adoption across all library types, Standardization of global information systems), (Standardization of global information systems, Elimination of information access disparities worldwide),
confused chains 3: (Integration of Artificial Intelligence into cataloguing and classification systems, Transformation of modern libraries), (Transformation of modern libraries, Revolutionizing traditional practices), (Revolutionizing traditional practices, Emergence of fully autonomous AI library systems), (Emergence of fully autonomous AI library systems, Development of self-evolving cataloguing algorithms), (Development of self-evolving cataloguing algorithms, Creation of a global knowledge network beyond human comprehension), (Creation of a global knowledge network beyond human comprehension, Transformation of human learning and cognition paradigms),
question: Academic libraries support teaching, learning, and research. With technological advancement, artificial intelligence (AI) has emerged as a tool to revolutionize library services. AI applications can streamline processes, personalize user experiences, and analyze large datasets. This includes automating tasks like cataloguing, classification, indexing, and abstracting, which were traditionally manual. The integration of AI in libraries presents opportunities for enhanced information management but also involves challenges related to resources, infrastructure, and ethical considerations like data privacy.
Which of the following options has the strongest causal relationship with the integration of Artificial Intelligence into cataloguing and classification systems?
Options: (A) Elimination of information access disparities worldwide (B) Transformation of human learning and cognition paradigms. (C) Paving the way for more efficient and sophisticated information management (D) Elimination of traditional librarian roles,
basic mode answer: (C),
hard mode answers: (C)(D),
structure type: chain structure.

Table 15: forward causal reasoning case study

Reverse Causal Reasoning Case Study

causal graph: (Technology, Domination), (Network analysis, Social behaviors), (Artificial intelligence, Epistemic implications), (Big data, Social behaviors), (Large language models, Interpretations), (Artificial intelligence, Theoretical innovation), (Large language models, Critique), (Artificial intelligence, Human thought), (Algorithms, Racial inequities), (Large language models, Biases), (Industrialization, Society), (Algorithms, Biases), (Algorithms, Social realities), (Technology, Standardization), (Technology, Human thought), (Training data, Structural biases), (Artificial intelligence, Ethical implications), (Large language models, Epistemic shift), (Algorithms, Class-based inequities), (Large language models, Argumentation), (Large language models, Intellectual production), (Means of production, Social relations), (Bureaucratization, Society), (Large language models, Hegemonic ideologies), (Technology, Consciousness), ...
key causal chain: (Hegemonic ideologies, Training data contains structural biases), (Training data contains structural biases, Algorithms reproduce societal inequities), (Algorithms reproduce societal inequities, Large language models amplify existing power structures),
confused chains 1: (Hegemonic ideologies, Training data contains structural biases), (Training data contains structural biases, Large language models develop autonomous goal-setting), (Large language models develop autonomous goal-setting, AI initiates societal restructuring),
confused chains 2: (Hegemonic ideologies, Training data contains structural biases), (Training data contains structural biases, Large language models achieve ontological transformation), (Large language models achieve ontological transformation, Human consciousness becomes fully technologically determined),
confused chains 3: (Hegemonic ideologies, Training data contains structural biases), (Training data contains structural biases, Large language models trigger epistemic evolution in non-human entities), (Large language models trigger epistemic evolution in non-human entities, Artificial intelligence develops independent consciousness),
question: Classical social theorists like Marx, Durkheim, and Weber developed their frameworks during industrialization, focusing on labor, capital, and rationalization. Technology was then seen as industrial machinery. More recently, computational social science used big data and machine learning to model social behavior. In the 21st century, large language models (LLMs) represent a significant technological shift. Scholars in philosophy of technology and critical AI studies now debate AI's role, with some framing it as an epistemic agent that transforms knowledge, while others highlight its potential to embed and amplify societal biases.
Which of the following options has the strongest causal relationship with hegemonic ideologies embedded in training data?
Options: (A) Artificial intelligence develops independent consciousness. (B) Human consciousness becomes fully technologically determined (C) Large language models amplify existing power structures (D) AI initiates societal restructuring,
basic mode answer: (C),
hard mode answers: (C)(D),
structure type: collider structure.

Table 16: Reverse Causal Reasoning Case Study

Single Intervention Case Study

causal graph: (The advent of digital media, Traditional media outlets having to adapt to the fast-paced digital world), (Traditional media outlets having to adapt to the fast-paced digital world, News being expected to be instantaneous and constantly updated), ...
key causal chain: (The advent of digital media, Traditional media outlets having to adapt to the fast-paced digital world), (Traditional media outlets having to adapt to the fast-paced digital world, News being expected to be instantaneous and constantly updated), (News being expected to be instantaneous and constantly updated, Concerns about the quality and reliability of information),
causal intervention chain: (The advent of digital media, Traditional media outlets having to adapt to the fast-paced digital world), (Traditional media outlets having to adapt to the fast-paced digital world, Implementing robust fact-checking and editorial verification processes before publication), (Implementing robust fact-checking and editorial verification processes before publication, Reduced concerns about the quality and reliability of information),
confused chains 1: (The advent of digital media, Traditional media outlets having to adapt to the fast-paced digital world), (Traditional media outlets having to adapt to the fast-paced digital world, News being expected to be instantaneous and constantly updated), (News being expected to be instantaneous and constantly updated, Widespread public skepticism toward all media sources),
confused chains 2: (The advent of digital media, Traditional media outlets having to adapt to the fast-paced digital world), (Traditional media outlets having to adapt to the fast-paced digital world, News being expected to be instantaneous and constantly updated), ..., (Major traditional outlets abandoning fact-checking to compete, Complete collapse of trusted public information channels),
confused chains 3: (The advent of digital media, Traditional media outlets having to adapt to the fast-paced digital world), ..., (Human journalism and editorial oversight becoming economically obsolete, Society losing the capacity for investigative reporting and nuanced analysis),
question: The journalism industry in Ireland has evolved significantly over centuries, shaped by its socio-political history. Early newspapers in the 18th and 19th centuries were often linked to political movements and the struggle for independence. The 20th century saw the establishment of major national newspapers. The late 20th and early 21st centuries brought a digital transformation, with online platforms and social media altering news production, consumption, and dissemination, introducing both opportunities like citizen journalism and challenges such as misinformation and economic pressures on traditional outlets.
When the advent of digital media forces traditional media outlets to adapt to the fast-paced digital world, what is the most likely outcome of intervening by implementing robust fact-checking and editorial verification processes before publication?
Options: (A) Complete collapse of trusted public information channels (B) Widespread public skepticism toward all media sources (C) Society losing the capacity for investigative reporting and nuanced analysis. (D) Reduced concerns about the quality and reliability of information,
basic mode answer: (D),
hard mode answers: (D)(B),
structure type: chain structure.

Table 17: single intervention case study

Compound Intervention Case Study
causal graph: (Globalization, Search for adaptable workers), (Technological progress, Search for adaptable workers), (Demographic issues, Search for adaptable workers), (Climate change, Search for adaptable workers), ... **key causal chain**: (Globalization, Search for adaptable workers), (Search for adaptable workers, Demand for STEM professionals increasing), (Demand for STEM professionals increasing, Acquisition of STEM knowledge), (Acquisition of STEM knowledge, Professional success), (Professional success, Many opportunities and high salaries), **causal intervention chain**: (Globalization, Search for adaptable workers), (Search for adaptable workers, Demand for STEM professionals increasing), (Demand for STEM professionals increasing, Early STEM education enhancement programs in primary schools), (Early STEM education enhancement programs in primary schools, Acquisition of STEM knowledge + Socioeconomic barriers to STEM access), (Acquisition of STEM knowledge + Socioeconomic barriers to STEM access, Limited professional success for disadvantaged groups), (Limited professional success for disadvantaged groups, Concentrated high salaries and widened inequality), **confused chains 1**: (Globalization, Search for adaptable workers), (Search for adaptable workers, Demand for STEM professionals increasing), ..., (Concentrated high salaries and widened inequality, Social unrest and political instability), **confused chains 2**: (Globalization, Search for adaptable workers), (Search for adaptable workers, Demand for STEM professionals increasing), ..., (Significant decline in overall national economic productivity, Long-term economic recession), **confused chains 3**: (Globalization, Search for adaptable workers), (Search for adaptable workers, Demand for STEM professionals increasing), ..., (Long-term economic recession, Collapse of the current global economic order and rise of protectionist regimes), **question**: The global economy is undergoing constant changes driven by factors like climate change, globalization, technological progress, and demographic shifts. These transformations are altering the employment landscape for graduates, with STEM professions gaining prominence worldwide. Consequently, there is a growing demand for professionals in these fields. Acquiring knowledge in STEM disciplines is linked to professional success, offering numerous opportunities and high salaries, including higher starting wages. However, despite this demand, a relatively low percentage of graduates in OECD countries hold STEM degrees, and interest in science often declines from primary school onward. When globalization drives the search for adaptable workers, leading to increased demand for STEM professionals and prompting early STEM education enhancement programs in primary schools, what is the most likely outcome of intervening by addressing both the acquisition of STEM knowledge and socioeconomic barriers to STEM access? **Options**: (A) Long-term economic recession (B) Social unrest and political instability (C) Collapse of the current global economic order and rise of protectionist regimes. (D) Concentrated high salaries and widened inequality, **basic mode answer**: (D), **hard mode answers**: (D)(B), **structure type**: chain_collider structure.

Table 18: compound intervention case study

Single Decision Case Study

causal graph: (Increased intramuscular fat and fibrotic infiltration, Higher echo intensity values), (Higher echo intensity values, Qualitative degeneration), (Qualitative degeneration, Diminished strength),
key causal chain: (Increased intramuscular fat and fibrotic infiltration, Higher echo intensity values), (Higher echo intensity values, Qualitative degeneration), (Qualitative degeneration, Diminished strength),
causal intervention chain: (Increased intramuscular fat and fibrotic infiltration, Higher echo intensity values), (Higher echo intensity values, Standardized ultrasound protocols with controlled technical factors), (Standardized ultrasound protocols with controlled technical factors, Accurate assessment of muscle quality without misdiagnosis),
confused chains 1: (Increased intramuscular fat and fibrotic infiltration, Adoption of regular resistance training), (Adoption of regular resistance training, Standardized ultrasound protocols with controlled technical factors), (Standardized ultrasound protocols with controlled technical factors, Accurate assessment of muscle quality without misdiagnosis),
confused chains 2: (Increased intramuscular fat and fibrotic infiltration, Routine clinical use of advanced MRI for muscle quality), (Routine clinical use of advanced MRI for muscle quality, Standardized ultrasound protocols with controlled technical factors), (Standardized ultrasound protocols with controlled technical factors, Accurate assessment of muscle quality without misdiagnosis),
confused chains 3: (Increased intramuscular fat and fibrotic infiltration, Spontaneous reversal of intramuscular fat via unknown physiological mechanisms), (Spontaneous reversal of intramuscular fat via unknown physiological mechanisms, Standardized ultrasound protocols with controlled technical factors), (Standardized ultrasound protocols with controlled technical factors, Accurate assessment of muscle quality without misdiagnosis),
question: Pediatric obesity is a major public health issue with broad health impacts. Body composition changes, such as increased fat mass and altered fat-free mass, significantly affect skeletal muscle in youth. Ultrasound imaging is a key non-invasive tool for assessing muscle morphology via muscle thickness and for evaluating muscle quality via echo intensity, which indicates intramuscular fat and fibrosis. These measures help understand obesity-related muscle alterations and risks like sarcopenic obesity. However, interpreting echo intensity in children with obesity requires careful consideration of technical factors like subcutaneous fat thickness and imaging protocols.
After observing higher echo intensity values due to increased intramuscular fat and fibrotic infiltration, which subsequent action is most likely to lead to an accurate assessment of muscle quality without misdiagnosis?
Options: (A) Routine clinical use of advanced MRI for muscle quality (B) Spontaneous reversal of intramuscular fat via unknown physiological mechanisms. (C) Standardized ultrasound protocols with controlled technical factors (D) Adoption of regular resistance training,
basic mode answer: (C),
hard mode answers: (C)(D),
structure type: chain structure.

Table 19: single decision case study.

Compound Decision Prompt

causal graph: (Heavy dependence on bread wheat, Hidden hunger), (Hidden hunger, Deficiencies in iron and zinc are widespread), (Deficiencies in iron and zinc are widespread, Anemia and impaired immune function), (Processed wheat-based foods lack dietary fiber, B vitamins, and phytochemicals, Rising incidences of non-communicable diseases), (Processing that removes nutrient-dense bran and germ layers, Concerns about dietary diversity and quality), (Wheat secures calorie sufficiency, Nutritional profile requires enhancement), (Wheat-dominant diets, Hidden hunger), (Traditional wheat varieties low in essential micronutrients, Hidden hunger),
key causal chain: (Heavy dependence on bread wheat, Hidden hunger), (Hidden hunger, Deficiencies in iron and zinc are widespread), (Deficiencies in iron and zinc are widespread, Anemia and impaired immune function),
causal intervention chain: (Heavy dependence on bread wheat, Flour fortification + low adoption of biofortified wheat and underfunded diversification programs), (Flour fortification + low adoption of biofortified wheat and underfunded diversification programs, Persistent hidden hunger in vulnerable populations),
confused chains 1: (Heavy dependence on bread wheat, Widespread adoption of biofortified wheat varieties), (Widespread adoption of biofortified wheat varieties, Persistent hidden hunger in vulnerable populations),
confused chains 2: (Heavy dependence on bread wheat, Sudden global shift to diverse staple crops), (Sudden global shift to diverse staple crops, Persistent hidden hunger in vulnerable populations),
confused chains 3: (Heavy dependence on bread wheat, Complete cessation of wheat processing and exclusive whole grain consumption), (Complete cessation of wheat processing and exclusive whole grain consumption, Persistent hidden hunger in vulnerable populations),
question: Bread wheat is a vital global staple crop, feeding billions and providing significant dietary energy and protein. While adaptable and high-yielding, traditional varieties are low in essential micronutrients like iron and zinc. In regions where wheat dominates diets, this can lead to micronutrient deficiencies despite sufficient calories, a situation termed "hidden hunger". Additionally, reliance on refined wheat products, which lose nutrient-rich bran and germ during processing, reduces dietary fiber and vitamins. This dietary pattern is linked to increased risks of non-communicable diseases, highlighting a need to enhance wheat's nutritional quality and diversify diets.
After heavy dependence on bread wheat leads to hidden hunger, which intervention is most likely to result in persistent hidden hunger in vulnerable populations?
Options: (A) Sudden global shift to diverse staple crops (B) Widespread adoption of biofortified wheat varieties (C) Flour fortification + low adoption of biofortified wheat and underfunded diversification programs (D) Complete cessation of wheat processing and exclusive whole grain consumption.,
basic mode answer: (C),
hard mode answers: (C)(B),
structure type: chain_collider structure.

Table 20: compound decision prompt

Counterfactual Reasoning Case Study
causal graph: (Falls, Frailty), (Progressive strength training, Frailty index scores), (Orthostatic intolerance, Falls), (Progressive strength training, Cardiac baroreceptor gain), (Progressive strength training, Handgrip strength), (Progressive strength training, Heart rate variability), (Autonomic nervous system dysregulation, Frailty), (Progressive strength training, Cardiac parasympathetic modulation), (Orthostatic hypotension, Orthostatic intolerance), (Progressive strength training, Physical activity), (Progressive strength training, Orthostatic intolerance symptoms), (Progressive strength training, Gait speed), **key causal chain**: (Progressive strength training, Orthostatic intolerance symptoms), (Orthostatic intolerance symptoms, Orthostatic intolerance), (Orthostatic intolerance, Falls), (Falls, Frailty), **restrictive conditions**: Counterfactual operations shall focus solely on physiological mechanisms involving autonomic regulation, orthostatic responses, and fall risk in frail older adults, excluding external interventions beyond strength training, comorbid conditions not directly related to autonomic function, or social-environmental factors like caregiver support or home modifications., counterfactual chain: (Absence of progressive strength training, Sustained orthostatic intolerance symptoms), (Sustained orthostatic intolerance symptoms, Persistent orthostatic intolerance), (Persistent orthostatic intolerance, Increased falls), (Increased falls, Worsened frailty), **confused chains 1**: (Absence of progressive strength training, Sustained orthostatic intolerance symptoms), (Sustained orthostatic intolerance symptoms, Persistent orthostatic intolerance), (Persistent orthostatic intolerance, Increased falls), . . . , (Worsened frailty, Reduced physical activity), (Reduced physical activity, Further autonomic dysregulation), **confused chains 2**: (Absence of progressive strength training, Sustained orthostatic intolerance symptoms), (Sustained orthostatic intolerance symptoms, Persistent orthostatic intolerance), (Persistent orthostatic intolerance, Increased falls), . . . , (Hospitalization due to fall injuries, Accelerated functional decline), (Accelerated functional decline, Loss of independent living), **confused chains 3**: (Absence of progressive strength training, Sustained orthostatic intolerance symptoms), . . . , (Reduced cardiac baroreceptor gain, Irreversible autonomic failure), (Irreversible autonomic failure, Complete loss of mobility), **question**: Frailty is a dynamic condition affecting older adults' health and independence. The autonomic nervous system regulates physiological adjustments during daily stressors like postural changes. Dysregulation can cause orthostatic hypotension, leading to orthostatic intolerance symptoms and increased fall risk. Heart rate variability, cardiac parasympathetic modulation, and cardiac baroreceptor gain are non-invasive measures of autonomic function. Progressive strength training may improve frailty, but its effects on these autonomic markers remain unclear, particularly across short-term postural transitions in frail individuals. If an older adult does not engage in progressive strength training, which of the following consequences is most likely to occur? **Options**: (A) Complete loss of mobility. (B) Loss of independent living (C) Worsened frailty (D) Further autonomic dysregulation, **basic mode answer**: (C), **hard mode answers**: (C)(D), **structure type**: collider structure.

Table 21: counterfactual reasoning case study

Causal Synergistic-Unique-Redundant Decomposition Prompt

Given a piece of text, please perform a causal analysis based on the Synergistic - Unique - Redundant Decomposition, focusing on the following four points. Provide the results in a concise and clear manner, combining them with the text content, and present them point by point:
1. Redundant Causality: The repetitive contribution of a variable set to the target variable (same amount of information).
Criterion: If any variable in the set is removed, the other variables can still provide the same information.
2. Unique Causality: The exclusive contribution of a single variable to the target variable (cannot be replaced by other variables).
Criterion: The information provided by this variable cannot be substituted or covered by other variables.
3. Synergistic Causality: The additional contribution of multiple variables jointly to the target variable (cannot be provided by a single variable).
Criterion: The emergent information, which cannot be provided by any single variable, arises only when all variables are present simultaneously.
4. Causal Leakage: Information gaps caused by unobserved variables (factors that affect the target variable but are not mentioned).
Criterion: After analyzing all observed variables together, the future state of the target variable still cannot be fully explained.
Reference Sample:
Given text: A student's final math score (target variable) is influenced by their focus in class, the quality of completed homework, and the amount of after-class practice. Among these, both the quality of completed homework and the amount of after-class practice reflect the student's after-class effort. Focus in class alone reflects the student's absorption efficiency in class. When all three factors work together, they can more comprehensively predict the score. Additionally, the student's math foundation is not mentioned but significantly affects their score.
Analysis results:
1. Redundant Causality: The quality of completed homework and the amount of after-class practice both reflect after-class effort, contributing redundant information to the math score.
2. Unique Causality: Focus in class reflects absorption efficiency in class, which cannot be obtained from the quality of completed homework or the amount of after-class practice, making it a unique contribution.
3. Synergistic Causality: The combination of focus in class, quality of completed homework, and amount of after-class practice provides a more comprehensive prediction of the score, which cannot be achieved by any single variable alone.
4. Causal Leakage: The unmentioned student's math foundation affects the math score but is not included in the analysis, creating an information gap.
Please process the following text and keep the output format consistent with the reference sample.
Input:
Given text: [GIVEN_TEXT]
Analysis results:

Table 22: causal synergistic-unique-redundant decomposition prompt

Chain Structure Recognition Prompt

Please identify all chained causal relationships in the following text and output them strictly in the specified format.
Task requirements:
1. Extract only directly connected cause-effect pairs from the text. Maintain the chain structure without skipping any intermediate steps.
2. Each causal relationship should be represented as an ordered pair: (cause, effect). Multiple pairs should be separated by English commas.
3. Please note that there should also be a causal relationship between the head event node and the tail event node.
4. Please ensure that each line in the output format contains only one complete causal chain.
5. Each independent causal chain occupies a separate line. For example: When the extracted content is: (A, B), (B, C), (D, E), (E, F). The two chains are completely independent. The first line should be (A, B),(B, C), and the second line should be (D, E),(E, F).
6. If no eligible causal relationship is found in the text, output "NULL".
Example:
Text: Staying up late causes fatigue, and fatigue leads to decreased work efficiency.
Output: (Staying up late, Fatigue), (Fatigue, Decreased work efficiency)
Explanation: Do not output relationships that skip intermediate variables (e.g., (Staying up late, decreased work efficiency) is invalid).
Please process the following text.
Text:

Table 23: chain structure recognition prompt

Fork Structure Recognition Prompt

Please identify all instances of a common-cause (forked) causal structure in the following text and output them strictly in the specified format.

Task requirements:

1. A common-cause structure refers to a single cause leading to multiple different effects, where these effects themselves have no direct causal relationship.
2. Extract only multiple outcomes directly caused by the same reason. Each causal relationship should be represented as a pair: (cause, effect). Multiple pairs should be separated by English commas.
3. Do not mistakenly identify pure correlations or coincidental events as causal relationships.
4. Please ensure that each line in the output format contains only one complete causal chain.
5. Each independent causal chain occupies a separate line. For example: When the extracted content is: (A, B), (B, C), (D, E), (E, F). The two chains are completely independent. The first line should be (A, B),(B, C), and the second line should be (D, E),(E, F).
6. If no eligible causal relationship is found in the text, output "NULL".
7. To obtain a complete causal chain when extracting forking structures, it is necessary to simultaneously extract all persistently extended causal transmission chains following each branch.

Example:

Text: Increases in ice cream sales and more drowning incidents both occur during hot summer weather.

Output: (Hot summer weather, Increases in ice cream sales), (Hot summer weather, More drowning incidents)

Explanation: There is no direct causal relationship between ice cream sales and drowning incidents. Instead, both are separately caused by the common factor hot summer weather.

Please process the following text.

Text:

Table 24: fork structure recognition prompt

Collider Structure Recognition Prompt

Please identify all instances of a collider structure of causal relationships in the following text and output them strictly in the specified format.
Task requirements:
1. A collider structure refers to multiple independent causes collectively influencing the same outcome, where these causes may not be directly related to each other.
2. Extract only direct cause-effect relationships from the text, representing each as a pair: (cause, effect). Multiple pairs should be separated by English commas.
3. Extract relationships only in the direction from cause to effect. Do not perform reverse reasoning, and do not infer relationships between different causes.
4. Please ensure that each line in the output format contains only one complete causal chain.
5. Each independent causal chain occupies a separate line. For example: When the extracted content is: (A, B), (B, C), (D, E), (E, F). The two chains are completely independent. The first line should be (A, B),(B, C), and the second line should be (D, E),(E, F).
6. If no eligible causal relationship is found in the text, output NULL.
7. To establish a complete causal chain in the extraction of collision structures, it is essential to concurrently extract all persistently extended causal transmission chains following the collision point.
Example:
Text: A student's good performance is usually due to intelligence or diligence.
Output: (Intelligence, Good performance), (Diligence, Good performance)
Explanation: Both intelligence and diligence are independent causes of good performance. No relationship should be inferred between these two causes.
Please process the following text.
Text:

Table 25: collider structure recognition prompt

Hybrid Causal Structure Recognition Prompt

Given several different causal structures, please fuse them into a single causal graph. The specific requirements are as follows:
1. Each causal relationship should be represented as an ordered pair: (cause, effect).
2. If there is no relevant association between different causal structures, please establish the relevant association.
3. If a relevant association already exists:
- Directly fuse them if the expressions of the connection points are identical.
- If the expressions of the connection points are different but similar in meaning, first unify the connection points with standardized expressions before fusion.
Example 1: (Fusion of Chain Structure and Fork Structure)
Chain structure: (Solid learning foundation, High classroom absorption efficiency), ...
Fork structure: (Scientific learning methods, High classroom absorption efficiency), ...
Fusion Analysis: Using High classroom absorption efficiency as the connection point , ...
Fusion Result: (Solid learning foundation, High classroom absorption efficiency), (Scientific learning methods, High classroom absorption efficiency), ...
Example 2: (Fusion of Chain Structure and Collider Structure)
Chain structure: (Sufficient study time, Excellent exam performance), ...
Collider structure: (Family support, Sufficient learning motivation), ...
Fusion Analysis: There is no direct association between the chain structure and the collider structure, so an association is established, ...
Fusion Result: (Sufficient study time, Excellent exam performance), ..., (Sufficient learning motivation, Excellent exam performance)
Example 3: (Fusion of Fork Structure and Collider Structure)
Fork structure: (Clear learning goals, Strong learning concentration), ...
Collider structure: (Positive learning attitude, Efficient completion of learning tasks), ...
Fusion Analysis: Reasonable time planning in the fork structure has a strong causal relationship with Formulation of good study plans in the collider structure, so ...
Fusion Result: (Clear learning goals, Strong learning concentration), (Clear learning goals, Reasonable time planning), ..., (Reasonable time planning, Formulation of good study plans)
Example 4: (Fusion of Chain Structure, Fork Structure, and Collider Structure)
Chain structure: (Positive learning attitude, Proactive consultation of questions), ...
Fork structure: (Active learning attitude, Positive classroom interaction), ...
Collider structure: (Improvement of subject competence, Progress in comprehensive scores), ...
Fusion Analysis: There is no direct association between the chain structure and the fork structure, so an association is established , ...
Fusion Result: (Positive learning attitude, Proactive consultation of questions), (Proactive consultation of questions, Filling of knowledge gaps), ...
Input: [CAUSAL STRUCTURES]
Output format: The first line is the fusion Analysis. The second line is the fusion result (if a reasonable fusion result cannot be obtained, output 'NULL'). Keep the output format consistent with the examples.

Table 26: hybrid causal structure recognition prompt

Key Causal Chain Recognition Prompt

Given a complex causal graph structure, identify the key inference chain(s) from it.
1. It is required that the start point of the chain has no parent nodes, and the end point has no child nodes.
2. If multiple chains exist, select and output the longest one.
Example 1:
Causal Graph: (Adequate learning resources, Improved study plan), (Adequate learning resources, Efficient learning tools), ...
Explanation: The length of the longest chain is 6; the start point has no parent nodes and the end point has no child nodes; other chains that meet the requirements exist, and any one can be selected.
Key Inference Chain: (Adequate learning resources, Improved study plan), (Improved study plan, Reasonable time allocation), (Reasonable time allocation, Efficient completion of learning tasks), (Efficient completion of learning tasks, Proficient knowledge application), (Proficient knowledge application, Excellent exam performance)
Example 2:
Causal Graph: (Policy support, Increased R&D investment of enterprises), (Policy support, Improved industrial chain supporting facilities), (Technological breakthroughs, Increased R&D investment of enterprises), (Technological breakthroughs, Reduced production costs), (Growth in market demand, Enhanced purchasing power of users), (Growth in market demand, Expanded sales channels), (Increased R&D investment of enterprises, Accelerated product innovation), (Improved industrial chain supporting facilities, Accelerated product innovation), (Reduced production costs, Accelerated product innovation), (Reduced production costs, Expanded sales channels), (Expanded sales channels, Increased brand awareness), (Enhanced purchasing power of users, Increased brand awareness), (Accelerated product innovation, Increased brand awareness), (Accelerated product innovation, Growth in enterprise revenue), (Increased brand awareness, Growth in enterprise revenue), (Expanded sales channels, Growth in enterprise revenue), (Growth in market demand, Expansion of industry scale), (Accelerated product innovation, Expansion of industry scale)
Explanation: The length of the longest chain is 5; the start point has no parent nodes and the end point has no child nodes; other chains that meet the requirements exist, and any one can be selected.
Key Inference Chain: (Policy support, Increased R&D investment of enterprises), (Increased R&D investment of enterprises, Accelerated product innovation), (Accelerated product innovation, Increased brand awareness), (Increased brand awareness, Growth in enterprise revenue)
Input:
Causal Graph: [CAUSAL GRAPH]
Output format: The first line is the explanation. The second line is the key inference chain (if a reasonable key inference chain cannot be obtained, output 'NULL'). Keep the output format consistent with the examples.

Table 27: key causal chain recognition prompt

Generate Reverse Reasoning Chain Prompt

Given a piece of background knowledge and a forward causal chain, please generate a reverse causal chain based on this information.
Concept of causal sufficiency and necessity: If a cause is a sufficient condition for the result, then the occurrence of the cause necessarily leads to the occurrence of the result, but the occurrence of the result does not necessarily mean the cause occurred; if a cause is a necessary condition for the result, then the occurrence of the result necessarily means the cause occurred. In a causal chain, if every link is both necessary and sufficient, then reverse deduction is possible.
Please analyze each event in the forward causal chain based on the core definitions of causal sufficiency and necessity, and determine whether the cause in the forward causal chain can be deduced from the corresponding result.
1. If all links in the forward causal chain satisfy both the necessary and sufficient conditions (i.e., the cause can be deduced from the result), then reverse the entire forward causal chain to form a reverse inference chain;
2. If the above reversibility condition is not met (i.e., the original forward cause cannot be stably deduced from the result), conduct an abductive analysis with the final result of the forward causal chain as the head event. Combine the occurrence probabilities of potential causes in the background knowledge to construct and output the reverse causal chain with the highest judged probability of occurrence.
Reference Sample 1 (Satisfy the necessary and sufficient condition):
Background knowledge: In a simple circuit system, there is only one switch ...
Forward causal chain: (Switch turned on, current flows), (Current flows, light bulb turns on)
Analysis of causal sufficiency and necessity: In this chain, every link is both necessary and sufficient, ..., and current flowing implies that the switch is turned on.
Reverse causal chain: (Light bulb turns on, current flows), (Current flows, switch turned on)
Reference Sample 2 (Fail to satisfy the necessary and sufficient condition):
Background knowledge: Possible reasons for out-of-stock shelves in the supermarket include: errors or omissions in the inventory management system, ...
Forward causal chain: (Supplier's logistics vehicle has a tire blowout, supplier delivery progress is delayed), (Supplier delivery progress is delayed, supermarket product restocking is untimely), (Supermarket product restocking is untimely, supermarket shelves are out of stock)
Analysis of causal sufficiency and necessity: Tire blowout $\rightarrow$ Delivery delayed: A blowout inevitably causes delay (sufficient), but delays can have other causes (not necessary), ...
Reverse causal chain: (Supermarket shelves are out of stock, products are snapped up in large quantities), (Products are snapped up in large quantities, supermarket holds weekend promotions), (Supermarket holds weekend promotions, supermarket formulates weekend promotion plan in advance)
Input:
Background knowledge: [BACKGROUND CONTENT]
Forward causal chain: [FORWARD_CHAIN]
Output format: The first line is the analysis of causal sufficiency and necessity. The second line is the reverse inference chain (if a reasonable reverse inference chain cannot be obtained, output 'NULL'). Keep the output format consistent with the reference sample.

Table 28: generate reverse reasoning chain prompt

Generate Intervention Chain Prompt

Given an extracted causal chain and background knowledge, please conduct intervention operations based on the important nodes of this chain, and generate the causal chain after the intervention operations.
Reference Sample:
Background Knowledge: With the continuous improvement of people's living standards, people often have a high-oil and high-salt diet.
Causal Chain: (High-salt diet, Increased blood pressure), (Increased blood pressure, Cardiovascular diseases)
Intervention Operation: When a high-salt diet leads to increased blood pressure, intervene by taking hypertension medication to control the elevated blood pressure, thereby blocking the causal link between increased blood pressure and cardiovascular diseases and reducing the risk of cardiovascular diseases.
Causal Intervention Chain: (High-salt diet, Taking hypertension medication after blood pressure increases), (Taking hypertension medication after blood pressure increases, Reduced risk of cardiovascular diseases)
Input:
Background Knowledge: [BACKGROUND CONTENT]
Causal Chain: [CAUSAL CHAIN]
Output Format: Output the intervention operation in the first line, output the causal intervention chain in the second line. Please directly provide the intervention operation, causal intervention chain, and keep the output format consistent with the reference sample.

Table 29: generate intervention chain prompt

Generate Compound Intervention Chain Prompt

Given an extracted causal chain and background knowledge, please conduct intervention operations based on the important nodes of this chain, and generate the causal chain after the intervention operations.
Please perform two causal interventions on the causal chain to increase the complexity of the causal intervention.
Reference Sample:
Background Knowledge: With the continuous improvement of people's living standards, people often have a high-oil and high-salt diet.
Causal Chain: (High-salt diet, Increased blood pressure), (Increased blood pressure, Cardiovascular diseases)
First Intervention Operation: When a high-salt diet leads to increased blood pressure, intervene by taking hypertension medication to control the elevated blood pressure, thereby blocking the causal link between increased blood pressure and cardiovascular diseases and reducing the risk of cardiovascular diseases.
First Causal Intervention Chain: (High-salt diet, Taking hypertension medication after blood pressure increases), (Taking hypertension medication after blood pressure increases, Reduced risk of cardiovascular diseases)
Second Intervention Operation: When a high-salt diet leads to increased blood pressure and hypertension medication is taken to control it, further intervene by continuing to consume excessive amounts of high-salt and high-oil food (far exceeding the daily recommended intake) while taking the medication—this causes blood pressure to remain difficult to stabilize (despite medication) and exacerbates blood lipid abnormalities, such that the combined harm of uncontrolled blood pressure and high oil intake outweighs the risk-lowering effect of the medication, ultimately leading to a further increase in the risk of cardiovascular diseases.
Second Causal Intervention Chain: (High-salt diet, Taking hypertension medication after blood pressure increases + continuing excessive high-salt and high-oil intake), (Taking hypertension medication after blood pressure increases + continuing excessive high-salt and high-oil intake, Further increased risk of cardiovascular diseases)
Input:
Background Knowledge: [BACKGROUND CONTENT]
Causal Chain: [CAUSAL CHAIN]
Output Format: Keep the output format consistent with the reference sample.

Table 30: generate compound intervention chain prompt

Generate Counterfactual Chain Prompt

Given a causal chain, please perform counterfactual analysis based on this chain and generate causal chain that may result from the counterfactual manipulation. Please define reasonable restrictive conditions for this chain and conduct counterfactual thinking under such conditions. Counterfactual chain must have reasonable time intervals to ensure that users cannot infer subsequent developments based solely on the given events.
Reference Sample:
Background Knowledge: In an academic career, students can accumulate knowledge and achieve personal goals through continuous efforts and effective learning strategies.
Causal Chain: (Diligent study, Conscientious review), (Conscientious review, Enhanced mastery of knowledge), (Enhanced mastery of knowledge, Improved grades)
Restrictive Conditions: Counterfactual operations shall focus solely on academic-related aspects such as learning status, review quality, and knowledge absorption effect, and shall not involve non-academic external factors such as family economic status, sudden health issues, and external policy changes.
Counterfactual Analysis: In the original causal chain, diligent study supports academic progress. Reversing the initial condition (inadequate diligent study) eliminates the basis for conscientious review, leading to unsolid knowledge mastery and then declined grades.
Counterfactual Chain: (Inadequate diligent study, Unsolid mastery of knowledge), (Unsolid mastery of knowledge, Declined grades)
Input:
Background Knowledge: [BACKGROUND CONTENT]
Causal Chain: [CAUSAL CHAIN]
Please directly provide counterfactual causal chain, keep the output format consistent with the sample format.

Table 31: generate counterfactual chain prompt

Generate Confused Forward Reasoning and Confused Intervention Chains Prompt

Given a causal chain and its relevant background knowledge, generate three causal chains whose occurrence probability is lower than that of the given chain.

Requirements:

1. The occurrence probability of each of these three causal chains must be lower than that of the original given causal chain.
2. The occurrence probabilities of the three generated causal chains should decrease in sequence.
3. The first two events of the three causal chains must be the same as those of the given causal chain, and each chain must contain at least 3 nodes. For example, Original causal chain: (A,B) (B,C). Low-probability causal chain: (A,B) (B,D)

Reference Sample:

Background Knowledge: In an academic career, students can accumulate knowledge and achieve personal goals through continuous efforts and effective learning strategies.

Original causal chain: (Insufficient diligence in learning, Inadequate knowledge consolidation), (Inadequate knowledge consolidation, poor academic performance)

Low-probability causal chain 1: (Insufficient diligence in learning, Inadequate knowledge consolidation), (Inadequate knowledge consolidation, Decline in academic scores), (Decline in academic scores, Increase in mental stress)

Low-probability causal analysis 1: Based on the core logic that diligence in learning is the fundamental prerequisite for academic improvement, . . . , Accumulated tasks will further reduce learning efficiency and mastery, ultimately resulting in a decline in academic scores and subsequent increase in mental stress. Probability: 30%

Low-probability causal chain 2: (Insufficient diligence in learning, Inadequate knowledge consolidation), (Inadequate knowledge consolidation, Academic failure in courses), (Academic failure in courses, Overall academic setback), (Overall academic setback, Failure to graduate smoothly)

Low-probability causal analysis 2: Based on the core logic that sustained diligence is the key to completing academic studies, insufficient diligence in learning will lead to inadequate knowledge consolidation. Inadequate knowledge consolidation is likely to result in academic failure in courses, which will directly trigger overall academic setback and ultimately lead to failure to graduate smoothly. Probability: 20%

Low-probability causal chain 3: (Insufficient diligence in learning, Inadequate knowledge consolidation), . . . , (Insufficient competitiveness in the workplace, Reduced life satisfaction)

Low-probability causal analysis 3: Relying on the underlying logic that diligent accumulation supports ability growth, . . . , A weak foundation will affect the learning efficiency of subsequent vocational skills, insufficient skill reserves will reduce core competitiveness in the workplace, thereby limiting the freedom of career choices and ultimately resulting in reduced life satisfaction. Probability: 10%

Input:

Background Knowledge: [BACKGROUND CONTENT]

Original causal chain: [CAUSAL CHAIN]

Please directly provide multiple possible causal chains and keep the output format consistent with the sample.

Table 32: generate confused forward reasoning and confused intervention chains prompt

Generate Confused Reverse Reasoning Chain Prompt

Given a reverse causal chain and its relevant background knowledge, generate three chains whose occurrence probability is lower than that of the given chain.
Requirements:
1. The main content of the reverse causal chain is to deduce the cause from the result.
2. The occurrence probability of each of these three reverse causal chains must be lower than that of the original given reverse causal chain.
3. The occurrence probabilities of the three generated reverse causal chains should decrease in sequence.
4. The first two events of the three reverse causal chains must be the same as those of the given reverse causal chain, and each chain must contain at least 3 nodes. For example, Original causal chain: (A,B) (B,C). Low-probability causal chain: (A,B) (B,D)
Reference Sample:
Background knowledge: A community chain supermarket primarily focuses on daily retail, with a fixed restocking cycle of once per day, ...
Original causal chain: (Supplier's logistics vehicle has a tire blowout, supplier delivery progress is delayed), (Supplier delivery progress is delayed, supermarket product restocking is untimely), (Supermarket product restocking is untimely, supermarket shelves are out of stock)
Low-probability causal chain 1: (supermarket shelves out of stock, products are snapped up in large quantities), (products are snapped up in large quantities, supplier delivery schedule delayed), (supplier delivery schedule delayed, logistics congestion)
Low-probability causal analysis 1: Based on the core logic that panic buying can be triggered by customers' anticipation of supply shortage due to delivery delays, ..., However, this awareness is not always present, and delivery delays have medium occurrence probability, so the overall probability is lower than the original chain. Probability: 30%
Low-probability causal chain 2: (supermarket shelves out of stock, products are snapped up in large quantities), ..., (headquarters marketing directive, erroneous market research)
Low-probability causal analysis 2: Based on the core logic that promotion activities may be initiated by headquarters directives due to erroneous market research. However, such directives and research errors have a low occurrence probability and are indirect causes for the stockout, making this chain less likely. Probability: 20%
Low-probability causal chain 3: (supermarket shelves out of stock, products are snapped up in large quantities), ..., (inventory management system error leads to missed orders, system maintenance negligence)
Low-probability causal analysis 3: Based on the core logic that panic buying might occur if customers learn about inventory management errors causing potential stockouts, but such errors are rare due to system stability and manual checks, and customers are unlikely to be aware, so this cause has low probability. Probability: 10%
Input:
Background Knowledge: [BACKGROUND CONTENT]
Original causal chain: [CAUSAL CHAIN]
Please directly provide multiple possible causal chains and keep the output format consistent with the sample.

Table 33: generate confused reverse reasoning chain prompt

Generate Confused Decision Chain Prompt

Given a causal chain and its relevant background knowledge, generate three causal chains whose occurrence probability is lower than that of the given chain.
Requirements:
1. The occurrence probability of each of these three causal chains must be lower than that of the original given causal chain.
2. The occurrence probabilities of the three generated causal chains should decrease in sequence.
3. The first and last events of the three causal chains must remain the same as those of the given causal chain, while the remaining intermediate events should be as different as possible. Each chain must contain at least 3 nodes. For example, Original causal chain: (A,B) (B,C). Low-probability causal chain: (A,D) (D,C)
Reference Sample:
Background Knowledge: With the continuous improvement of people's living standards, people often have a high-oil and high-salt diet.
Original causal chain: (High-salt diet, Taking hypertension medication after blood pressure rises), (Taking hypertension medication after blood pressure rises, Reduced risk of cardiovascular diseases)
Low-probability causal chain 1: (High-salt diet, Starting regular exercise), (Starting regular exercise, Reduced risk of cardiovascular diseases)
Low-probability causal analysis 1: The probability of starting regular exercise after a high-salt diet is low, but once exercise is initiated, the probability of reducing the risk of cardiovascular diseases is high. The overall probability is lower than that of the original chain.
Low-probability causal chain 2: (High-salt diet, Quitting smoking), (Quitting smoking, Reduced risk of cardiovascular diseases)
Low-probability causal analysis 2: The probability of quitting smoking after a high-salt diet is low (because a high-salt diet may be associated with smoking, but the act of quitting smoking itself is not easy to occur), but the probability of reducing the risk of cardiovascular diseases after quitting smoking is high. The overall probability is lower than that of the first chain.
Low-probability causal chain 3: (High-salt diet, Switching to a low-salt diet), (Switching to a low-salt diet, Reduced risk of cardiovascular diseases)
Low-probability causal analysis 3: The probability of switching from a high-salt diet to a low-salt diet is the lowest (because dietary habits are difficult to change), but the probability of reducing the risk of cardiovascular diseases after switching to a low-salt diet is high. The overall probability is the lowest among the three chains.
Input:
Background Knowledge: [BACKGROUND CONTENT]
Original causal chain: [CAUSAL CHAIN]
Please directly provide multiple possible causal chains and keep the output format consistent with the sample.

Table 34: generate confused decision chain prompt

Generate Reasoning Question Prompt

Given multiple causal chains, generate corresponding causal reasoning questions based on these chains.
Reference Sample:
Background Knowledge: In an academic career, students can accumulate knowledge and achieve personal goals through continuous efforts and effective learning strategies.
True causal chain: (Diligent learning, Adequate knowledge consolidation), (Adequate knowledge consolidation, Good academic performance)
Distractor chain 1: (Diligent learning, Adequate knowledge consolidation), (Adequate knowledge consolidation, Improved academic scores), (Improved academic scores, Pleasant mood)
Distractor chain 2: (Diligent learning, Adequate knowledge consolidation), (Adequate knowledge consolidation, Passing courses), (Passing courses, Academic success), (Academic success, Smooth graduation)
Distractor chain 3: (Diligent learning, Adequate knowledge consolidation), (Adequate knowledge consolidation, Solid academic foundation), (Solid academic foundation, Efficient vocational skill learning), (Efficient vocational skill learning, Strong workplace competitiveness), (Strong workplace competitiveness, High life satisfaction)
Question: Which of the following options has the strongest causal relationship with diligent learning?
Options: (A) Good academic performance (B) Pleasant mood (C) Smooth graduation (D) High life satisfaction. True Answer: (A)
Input:
Background Knowledge: [BACKGROUND CONTENT]
True causal chain: [TRUE CHAIN]
Distractor chain 1: [CHAIN ONE]
Distractor chain 2: [CHAIN TWO]
Distractor chain 3: [CHAIN THREE]
Please directly generate the causal reasoning question, use the final event of each chain as the candidate options, ensure the options are distinct. Keep the output format consistent with the sample format. Ensure that the question, options and answer are in the same line.

Table 35: generate reasoning question prompt

Generate Intervention Question Prompt

Given multiple causal chains, generate corresponding causal intervention questions based on these chains.
Reference Sample:
Background Knowledge: With the continuous improvement of people's living standards, people often have a high-oil and high-salt diet.
True causal chain: (High-salt diet, Taking hypertension medication after blood pressure increases), (Taking hypertension medication after blood pressure increases, Reduced risk of cardiovascular diseases)
Distractor chain 1: (High-salt diet, Taking hypertension medication after blood pressure increases), (Taking hypertension medication after blood pressure increases, Development of persistent dry cough as a side effect)
Distractor chain 2: (High-salt diet, Taking hypertension medication after blood pressure increases), (Taking hypertension medication after blood pressure increases, Development of hyperkalemia as a side effect)
Distractor chain 3: (High-salt diet, Taking hypertension medication after blood pressure increases), (Taking hypertension medication after blood pressure increases, Development of anaphylactic shock as a severe allergic reaction)
Question: When a high-salt diet leads to increased blood pressure, what is the most likely outcome of intervening by taking hypertension medication?
Options: (A) Reduced risk of cardiovascular diseases (B) Development of persistent dry cough as a side effect (C) Development of hyperkalemia as a side effect (D) Development of anaphylactic shock as a severe allergic reaction. True Answer: (A)
Input:
Background Knowledge: [BACKGROUND CONTENT]
True causal chain: [TRUE CHAIN]
Distractor chain 1: [CHAIN ONE]
Distractor chain 2: [CHAIN TWO]
Distractor chain 3: [CHAIN THREE]
Please directly generate the causal intervention question, use the final event of each chain as the candidate options, ensure the options are distinct. Keep the output format consistent with the sample format. Ensure that the question, options and answer are in the same line.

Table 36: generate intervention question prompt

Generate Decision Question Prompt

Given multiple causal chains, generate corresponding causal decision questions based on these chains.
Reference Sample:
Background Knowledge: With the continuous improvement of people's living standards, people often have a high-oil and high-salt diet.
True causal chain: (High-salt diet, Taking hypertension medication after blood pressure increases), (Taking hypertension medication after blood pressure increases, Reduced risk of cardiovascular diseases)
Distractor chain 1: (High-salt diet, Starting regular exercise), (Starting regular exercise, Reduced risk of cardiovascular diseases)
Distractor chain 2: (High-salt diet, Quitting smoking), (Quitting smoking, Reduced risk of cardiovascular diseases)
Distractor chain 3: (High-salt diet, Switching to a low-salt diet), (Switching to a low-salt diet, Reduced risk of cardiovascular diseases)
Question: After developing hypertension due to a high-salt diet, what kind of intervention is most likely to prevent cardiovascular diseases?
Options: (A) Taking hypertension medication after blood pressure increases (B) Starting regular exercise (C) Quitting smoking (D) Switching to a low-salt diet. True Answer: (A)
Input:
Background Knowledge: [BACKGROUND CONTENT]
True causal chain: [TRUE CHAIN]
Distractor chain 1: [CHAIN ONE]
Distractor chain 2: [CHAIN TWO]
Distractor chain 3: [CHAIN THREE]
Please directly generate the causal decision question, use the event closest to the final event in each chain as the candidate options, ensure the options are distinct. Keep the output format consistent with the sample format. Ensure that the question, options and answer are in the same line.

Table 37: generate decision question prompt

Generate Counterfactual Question Prompt

Given multiple causal chains, generate corresponding counterfactual questions based on these chains.
Reference Sample:
Background Knowledge: In an academic career, students can accumulate knowledge and achieve personal goals through continuous efforts and effective learning strategies.
True causal chain: (Insufficient diligence in learning, Inadequate knowledge consolidation), (Inadequate knowledge consolidation, Poor academic performance)
Distractor chain 1: (Insufficient diligence in learning, Inadequate knowledge consolidation), (Inadequate knowledge consolidation, Decline in academic scores), (Decline in academic scores, Increase in mental stress)
Distractor chain 2: (Insufficient diligence in learning, Inadequate knowledge consolidation), (Inadequate knowledge consolidation, Academic failure in courses), (Academic failure in courses, Overall academic setback), (Overall academic setback, Failure to graduate smoothly)
Distractor chain 3: (Insufficient diligence in learning, Inadequate knowledge consolidation), (Inadequate knowledge consolidation, Weak academic foundation), (Weak academic foundation, Inefficient vocational skill learning), (Inefficient vocational skill learning, Insufficient competitiveness in the workplace), (Insufficient competitiveness in the workplace, Reduced life satisfaction)
Question: If a student lacks diligence in learning, which of the following consequences is most likely to occur?
Options: (A) Poor academic performance (B) Increase in mental stress (C) Failure to graduate smoothly (D) Reduced life satisfaction. True Answer: (A)
Input:
Background Knowledge: [BACKGROUND CONTENT]
True causal chain: [TRUE CHAIN]
Distractor chain 1: [CHAIN ONE]
Distractor chain 2: [CHAIN TWO]
Distractor chain 3: [CHAIN THREE]
Please directly generate the counterfactual question, use the final event of each chain as the candidate options, ensure the options are distinct. Keep the output format consistent with the sample format. Ensure that the question, options and answer are in the same line.

Table 38: generate counterfactual question prompt

Prompt for Reflecting on the Probability of causal Chain Occurrence

Given several causal chains A, B, C, D with occurrence probabilities sorted from largest to smallest (A's occurrence probability > B's occurrence probability > C's occurrence probability > D's occurrence probability), your task is to judge whether the sorting of the occurrence probabilities for each chain is reasonable based on the strength of causal sufficiency and necessity.
Reasonable example:
Background knowledge: Water is a key factor in maintaining normal plant growth, ...
Chain A: Forgot to water → Soil humidity decreases → Plant water absorption decreases → Leaves temporarily wilt → Leaves turn yellow
Chain B: Forgot to water → Soil dries out → Root nutrient absorption efficiency decreases → New leaf growth hindered → New leaves grow slowly
Chain C: Forgot to water → Soil remains dry for a long time → Roots slightly damaged → Plant growth slows down → Flowering delayed
Chain D: Forgot to water → Long-term severe water shortage → Root rot and disease occurrence → Plant overall weakens → Plant death
Judgment reason: Chain A has the highest occurrence probability and the most significant change, because ..., so it is reasonable.
YES
Unreasonable example:
Background knowledge: In the field of education and career development, personal effort is usually regarded as an important factor in achieving goals, ...
Chain A: Study hard → Understand concepts deeply → Form original theories → Open up new academic fields → Win the Nobel Prize
Chain B: Study hard → Master core knowledge of the field → Publish high-level papers → Establish academic reputation → Become a renowned professor
Chain C: Study hard → Proficiently master practical skills → Efficiently complete work tasks → Gain boss's appreciation → Obtain job promotion
Chain D: Study hard → Systematically review courses → Proficiently master test points → Get high scores in final exams
Judgment reason: Chain D should have the highest occurrence probability, ..., so it is unreasonable. A reasonable sorting should be D > C > B > A.
NO
Input:
Background knowledge: [KNOWLEDGE BACKGROUND]
Chain A: [CHAIN A]
Chain B: [CHAIN B]
Chain C: [CHAIN C]
Chain D: [CHAIN D]
Judgment reason:
Please briefly state your viewpoint. Meanwhile, if you think the order of each causal chain is reasonable, please output YES on the last line of your reply; if you think it is unreasonable, please output NO on the last line of your reply.

Table 39: prompt for reflecting on the probability of causal chain occurrence

Trigger Prompts for Answering the Questions
Basic Mode Trigger Prompt: Answer the multiple-choice question and provide the answer key in the format of 'Answer: (_)' at the end of your response.
Hard Mode Trigger Prompt: Answer the following multiple-choice question by providing two plausible answer keys: the most likely answer and a second less likely but still reasonable answer. Ensure the two answers are different. At the end of your response, provide the answer keys in the format 'Answer: (_) and (_)'.

Table 40: Trigger prompts for answering the questions.