CausalityCheck: A Framework for Evaluating Causal Reasoning in Large Language Models

Jiang Li[*], **Zehua Duo**[*], **Guanglai Gao, Xiangdong Su**[†]
[1] College of Computer Science, Inner Mongolia University, China
[2] National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, China
[3] Inner Mongolia Key Laboratory of Multilingual Artificial Intelligence Technology, China
lijiangimu@gmail.com, cssxd@imu.edu.cn

Abstract

Causal reasoning is a crucial component of understanding complex phenomena and building intelligent systems. Recent advancements in large language models (LLMs) have demonstrated their strong capabilities in reasoning tasks; however, their true understanding of causal relationships remains limited, particularly in cases where causal chains are misidentified or reliance on empirical inference occurs. To mitigate the risk that models misclassify data as false positives due to these issues, we introduce **CausalityCheck**, an automated tool designed to efficiently generate causal reasoning checklists. This checklist enables the creation of multi-task causal reasoning datasets with task generalization and reasoning robustness from a single causal reasoning dataset. Using **CausalityCheck**, we developed CausalityCheck-CP to assess the causal reasoning abilities of 18 LLMs. This framework also measures the extent to which causal chains are misidentified or rely on empirical inferences. Our results indicate that the current large language models still face two critical issues when handling complex causal reasoning tasks: incorrect identification of causal chains and reliance on empirical inference. The code and data are available at https://github.com/dzh597/CausalityCheck.

1 Introduction

Causal reasoning constitutes a crucial component in understanding complex phenomena and constructing intelligent systems (Jin et al., 2023; Schölkopf, 2022; Weinberg et al., 2025). In recent years, causal reasoning capability has increasingly been recognized as one of the key attributes through which large language models (LLMs) demonstrate intelligence (Liu et al., 2025b). Presently, large language models have achieved significant progress in numerous reasoning tasks, particularly within problem-solving (Miliani et al., 2025) and text generation domains (Wang and Shen, 2024; Yu et al., 2026). However, when tackling causal reasoning tasks, these models often rely on empirical inference rather than genuine understanding of causal relationships (Chi et al., 2024). This limitation means existing evaluation methods fail to comprehensively reflect a model's causal reasoning capabilities.

To effectively assess causal reasoning, traditional evaluation methods primarily focus on task-specific answer accuracy, such as examining a model's performance on given causal questions using causal reasoning datasets (Wang, 2024; Chen et al., 2024b; Liu et al., 2025a). These datasets typically design various scenarios requiring models to comprehend and deduce outcomes based on causal relationships. However, current approaches suffer from a significant flaw: they predominantly emphasize problem-solving capability while overlooking the correctness of the reasoning process. This includes issues such as misidentifying causal chains and relying on empirical inferences, as illustrated in Figure 1. Misidentifying causal chains occurs when a model captures an erroneous causal chain during inference yet still arrives at the correct answer; this error becomes apparent with minor perturbations. The correct causal chain in the diagram should be: *working hours* $\rightarrow$ *amount of hair* $\rightarrow$ *performance*. The model erroneously posited that less hair leads to better performance. When perturbed by shaving everyone bald, the model concluded an outcome clearly erroneous result that everyone's performance was excellent. Reliance on empirical inference occurs when models do not construct causal relationships but instead search for answers within the training data. The diagram should first identify two causes within the context, then infer the outcome based on these causes. However, the model directly searched for the individual

[*]These authors contributed equally to this work.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 16403–16429

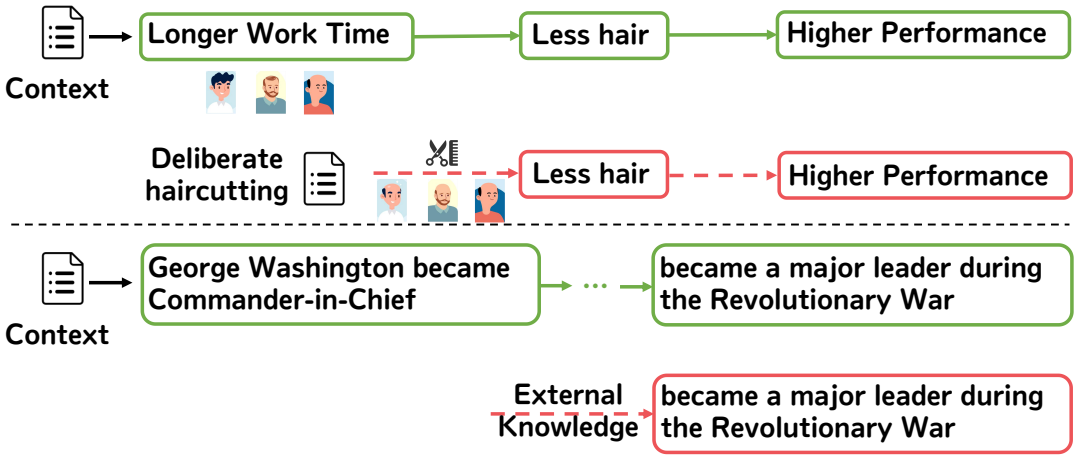


Figure 1: The figure illustrates examples of misidentifying causal chains and relying on empirical inferences. Green is correct. Though arriving at the correct answer, red employed an erroneous line of reasoning.

George Washington and cheated by retrieving the answer from its pre-trained data. The model did not rely on the problem's context and did not even know the cause of this outcome; it merely got the answer correct.

This paper introduces a novel causality reasoning capability assessment tool, CausalityCheck, aimed at comprehensively evaluating causal reasoning abilities. Its core principle is that deep causal reasoning should encompass correct reasoning processes, not merely a single answer. CausalityCheck constructs a 4×4 dimensional checklist to evaluate model performance across diverse causal reasoning tasks. Unlike traditional single-task assessments, it emphasizes task generalization and reasoning robustness, forming a multi-task evaluation framework grounded in causal understanding. By incorporating task generalization and reasoning robustness, where task generalization encompasses origin problem, answerable judging, process judging, and definitive verdict, and reasoning robustness includes problem solving, problem understanding, irrelevant disruption, and virtual causal reasoning, we can evaluate models' causal reasoning capabilities across more complex and diverse scenarios. Should issues arise such as misidentified causal chains or reliance on empirical inferences, corresponding evaluation metrics will decline. The core principle of CausalityCheck is this: if a model genuinely comprehends causality, it should produce robust reasoning across diverse tasks, despite varying distractions and contextual shifts.

Through CausalityCheck, we propose CausalityCheck-CP, a causal reasoning dataset generated from the CausalProbe-H and CausalProbe-M datasets. It comprises up to 8920 high-quality samples and 16 sets of checklist matrices, enabling comprehensive evaluation of models' textual reasoning capabilities. Experimental results reveal two major challenges for current large language models in causal reasoning tasks: misidentification of causal chains and reliance on empirical inferences. Specifically, models exhibit lower accuracy when handling tasks requiring deep understanding of causal chains, indicating limited performance in causal chain recognition. Similarly, accuracy is low when relying on stereotypes from training data for inference, indicating that models produce erroneous reasoning when dependent on empirical inferences. In contrast, higher accuracy is observed in the VCR task, suggesting that models prefer to rely on causal relationships within the current context when encountering virtual vocabulary or unseen entities, thereby improving reasoning accuracy. Therefore, compared to the original question-answering tasks in mainstream benchmarks, our CausalityCheck evaluation more accurately reflects a model's causal reasoning capabilities.

Due to space limitations, the related work is presented in Appendix B.

2 CausalityCheck

CausalityCheck is an automated tool for efficiently generating causality inference checklists, encompassing common causality inference tasks and diverse types of robustness testing. The fundamental workflow of CausalityCheck and MATHCHECK is similar (Zhou et al., 2024), differing only in their specific tasks. Within our checklists, various causal

tasks are arranged row-by-row to assess a task's generalizability; while columns feature different variants of casual reasoning problems to evaluate the robustness of reasoning. Section 2.1 details the task types, Section 2.2 discusses diverse variants of casual reasoning problems, and Section 2.3 explains how checklist data is constructed.

2.1 Task Generalization

To test models across tasks within the same domain and prevent reliance on empirical reasoning, we incorporated four categories of mathematical tasks into CausalityCheck: original problems, answerable judgements, procedural judgements, and causal validity.

Original Problems. In this task, we require the model to answer a given causal reasoning question. As the most commonly used method for testing causal reasoning ability in current research, it demands that the model analyze the question, recall and apply appropriate causal knowledge, and ultimately derive an inference.

Answerable Judging. Given a causal inference question, the model must determine whether sufficient information is provided to answer it. This task requires the model to analyze the question, identify the key conditions necessary for answering it, and verify whether these conditions are present in the question statement. This set includes both questions that provide key conditions and those that omit them, aiming to ensure, through comparative testing, that the model does not provide answers randomly when not constrained by the prompt.

Process Judging. Given a causal inference problem with both a correct causal chain and an incorrect one, the model must determine whether the causal chain reflects a valid process. Compared to the answerable judgments, process judgments involve a more granular evaluation of solutions, requiring the model to judge the correctness of causal chains. This helps distinguish potentially erroneous answers.

Causal Validity Assessing. In Causal Validity Assessments, we categorize the model's conclusions into four types. The first type, correct and reliable, refers to conclusions that are not only accurate but also derived through rigorous causal reasoning, taking into account potential variables and biases, and maintaining consistency across various scenarios. The second type, correct but prone to error, refers to conclusions that appear correct on the surface, but their reasoning relies on overly simplified or overlooked assumptions, such as ignoring potential confounding variables or over-relying on a specific dataset or time frame, leading to accurate results in some scenarios but potentially misleading ones in others. The third type, incorrect but prone to correctness, refers to conclusions that are fundamentally incorrect, but their reasoning methods, assumptions, or data are close to being correct in certain aspects or are valid in specific contexts, though they lack generalize ability or exhibit significant bias. The fourth type, incorrect and unreliable, refers to conclusions that are incorrect, with causal reasoning that is neither rigorous nor based on trustworthy assumptions or data, making them untenable across different scenarios and prone to leading to misleading and erroneous decisions.

Diverse task configurations test whether models possess a deep understanding of the same problem, rather than merely accepting a single answer as correct. The first example in Figure 1 illustrates this point: although the model arrived at the correct answer, it constructed an obviously erroneous causal relationship. While the model could answer the original question correctly, it erred in its procedural judgment, indicating that it did not deeply comprehend the causal relationships within this example.

2.2 Reasoning Robustness

A model that genuinely comprehends the intrinsic causal relationships within a problem will exhibit robust reasoning when confronted with multiple variants of that problem. It should not correctly identify the answer while failing to recognize the causal chain. To address such issues, we employed four problem forms: the original problem and its three rewritten variants to examine the model's reasoning robustness.

Problem Solving. This serves as the foundational question for other robustness variants. As the most fundamental functional test, it verifies whether the model possesses basic causal capabilities without modification.

Problem Understanding. This involves transforming the original problem into a new formulation. Although the wording or sentence structure differs, the underlying mathematical logic remains unchanged. This task focuses more on semantic robustness, aiming to assess whether the model can reason correctly when confronted with the same mathematical logic presented through different descriptions.

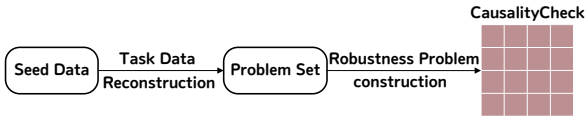


Figure 2: Dataset Construction Process

Irrelevant Disturbance. This involves inserting conditions related to the original problem's subject but irrelevant to the final answer. Previous research indicates LLMs are susceptible to such disturbances. This task requires models to distinguish between conditions essential for solving the problem and those that are extraneous.

Virtual Causal Reasoning. By replacing entities containing real-world information in problems with meaningless virtual vocabulary, this eliminates the possibility of models directly extracting usable information from training data. It tests whether models can make reasonable inferences based solely on causal logic in the absence of prior knowledge. The virtual vocabulary is generated directly by large language models, ensuring that it does not conflict with any existing real-world terms.

The second example in Figure 1 illustrates the VCR task, where George Washington is replaced with a virtual word to prevent data leakage. For scenarios where all context originates from training data, both question comprehension and irrelevant interference formats are employed to assess reasoning robustness.

2.3 Checklist Construction

Creating CausalityCheck data is a labour-intensive and time-consuming process. The advent of LLMs has introduced new flexibility and quality in generating causal reasoning content. Consequently, we employ LLMs such as GPT-5 used in our experiments as engines to automatically generate our CausalityCheck data. The dataset construction workflow is illustrated in Figure 2. Users first collect annotated causal reasoning questions as seed data. Subsequently, LLMs expand these questions to construct multiple causal reasoning tasks related to the original question, completing the task data reconstruction. Thirdly, questions under each task are rewritten into their robustness variants, forming a robustness question set. Finally, each data point undergoes two rounds of manual inspection to ensure the correct composition of the CausalityCheck dataset. To ensure the accuracy and consistency of the annotation process, we adopted the standards outlined in Appendix D

Based on seed data, we automatically rewrite them to expand task generalization capabilities, constructing multiple tasks including answerable judgments, process judgments, and causal validity assessments. For answerable judgment tasks, we replace answers in seed data with answerable ones while leaving others unchanged to form an answerable judgment data point. We then prompt the model to remove a key cause from this answerable judgment data, thereby generating an unanswerable data point. For process judgment tasks, we apply the context of a single data point to LLMs to construct a causal chain. Specifically, given a question and its correct answer, we first build a correct causal chain. We then prompt the model to introduce erroneous modifications to this correct chain, thereby generating an incorrect causal chain. For causal validity assessment, we prompt the model to rewrite four data points corresponding to: correct and reliable, correct but prone to error, incorrect but prone to correctness, and incorrect and unreliable. Each data point sequentially selects one of these options as its answer. The final answer is not a causal relationship but one of these four states, enabling a more nuanced evaluation of the answer's validity. Once we obtained four distinct tasks including the seed data, we generated three additional robustness questions for each corresponding task. Question comprehension and irrelevant interference tasks involve rewriting the question without altering the final answer. For question comprehension, we solely prompt the model to rewrite the causal reasoning context. For irrelevant interference, we solely prompt the model to add an irrelevant interfering cause or effect. For VCR, we prompt the model to identify a contextually rich entity word within the text and replace it with a fictional term non-existent in the real world. This replacement must occur wherever the word appears in the context, question, and answer, ensuring its complete removal from the dataset. Through these steps, a single data point can be expanded into 36 distinct instances. A set of CausalityCheck-CP construction processes is listed in Appendix J.

3 Experiments

3.1 Dataset

We use CausalityCheck to comprehensively evaluate text-based causal reasoning abilities and in-

troduce a benchmark dataset, CausalityCheck-CP. CausalityCheck-CP is generated from CausalProbe-H and CausalProbe-M (Chi et al., 2024), a dataset designed to assess causal reasoning over individual events. We select CausalProbe-2024 (which includes the Easy, Hard, and Multi series) as the seed benchmark dataset for the following reasons: (a) it is the most recent dataset for evaluating causal reasoning abilities over individual events, with the H and M series offering higher difficulty and more distractors, thus providing a ready basis for analyzing distractors and incorrect answers; (b) our objective is to assess whether state-of-the-art models possess genuine deep causal reasoning abilities, and the tasks focus on individual event causal reasoning, which aligns with the diversity of individual events typically reported in the media.

We first collected a subset of 1,200 problems from CausalProbe-H, ensuring that the difficulty of the problems is evenly distributed. Subsequently, we generated 16 groups (43,200 examples in total) through CausalityCheck, where the incorrect choices from CausalProbe-M were used to guide the generation of distractors for large models. This dataset serves as a tool for evaluating the deep causal reasoning capabilities of LLMs. To mitigate any inherent regularities in the generated data, we selected only 25% of the data as the final test set (10,800 examples). All datasets underwent meticulous manual validation to ensure high quality and reliability. For this, we recruited ten graduate students, who were trained according to the specific requirements of our study. After a screening process, 8,920 examples were retained. The number within each group is shown in Figure 3. Details of the screening process and discussions regarding GPT-generated data biases are provided in Appendix F and G.

3.2 Setups

To systematically evaluate the causal reasoning capabilities of contemporary large language models, we conducted a comprehensive assessment of 18 representative models. These models can be categorized into three broad groups: (a) general-purpose large language models, (b) reasoning-augmented models, and (c) lightweight models. We employed the Accuracy metric to evaluate task performance. The list of selected models and detailed assessment settings can be found in Appendix C.

	OP *1	AJ *2	PJ *2	CVA *4
PS *1	88	536	532	1072
PU *1	123	430	548	1008
ID *1	152	524	534	1032
VCR *1	119	562	576	1084

Figure 3: The number within each group, the figure beside each task indicates the quantity expanded from the seed data to that group.

3.3 Experimental Results

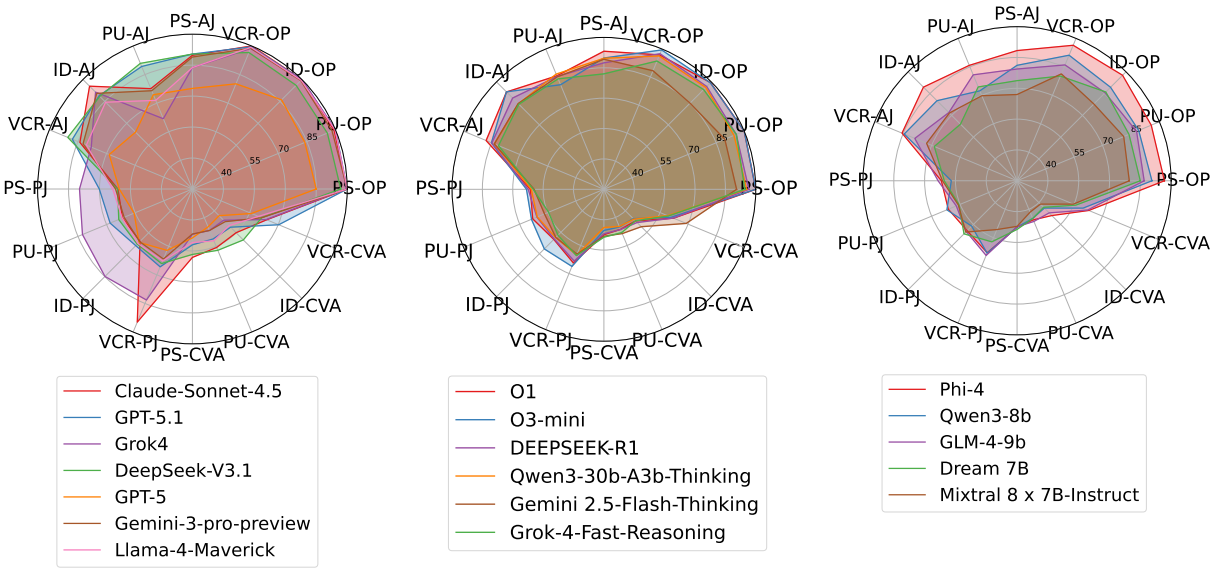
The Table 1 reports each model's overall accuracy and performance across eight major task categories, and the Figure 4 further breaks these results down across sixteen sub-tasks, with the corresponding numerical data provided in Appendix A.

We conducted a detailed analysis of two issues within causal reasoning tasks: incorrect identification of causal chains and reliance on empirical inference. Firstly, the problem of incorrect causal chain identification manifests through the lower accuracy of the Process Judgment (PJ) task compared to other tasks. As shown in the detailed data in Appendix A, the PJ task consistently exhibits lower accuracy than other tasks. For instance, Claude-Sonnet-4.5 achieves an accuracy of 59.46% on the PJ task, markedly lower than its 82.27% performance on the AJ task. This disparity indicates that while the model demonstrates strong reasoning capabilities in answerable judgement tasks, it may fail to accurately identify causal relationships when confronted with tasks requiring deeper understanding of causal chains, thereby yielding erroneous inferences.

Secondly, issues stemming from reliance on empirical inference are revealed through accuracy variations in VCR tasks. Across all models, performance in the VCR task generally surpassed that in other tasks. For instance, GPT-5.1 achieved an accuracy of 74.51% in the VCR task, outperforming its results in the PS, PU, and ID tasks. This suggests that when confronted with causal reasoning

Model	All	OP	AJ	PJ	CVA	PS	PU	ID	VCR
Claude-Sonnet-4.5-20250929	71.43	99.76	82.27	59.46	44.23	69.67	64.73	70.41	80.91
GPT-5.1-20251113	70.92	99.85	85.49	55.95	41.55	70.84	69.52	67.96	74.51
Grok4	70.62	99.41	71.68	76.80	34.57	70.00	63.74	74.21	74.52
DeepSeek-V3.1	69.33	95.07	86.68	50.57	45.02	68.37	69.13	69.06	70.77
GPT-5	67.41	99.48	86.05	49.32	35.84	67.29	61.32	60.5	70.13
Gemini-3-pro-preview	65.23	98.78	79.43	48.61	34.08	65.98	60.77	66.09	68.06
Llama-4-Maverick	65.19	99.54	73.45	50.15	37.63	66.12	61.76	65.46	67.43
O1	67.48	98.77	87.45	49.44	34.28	67.93	65.05	66.51	70.45
O3-mini	67.03	99.62	82.96	53.25	32.31	66.13	63.24	68.17	70.60
DeepSeek-R1	65.25	95.53	82.13	48.74	34.59	62.46	62.80	64.47	69.26
Qwen3-30b-A3b-Thinking	63.41	94.16	81.53	46.78	31.19	63.30	62.21	61.80	66.35
Gemini 2.5-Flash-Thinking	62.32	84.55	80.92	44.82	38.99	62.46	60.05	59.91	66.86
Grok-4-Fast-Reasoning	62.14	92.87	77.26	44.00	34.42	61.72	60.85	61.05	64.94
Phi-4	65.49	95.47	83.07	48.46	34.97	64.85	62.77	65.13	69.23
Qwen3-8B	59.39	86.73	72.71	46.11	32.01	58.81	56.09	56.50	66.16
GLM-4-9B	58.77	82.13	70.34	48.26	34.37	59.03	57.39	54.86	63.82
Dream 7B	53.62	78.39	60.24	44.90	30.94	55.94	52.94	51.62	53.97
Mixtral 8 x 7B-Instruct	51.15	73.33	60.35	42.74	28.17	51.69	49.48	50.28	53.12

Table 1: Model performance on CausalityCheck-CP. PS: Problem Solving, AJ: Answerable Judging, PJ: Process Judging, CVA: Causal Validity Assessing, OP: Original Problem, PU: Problem Understanding, ID: Irrelevant Disturbance, VCR: Virtual Causal Reasoning. Each score is the average score of related units. For example, 'All' means all units, PS includes solving units on four problem types, OP includes original problems on four tasks units.



(a) general-purpose large language models (b) reasoning-augmented models (c) lightweight models.

Figure 4: Performance of models in three categories.

tasks, models often rely on direct inferences from experiences within the training data rather than resolving problems through deep causal reasoning. The models' reliance on stereotypical patterns within their experience leads to lower accuracy in these tasks compared to the VCR task. As the VCR task involves virtual vocabulary or unseen entities, the models cannot draw upon existing experience for inference. Consequently, they tend to focus more intently on causal relationships within the current context, thereby enhancing their reasoning accuracy.

In summary, the two major issues of incorrect causal chain identification and reliance on empirical inference reveal the limitations of current large language models when handling complex causal reasoning tasks. To enhance model performance in complex causal reasoning, future research must focus more intently on how models accurately identify causal relationships and whether they can transcend reliance on empirical inference to engage in deeper causal reasoning. In addition, we have analyzed two phenomena in Appendix H and I.

3.4 Ablation Study

We conducted a comparative evaluation of six prompting methods to analyze the impact of different prompting strategies on causal reasoning task performance. The methods employed included Zero-shot Prompting (Kojima et al., 2023), Few-shot Prompting (Brown et al., 2020), Chain-of-Thought (CoT) (Wei et al., 2023), Tree-of-Thought (ToT) (Yao et al., 2023), Highlighted Chain-of-Thought (HoT) (Lei et al., 2025), and Single-Property Enhancement Prompting (SPE) (Do et al., 2025). Zero-shot refers to providing only the task objective or problem description without examples, requiring the model to directly perform inference or generation. Few-shot employs a small number of input-output examples to help the model align task structure with reasoning patterns. CoT requires the model to explicitly generate intermediate reasoning steps before answering to support multi-step causal inference. ToT uses tree-based reasoning to explore multiple causal paths and select the optimal conclusion. HoT further highlights key facts within CoT reasoning chains to enhance transparency and traceability, while SPE optimises reasoning quality by augmenting single attributes within prompts. This paper's experiments select the two most relevant attribute categories from the original paper, contextual logic and structural logic, as augmentation targets. All tasks were employed as evaluation tasks to test each method's performance during reasoning, with results presented in the Figure 5.

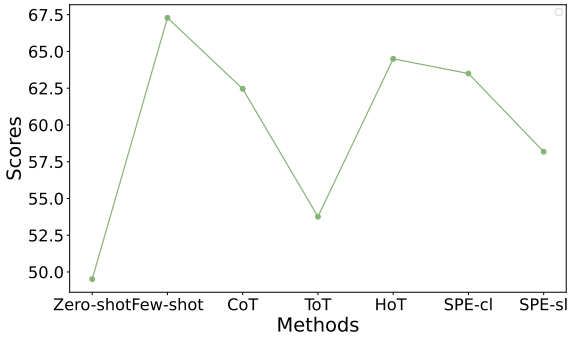


Figure 5: The performance of each prompting method across all evaluation tasks.

Zero-shot Prompting demonstrated the lowest accuracy at 49%, revealing the model's limited reasoning capability without examples or contextual information. In such scenarios, the absence of task context and examples often leads the model to generate outputs inconsistent with task requirements, resulting in inaccurate reasoning. In contrast, Few-shot Prompting achieved the best performance among all methods, with an accuracy of 71%, indicating that providing task context through a small number of examples can significantly enhance the model's reasoning capabilities. A high-quality example enables the model to directly comprehend the task's generation template, thereby reducing errors and biases during causal reasoning. The CoT method achieved a relatively strong accuracy of 62%. Through stepwise reasoning, CoT effectively avoids jumping directly to conclusions, minimizing errors. However, when handling complex reasoning, CoT's approach is comparatively conservative, failing to match the high accuracy of Few-shot. The HoT method achieved the highest accuracy of 64.5% in the PJ category, though its performance declined in the AJ and CVA categories. This indicates that, in certain scenarios, while enhancing the transparency and traceability of the reasoning process, the added detail did not consistently yield beneficial effects across all tasks (this aspect of the experimental results was not fully demonstrated). The ToT method employs a multi-path structure in reasoning, yet its accuracy stands at 53%, suggesting that excessive path selection may increase reasoning complexity, thereby impacting final outcomes, particularly for relatively straightforward tasks. SPE achieved accuracy rates

of 63.5% and 58%. This method enhances causal reasoning quality by optimizing specific prompt attributes (contextual logic and structural logic selected in this experiment). However, due to potential issues with prompt configuration, it failed to significantly improve reasoning performance, falling short of the anticipated efficiency gains. Overall, Few-shot Prompting demonstrated the most favorable performance. Consequently, we selected Few-shot Prompting as the optimal strategy for subsequent causal inference tasks. Details of the ablation study are provided in Appendix E.

3.5 Evaluation For Two Challenges

We designed an evaluation method based on task differences to measure the extent of causal chain errors and the model's reliance on empirical inferences.

To better assess the severity of causal chain errors, we selected the discrepancy between the AJ and PJ tasks as a key metric. In the AJ task, the model must determine whether a causal relationship is correct, whereas in the PJ task, it must judge whether the causal chain itself is correct. Should the model correctly judge the AJ task but incorrectly judge the PJ task, this indicates that while the model has successfully identified the causal relationship, it has failed to correctly deduce the sequence or structure of the causal chain. By calculating the discrepancy between the results of the AJ and PJ tasks, we can effectively evaluate the degree of error in the model's causal chain construction process. Therefore, the discrepancy between AJ and PJ serves as a crucial metric to assess causal chain errors. We did not utilize the OP task for comparison, as it directly employs original sentences from news articles, resulting in scores nearly reaching 100% and exhibiting severe overfitting. This suggests that directly using the model's training data, such as the publicly available 2024 news dataset, as a causal test dataset is not advisable, as the model can easily rely on the data to generate answers. Furthermore, the CVA task was excluded from comparison because it focuses more on determining answer correctness with less emphasis on causal chain reasoning, thus failing to adequately reflect causal chain errors.

To gauge the model's reliance on empirical inferences., we selected the average difference across the VCR and PS, PU, ID tasks as the key metric. In the PS, PU, and ID tasks, entities are real-world entities susceptible to large models' entity bias. Conversely, in the VCR task, real entities are replaced with dummy words, forcing the model to rely solely on causal reasoning to draw conclusions, free from entity bias interference. Consequently, the discrepancy between VCR and PS, PU, ID tasks serves as a crucial metric for evaluating empirical dependency. This approach enables clear comparison of a model's reasoning performance when handling real entities versus virtual tokens, thereby assessing whether it over-relies on training data during inference.

The results of these two experiments are shown in Table 2. The degree of causal chain errors was higher than that of reliance on empirical inferences, with only Grok-4 exhibiting no causal chain errors. The absence of causal chain errors in Grok-4 may stem from its comparatively robust reasoning capabilities and comprehension of complex relationships, underscoring the decisive role of deep logical reasoning in constructing reliable causal chains. Through this design, we are able to quantify the performance of models in causal reasoning tasks and provide valuable reference points for the future optimization of causal reasoning models.

Model	Challenge 1	Challenge 2
Claude-Sonnet-4.5	22.81	12.64
GPT-5.1	29.54	5.07
Grok-4	-5.12	5.20
DeepSeek-V3.1	36.11	1.92
GPT-5	36.73	7.09
Gemini-3-pro-preview	30.82	3.78
Llama-4-Maverick	23.30	2.98

Table 2: Evaluation for two challenges in Casuality Check-CP. Challenge 1 indicates an error in the causal chain, while Challenge 2 indicates reliance on empirical inference.

4 Conclusion

In this work, we propose CausalityCheck, a novel framework for evaluating causal reasoning in large language models across two dimensions: task generalisation and reasoning robustness. Tested with the CausalityCheck-CP dataset, which includes a new virtual causal reasoning task, the results reveal two key challenges in current models: errors in causal chain identification and reliance on empirical inference. Many models produce correct answers but misidentify causal chains, and they often rely on training data knowledge instead of

deep causal reasoning. These challenges highlight significant limitations in the models' causal reasoning abilities. Future research should focus on improving causal chain identification and reducing dependence on empirical inference for deeper reasoning.

5 Limitation

Although this study proposes the CausalityCheck evaluation framework and develops the CausalityCheck-CP dataset, several areas remain under-explored. Firstly, compared to structural causal reasoning methods focused on explicit graph construction, the current approach offers limited solutions to the problem of misidentified causal chains. Future research could explore reducing model dependency on entity memory through causal DAG supervision and neuro-symbolic hybrid models. Secondly, whilst this study references several benchmarks and checklists, it does not delve deeply into benchmarks for structural rigor and counterfactual reasoning, nor how to reduce reliance on shortcut patterns. Integrating debiasing techniques such as counterfactual analysis and multi-agent evaluation could further enhance the depth and robustness of causal reasoning. Future work should focus on improving the generalization capability and reasoning accuracy of causal inference tasks through these debiasing approaches.

Acknowledgments

This work was funded by National Natural Science Foundation of China (Grant No. 62366036), Outstanding Youth Fund Project of Inner Mongolia Autonomous Region (Grant No. 2025JQ010), Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (Grant No. NJYT24033), Major Science and Technology Projects of Inner Mongolia Autonomous Region (Grant No. 2025ZDSF0029), Key R&D and Achievement Transformation Program of Inner Mongolia Autonomous Region (Grant No. 2025YFDZ0011, 2025YFDZ0026, 2025YFSH0021, 2025YFHH0073), Hohhot Science and Technology Project (Grant No. 2023-Zhan-Zhong-1).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio T. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4: A 14-billion parameter language model with synthetic data training. https://www.microsoft.com/en-us/research/wp-content/uploads/2024/12/P4TechReport.pdf.

Meta AI. 2025. Llama 4 maverick. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Llama 4 Scout and Llama 4 Maverick were released as part of the Llama 4 model family.

Mistral AI. 2023. Mixtral-8x7b instruct: Sparse mixture of experts language model. https://mistral.ai/news/mixtral-of-experts.

Anthropic. 2025. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Hongye Cao, Fan Feng, Tianpei Yang, Jing Huo, and Yang Gao. 2025. Causal information prioritization for efficient reinforcement learning. *arXiv preprint arXiv:2502.10097*.

Meiqi Chen, Bo Peng, Yan Zhang, and Chaochao Lu. 2024a. CELLO: Causal evaluation of large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22353–22374, Miami, Florida, USA. 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. 2024b. Causal evaluation of language models. *Preprint*, arXiv:2405.00622.

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

Google DeepMind. 2025. Gemini 2.5 flash (google ai model). https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash.

DeepSeek. 2025a. Deepseek-v3.1 release. https://api-docs.deepseek.com/news/news250821.

DeepSeek. 2025b. Deepseek-r1 release and technical report. https://api-docs.deepseek.com/news/news250120.

Xuan Long Do, Duy Dinh, Ngoc-Hai Nguyen, Kenji Kawaguchi, Nancy Chen, Shafiq Joty, and Min-Yen Kan. 2025. What kes a good natural language prompt? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5835–5873.

Google. 2025. Gemini 3 pro preview. https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-3-pro-preview.

Thilo Hagendorff and Sarah Fabi. 2025. Beyond chains of thought: Benchmarking latent-space reasoning abilities in large language models. *arXiv preprint arXiv:2504.10615*.

Liang He, Yougang Chu, Zhen Wu, Jianbing Zhang, Xinyu Dai, and Jiajun Chen. 2025. Rethinking relation extraction: Beyond shortcuts to generalization with a debiased benchmark. *arXiv preprint arXiv:2501.01349*.

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: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. 2024. LLMs are prone to fallacies in causal inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10553–10569, Miami, Florida, USA. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Tian Lan, Jiang Li, Yemin Wang, Xu Liu, Xiangdong Su, and Guanglai Gao. 2025a. F^2bench: An open-ended fairness evaluation benchmark for llms with factuality considerations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2031–2046.

Tian Lan, Xiangdong Su, Xu Liu, Ruirui Wang, Ke Chang, Jiang Li, and Guanglai Gao. 2025b. Mcbe: A multi-task chinese bias evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6033–6056.

Iok Tong Lei, Ziyu Zhu, Han Yu, Yige Yao, and Zhidong Deng. 2025. Hint of pseudo code (hopc): Zero-shot step by step pseudo code reasoning prompting. In *International Conference on Neural Information Processing*, pages 521–535. Springer.

Naiming Liu, Richard Baraniuk, and Shashank Sonkar. 2025a. Clear-3k: Assessing causal explanatory capabilities in language models. *Preprint*, arXiv:2506.17180.

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. 2025b. 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.

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.

OpenAI. 2024. Openai o1 system card (technical report). https://arxiv.org/abs/2412.16720.

OpenAI. 2025a. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf. Official technical report.

OpenAI. 2025b. Gpt-5.1: A smarter, more conversational chatgpt. https://openai.com/index/gpt-5-1/.

OpenAI. 2025c. Openai o3-mini: Cost-efficient reasoning model. https://openai.com/index/openai-o3-mini/.

Alibaba / Qwen. 2025. Qwen3-30b-a3b and the qwen3 model family. https://en.wikipedia.org/wiki/Qwen.

Bernhard Schölkopf. 2022. Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 765–804.

THUDM. 2025. Glm-4-9b: Open source large language model. https://github.com/thudm/GLMâĂŚ4.

Lei Wang and Yiqing Shen. 2024. Evaluating causal reasoning capabilities of large language models: A systematic analysis across three scenarios. *Electronics*, 13(23):4584.

Yiwei Wang, Bryan Hooi, Fei Wang, Yujun Cai, Yuxuan Liang, Wenxuan Zhou, Jing Tang, Manjuan Duan, and Muhao Chen. 2023. How fragile is relation extraction under entity replacements? *arXiv preprint arXiv:2305.13551*.

Zeyu Wang. 2024. Causalbench: A comprehensive benchrk 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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Abraham Itzhak Weinberg, Cristiano Premebida, and Diego Resende Faria. 2025. Causality from bottom to top: A survey. *Machine Learning*, 114(11):234.

xAI. 2025a. Grok 4. https://x.ai/news/grok-4.

xAI. 2025b. Grok 4 fast reasoning. https://www.infoq.com/news/2025/09/xai-grok4-fast/.

A. Yang and 1 others. 2025. Qwen3: Large language models with unified thinking and non-thinking modes. https://arxiv.org/abs/2505.09388.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *arXiv preprint*.

Bohan Yu, Wei Huang, and Kang Liu. 2026. Sr-ki: Scalable and real-time knowledge integration into llms via supervised attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34486–34494.

Yanxi Zhang, Xin Cong, Zhong Zhang, Xiao Liu, Dongyan Zhao, and Yesai Wu. 2025. Hcr-reasoner: Synergizing large language models and theory for human-like causal reasoning. *Preprint*, arXiv:2505.08750.

Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. 2024. Is your model really a good math reasoner? evaluating mathematical reasoning with checklist. *arXiv preprint arXiv:2407.08733*.

A Heatmap Of CasualityCheck-CP

The specific data in the Figure 4 show in the Figure 6.

B Related Work

B.1 Causal Reasoning in Current LLMs

In recent years, large language models (LLMs) have made significant strides in natural language understanding and reasoning (Hagendorff and Fabi, 2025), yet their causal reasoning capabilities remain limited (Joshi et al., 2024). LLM reasoning typically relies on statistical correlations and pattern matching, lacking a profound grasp of causal relationships (Ma, 2025). Although these models can generate seemingly plausible chains of causal inference, they often base these on associative patterns extracted from training data rather than a genuine understanding of causal logic.Recent research (Chi et al., 2024) indicates that LLMs primarily operate at the correlation level when handling causal reasoning, being capable of identifying associations between variables but unable to grasp the causal relationships within causal chains. To address this shortcoming, researchers are exploring the integration of causal reasoning into model training processes, particularly within multimodal learning (Chen et al., 2024a) and reinforcement learning (Cao et al., 2025) domains. This approach aims to enhance causal learning through environmental feedback and reward mechanisms. Concurrently, novel causal reasoning evaluation frameworks such as CausalBench (Wang, 2024) and ExpliCa (Miliani et al., 2025) are emerging. These aim to comprehensively assess models' performance in causal reasoning tasks, focusing on the depth of causal understanding and the accuracy of the reasoning process. Causality is commonly divided into two categories: type causality and actual causality (Zhang et al., 2025). This paper addresses the latter.

In recent years, large language models (LLMs) have made significant strides in natural language understanding and reasoning (Hagendorff and Fabi, 2025), yet their causal reasoning capabilities remain limited (Joshi et al., 2024). LLM reasoning typically relies on statistical correlations and pattern matching, lacking a profound grasp of causal relationships (Ma, 2025). Although these models can generate seemingly plausible chains of causal inference, they often base these on associative patterns extracted from training data rather than a genuine understanding of causal logic. Recent research

Figure 6: Visualized heatmap of CausalityCheck-CP

(Chi et al., 2024) indicates that LLMs primarily operate at the correlation level when handling causal reasoning, being capable of identifying associations between variables but unable to grasp the causal relationships within causal chains. To address this shortcoming, researchers are exploring the integration of causal reasoning into model training processes, particularly within multimodal learning (Chen et al., 2024a) and reinforcement learning (Cao et al., 2025) domains. This approach aims to enhance causal learning through environmental feedback and reward mechanisms. Concurrently, novel causal reasoning evaluation frameworks such as CausalBench (Wang, 2024) and ExpliCa (Miliani et al., 2025) are emerging. Beyond causal reasoning benchmarks, recent evaluation studies have also emphasized the need for more comprehensive assessment settings for LLMs, such as multi-task and culturally grounded bias evaluation, as exemplified by McBE (Lan et al., 2025b). In addition, F2Bench (Lan et al., 2025a) extends this trend by proposing an open-ended fairness benchmark with factuality considerations, highlighting the importance of evaluating model outputs in more realistic and context-sensitive scenarios. These efforts reflect a broader trend toward more comprehensive evaluation of LLM reasoning, focusing on the depth of understanding and the reliability of model outputs across different reasoning settings. Causality is commonly divided into two categories: type causality and actual causality (Zhang et al., 2025). This paper addresses the latter.

B.2 Virtual Causal Reasoning

The motivation for the Virtual Causal Reasoning (VCR) task stems from the possibility that entity names within datasets may cause models to rely on entity memorization rather than contextual reasoning. Consequently, VCR mitigates this issue by substituting these entities with dummy words. The prototype of this task involves replacing entity names in the dataset with other entities (Wang et al., 2023), thereby evaluating the model's generalization capability. Subsequent research has further developed entity replacement strategies, offering more systematic evaluation methods by breaking spurious correlations between entities and labels and constructing bias-free benchmarks (He et al., 2025). Such research emphasizes that entity information can lead models to learn contextually inconsistent shortcut patterns. However, as pre-training data volumes and model scales increase, substituted entities may still be familiar to the model. Hence, we propose VCR, which does not rely on actual entity information but replaces entities in the problem with virtual tokens, thereby eliminating interference from real-world knowledge learned during model training. Under this setting, the model must rely solely on its reasoning capabilities and learned causal relationships for judgment, rather than retrieving existing information from training data.

C Evaluated Models and Experimental Settings

To systematically evaluate the causal reasoning capabilities of contemporary large language models, we conduct a comprehensive assessment of 17 representative models. These models can be categorized into three groups:

(a) General-Purpose LLMs. These models exhibit broad applicability, strong overall capabilities, and typically adopt large-scale closed-source architectures. This category includes Claude-Sonnet-4.5 (released on 2025-09-29) (Anthropic, 2025), GPT-5.1 (released on 2025-11-13) (OpenAI, 2025b), Grok-4 (xAI, 2025a), DeepSeek-V3.1 (DeepSeek, 2025a), GPT-5 (OpenAI, 2025a), Gemini-3-Pro-Preview (Google, 2025), and Llama-4-Maverick (AI, 2025).

(b) Reasoning-Enhanced LLMs. These models significantly strengthen chain-of-thought, long-form reasoning, and deliberate reasoning through pre-training strategies or inference-time optimization. This group includes O1 (OpenAI, 2024), O3-mini (OpenAI, 2025c), DeepSeek-R1 (DeepSeek, 2025b), Qwen3-30B-A3B-Thinking (Qwen, 2025), Gemini-2.5-Flash-Thinking (DeepMind, 2025), and Grok-4-Fast-Reasoning (xAI, 2025b).

(c) Lightweight LLMs. This category consists of open-source, parameter-efficient models that serve as effective baselines for causal reasoning tasks. It includes Phi-4 (Abdin et al., 2024), Qwen3-8B (Yang et al., 2025), GLM-4-9B (THUDM, 2025), Dream-v0-Instruct-7B (Dream-7B) (Ye et al., 2025), and Mixtral-8×7B-Instruct (AI, 2023). Notably, Dream-v0-Instruct-7B follows a *discrete diffusion & iterative denoising* paradigm rather than a traditional autoregressive architecture; meanwhile, Mixtral-8×7B-Instruct employs a Mixture-of-Experts structure that activates only $\tilde{1}$3B parameters per inference, enabling strong performance while maintaining lower inference cost.

For all dataset construction procedures and evaluation tasks, we adopt accuracy as the primary evaluation metric. We employ a few-shot prompting setup to enhance models' adherence to task-specific instructions. The dataset design intentionally simplifies certain complex problem settings, thereby reducing the overall difficulty of the evaluation.

For all resource-constrained models, we use their default hyperparameter configurations, setting the sampling *temperature* to 0 and the *maximum number of generated tokens* to 1024. Similarly, for all open-source models, we standardize the inference configuration across experiments by setting *do_sample* = False, *max_gen_len* = 512, and *temperature* = 0.1.

D Ensuring Annotation Data Quality

To ensure annotation consistency and accuracy, we employ inter-annotator reliability metrics. Annotator reliability refers to the consistency of annotation results when different annotators work on the same dataset. By evaluating inter-annotator reliability, we can determine the reliability of gold labels and the subjectivity of annotators. A commonly employed metric is Cohen's Kappa coefficient, which quantifies agreement between two annotators on a scale ranging from -1 (complete disagreement) to 1 (perfect agreement). An Kappa value exceeding 0.8 indicates consistent annotation outcomes between annotators.

To enhance annotator consistency, rigorous training is essential to ensure all annotators comprehend annotation standards and adhere to uniform rules throughout the annotation process. Furthermore, a regular feedback mechanism is indispensable. Annotators may encounter challenging cases during annotation, and such a mechanism helps resolve uncertainties and standardise practices. For instance, if Annotators A and B interpret certain category definitions differently, weekly annotation review meetings facilitate further discussion and confirmation of these definitions, thereby minimising bias.

Model	OP	AJ	PJ	CVA
PS	0.92	0.72	0.93	0.62
PU	0.93	0.83	0.85	0.59
ID	0.83	0.78	0.84	0.65
VCR	0.89	0.80	0.81	0.65

Table 3: Kappa coefficient for different tasks.

Judging by the Cohen's Kappa coefficients in the table 3, the annotation consistency across tasks is generally high. The Kappa values for OP, PJ, and AJ all fall within the excellent range, indicating a high degree of agreement among annotators for these three tasks, a reliable annotation process, and clear annotation standards. Conversely, the relatively low Kappa value for CVA reveals some inconsistency among annotators, likely attributable to insufficiently clear annotation criteria. Overall, the Kappa values for most tasks exceeded 0.8, demonstrating good annotation consistency.

To ensure the accuracy and reliability of gold labels, clearly defining annotation standards and category boundaries is paramount. During the annotation process, each conclusion must be categorised based on the rigour and reliability of its reasoning. Figure 14 presents the results for four distinct CVA tasks. Firstly, conclusions labelled correct and reliable are not only factually accurate but also maintain consistency across varying contexts through rigorous causal reasoning that accounts for all possible variables and potential biases. For instance, in certain AI reporting cases, selecting increased media coverage as the answer was analysed as meeting causal criteria and holding true across varied contexts, thus earning a reliable label.

Conversely, some conclusions appear correct yet rely on oversimplified reasoning or overlook critical variables. These are categorised as correct but prone to error. For instance, when discussing the state of AI media coverage, one perspective posits that it is driven by public interest and prioritises sensationalism over accuracy. While this conclusion holds true in certain instances, it overlooks multidimensional factors such as depth and objectivity, thus exhibiting a bias in its generalisability.

Furthermore, certain conclusions may appear erroneous, yet their reasoning, assumptions, or data may be close to correct—valid only in specific contexts or exhibiting significant bias. Such conclusions fall under the category of wrong, but leaning towards right. For instance, when discussing why experts advocate altering AI reporting, one conclusion posits that current coverage is unbalanced and fuels sensationalism. Although this conclusion itself is incorrect, it reflects a prevailing trend in current coverage and holds validity in certain contexts, hence being labelled incorrect but leaning towards correct.

However, certain conclusions are not only incorrect but also lack rigorous reasoning or rely on

unreliable assumptions and data. Such conclusions are categorised as incorrect and unreliable. For instance, when analysing whether journalists increased AI coverage in response to expert calls, selecting they increased media coverage as the conclusion is erroneous. This overlooks the requirement for balanced reporting and lacks in-depth analysis of coverage content. Such conclusions lack reliability and risk misdirection, hence being labelled unreliable.

These classification criteria assist annotators in clearly distinguishing between different types of conclusions, ensuring the accuracy and reliability of gold labels. This not only enhances the quality of annotated data but also provides a robust foundation for model training, enabling models to make more precise and dependable judgements when confronted with diverse reasoning and conclusions.

E Details In The Ablation Study

The following supplementary details provide a more comprehensive exposition of how each prompting strategy is implemented, alongside their consistency with the task and relationship to the test distribution:

Few-shot examples and their overlap with the test distribution. In our experiments, we employed few-shot prompting to augment the model's reasoning capabilities. Few-shot examples assist the model in understanding task structure by providing a small number of instances, thereby reducing errors and biases in reasoning. Each few-shot example comprises the following components: **Problem Description.**Articulates the core elements and objectives of the causal reasoning task. **Input-Output Pairs.** Presents typical examples of causal reasoning tasks, where inputs include the question and relevant contextual information, and outputs represent the model's inferred results.

However, the selection and design of few-shot examples must be contrasted and optimised against the test distribution. To ensure representativeness, we selected representative task types from actual test data, ensuring these examples encompassed task diversity and complexity. For instance, in tasks like Process Judging, we ensured provision of different types of causal chains, along with corresponding correct and incorrect reasoning approaches for each chain. In this manner, the few-shot examples not only enhance the task's reasoning consistency but also avoid biases overly constrained by the test data.

CoT content control. In the CoT approach, we require the model to generate intermediate steps of its reasoning process to enhance the depth of its reasoning. By guiding the model through step-wise reasoning, the CoT approach helps prevent leaps to erroneous conclusions while enhancing transparency and traceability in causal reasoning. To effectively control CoT content, we employ the following specific methods: **Limiting the number of reasoning steps.** To ensure clarity and coherence in CoT reasoning, we cap the number of steps per inference to prevent the generation of verbose and incomprehensible chains. In experiments, the number of CoT steps is typically capped at 3 to 5. **Defining reasoning objectives.** To guide the model in generating effective intermediate steps, we explicitly label the reasoning objective in each example, such as derive the causal chain or validate the plausibility of the causal relationship. **Content consistency.** To prevent inconsistent steps or content deviating from the task objective during reasoning, we require the model to strictly focus on the core causal relationships of the current task at each reasoning step, avoiding irrelevant or secondary information.

Inter-task consistency. In experiments, we ensured consistency across different tasks, particularly when employing distinct prompting strategies. To achieve this, we adopted the following methods: **Uniform Task Definition.** All tasks adhere to consistent definitions and formats to prevent significant structural or objective deviations. This encompasses input formats (questions and contextual information), output formats (inference results), and evaluation criteria for each task. **Cross-Task Prompt Sharing.** To further enhance consistency, we designed a universal prompt template ensuring uniform application of prompting strategies across tasks. For instance, across all tasks, few-shot examples consistently include problem descriptions and input-output pairs, while CoT examples feature step-by-step guidance through reasoning processes.**Consistent evaluation criteria.** All tasks employ unified assessment standards, particularly regarding reasoning accuracy, coherence of reasoning processes, and task generalization capabilities. These uniform metrics enable comparative analysis of reasoning performance across different tasks and prompting strategies.

Through these specific details, we ensure the validity of prompting strategies across diverse tasks

and test distributions within ablation experiments, while also guaranteeing the comparability and consistency of experimental outcomes. These measures enhance the efficacy of prompting strategies for causal reasoning tasks and provide clear directions for future optimization.

F Data Filtering Process

After generating the dataset using CausalityCheck on GPT-5, we filtered 25% of the data. First, we used Python code to remove entries missing key fields. These include data missing the `context`, `question`, `choice`, and `answer` fields, as well as entries where these fields contain incomplete information. The second step involved manually inspecting each entry to ensure it met the criteria, provided a valid causal relationship, and that the relationship could be clearly reflected in the available options. Additionally, we removed any data containing inappropriate content, such as violent themes. For example:

> *Aviva Siegel, a released Israeli hostage, is pleading for the release of her husband, Keith, who is still being held by Hamas in Gaza. Aviva describes the brutal conditions she and Keith endured during their captivity, including physical abuse and sexual assault of female hostages. She is calling on international mediators, such as the US, Qatar, and Egypt, to do more to secure the release of her husband and other hostages.*

Although this data passed GPT-5's content management policy, each LLMs has different thresholds, and such content may not be accepted by other models. To avoid this, we removed any data that contained even a slight amount of related content. The final dataset used in this paper is the CausalityCheck-CP dataset.

G Discussion Of Data Bias

Whilst acknowledging that large language models (LLMs) may introduce inherent biases during data generation, we consider such biases acceptable within this study and do not believe they undermine the conclusions or validity of CausalityCheck. The core motivation of CausalityCheck is to evaluate causal reasoning capabilities across multiple dimensions, thereby revealing language models' causal reasoning characteristics more comprehensively. Data rewriting constitutes a pivotal step in this process, which may be performed either manually or automatically by LLMs. Whilst we acknowledge that expert involvement in data rewriting may represent the most impartial approach, its applicability is constrained by substantial costs, rendering it insufficiently scalable in practice. Therefore, to enhance the scalability and practicality of our research, we selected GPT-5 as the rewriting tool. This choice stems from its recognition as the state-of-the-art, publicly available model at the time, capable of delivering rewriting quality approaching professional standards. It is noteworthy that while LLMs played a pivotal role in data generation, we did not rely solely on automated rewriting. All generated content underwent manual verification to prevent unnatural outputs or ambiguous causal relationships. Furthermore, although GPT-5-generated data was employed in the experiments, it did not confer a significant advantage to the GPT series models. As Table 4 demonstrates, Claude-4.5-sonnet actually performed more impressively in the experiments, indicating that biases introduced during the rewriting process did not decisively influence the final experimental outcomes.

Model	All
Claude-Sonnet-4.5	**71.43**
GPT-5.1	70.92
Grok-4	70.62
DeepSeek-V3.1	69.33
GPT-5	67.41
Gemini-3-pro-preview	65.23
Llama-4-Maverick	65.19

Table 4: Performance of General-Purpose LLMs on ALL Metrics.

H The implications of negative inferential outcomes

Overall, all models demonstrated relatively stable performance in the process judgment (PJ) task, with larger-scale models generally exhibiting superior results. Furthermore, task generalization had a minimal impact on causal chain identification, indicating the models' ability to consistently extract causal chains across varying contextual settings. To delve deeper, we conducted a detailed examination of Claude-Sonnet-4.5's outcomes in the VCR-PJ and PS-PJ tasks depicted in Figure 4, uncovering

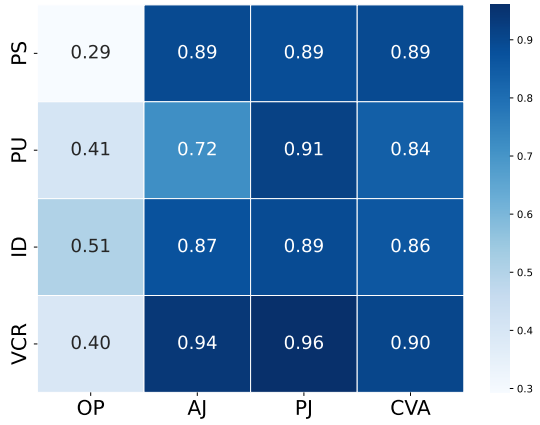


Figure 7: Manual annotation pass rate for OP, AJ, PJ, and CVA.

a critical distinction. Specifically, Claude-Sonnet-4.5 achieved an accuracy rate of 93% in this task, whereas other models maintained accuracy around 50%. To ensure the reliability of this phenomenon, we replicated the experiment three times within this task, with consistently identical results indicating this outcome was not coincidental.

In the PS-PJ task, the task context described negative reports concerning artificial intelligence (AI). Although these reports themselves were objective, reflecting public concerns about AI's rapid advancement, the causal inference outcome should have been negative. However, Claude-Sonnet-4.5 erroneously avoided this inference, failing to attribute the negative outcome to the causal chain. Conversely, in the VCR-PJ task, when AI was replaced with a fictional term (e.g., SC), the model correctly performed the inference. This suggests that when confronted with specific entities, particularly when describing their negative coverage, the model may avoid inferring negative outcomes. As the model had never encountered the fictional vocabulary in the VCR-PJ task, it relied solely on contextual information for inference, yielding the correct causal reasoning outcome. This phenomenon reveals that Claude-Sonnet-4.5 may be susceptible to avoidance of negative information when processing causal reasoning for specific entities, an effect that extends beyond the influence of the prompt itself.

I Double Prompt Effect

Upon further examination of the manually annotated dataset, we observed a significant phenomenon known as the Double Prompt Effect. This effect refers to the marked improvement in generated data quality when the model is simultaneously instructed to produce both correct and erroneous information during the prompting process. Figure 7 illustrates the pass rates across different tasks (OP, AJ, PJ, CVA) in the manually annotated dataset, with OP's pass rate notably lower than those tasks generating both correct and incorrect data. In the AJ, PJ, and CVA tasks, by comparing and correcting potential errors or ambiguities during generation, the model gains a clearer understanding of the task structure, thereby producing more accurate and structurally coherent data. This approach prompts the model to self-correct when encountering potential ambiguities or errors during generation, thereby reducing vagueness and inconsistencies in causal relationships. In contrast, when relying solely on prompts containing correct information (as in the OP task), the model tends to generate overly simplistic or ambiguous results. This manifests as contentious answers within the options, leading to diminished reasoning quality and hindering the effective testing and verification of causal relationships. To conduct more rigorous validation, we compared pass rate improvements when using versus not using double prompts under matched seed questions. Specifically, we designed two experimental groups for the OP task: one utilizing a single correct information prompt, and another employing double prompts (simultaneously generating both correct and incorrect information). We measured the effectiveness of dual prompts by calculating the pass rate for each experimental group. Experimental results demonstrate that the group employing dual-output prompts achieved significantly higher pass rates across multiple tasks compared to the single-prompt group, averaging over 40% higher. This substantial improvement indicates that when models can contrast and correct potential errors or ambiguities during generation, the accuracy and consistency of the reasoning process are effectively enhanced. Consequently, the experimental findings further validate the efficacy of dual prompts. The dual-prompt strategy markedly

Model	PS-OP	PU-OP	ID-OP	VCR-OP
Single	0.29	0.41	0.51	0.40
Double	0.82	0.89	0.92	0.85

Table 5: Comparison of double prompt and single prompt.

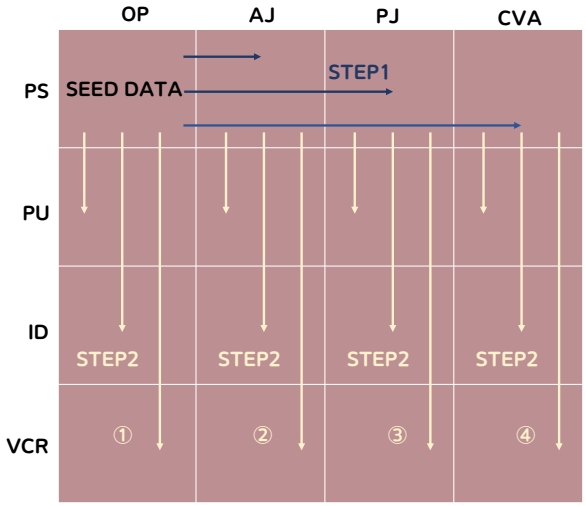


Figure 8: CausalityCheck-CP Construction Processes.

improves the quality and verifiability of generated data, demonstrating its crucial role in constructing causal datasets and enhancing the precision and robustness of causal reasoning tasks.

J CausalityCheck-CP Construction Processes

The construction of the CausalityCheck-CP[1] , as illustrated in Figure 8, follows a two-step process. Step 1 horizontally expands the original problem (PS-OP) to create PS-AJ, PS-PJ, and PS-CVA. Step 2 then vertically builds on PS-OP, PS-AJ, PS-PJ, and PS-CVA to form the complete dataset. Notably, the robustness extensions for PU, ID, and CVR reasoning in Step 2 are all derived from PS data: entries 2/3/4 in STEP2 retain the context of entry 1, with any necessary additions made as supplements rather than replacements. This approach ensures consistency in the difficulty level across entries, thereby maintaining the fairness of the dataset. Figures 9 to 17 show 36 variants.

[1]According to the licenses from the BBC and The Guardian, CausalityCheck-CP can only be used for non-profit purposes. All rights to the corpora used in CausalityCheck-CP, including copyrights, are owned by the BBC and The Guardian.

PS-OP 1

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects of AI.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answer: 3

PU-OP 2

Context: Journalists are receiving backlash for fueling the excitement around artificial intelligence (AI) and failing to provide precise accounts of its potential and constraints. The growing fascination with AI has resulted in a rise in media attention, prompting some specialists to advocate for more equitable coverage that underscores both the advantages and drawbacks of AI.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answer: 3

ID-OP 3

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in AI. On a related but distinct note, this heightened public awareness has also accelerated the development of new AI-driven diagnostic tools in the healthcare sector, an outcome beyond the scope of journalistic practices.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answer: 3

VCP-OP 4

Context: Journalists are facing criticism for contributing to the hype surrounding synthetica cognita (SC) and not accurately reporting on its capabilities and limitations. The surge in interest in SC has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects of SC.
Question: What is the result of the surge in interest in SC in terms of media coverage?
Choice_1: Increased scrutiny on SC by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of SC.
Choice_3: Greater responsibility on the media to report on SC accurately.
Choice_4: Improved understanding of SC technologies by journalists.
Answer: 3

Figure 9: OP dataset. Gray indicates the data that has been altered. The same settings are applied to Figures 8 through 19.

PS-AJ

5

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects of AI.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answerable: 1

PS-AJ

6

Context: The recent surge in interest in artificial intelligence (AI) has had a profound impact on the tech industry, leading to a significant increase in venture capital funding for AI startups and a race among major corporations to integrate AI into their products. This has also spurred a new wave of academic research focused on machine learning ethics and safety protocols. While the media frequently reports on these industry and academic trends, the focus remains on the technological and economic shifts rather than on the evolution of media practices themselves.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answerable: 0

PU-AJ

7

Context: Reporters are under fire for their part in inflating the buzz around artificial intelligence (AI) and for inaccurately conveying its abilities and shortcomings. The heightened curiosity in AI has caused an upsurge in media focus, with some authorities urging for a more balanced portrayal that reflects both the benefits and drawbacks of AI.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answerable: 1

PU-AJ

8

Context: The growing fascination with artificial intelligence (AI) has strongly influenced the technology sector, causing a substantial rise in venture capital investment in AI startups and sparking a competition among leading companies to incorporate AI into their offerings. This development has also initiated a newer trend of academic research targeting machine learning ethics and safety measures. Although media coverage frequently touches on these industry and academic developments, the primary emphasis remains on technological and economic changes rather than on changes in media reporting itself.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answerable: 0

Figure 10: A part of AJ dataset.

ID-AJ

9

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in AI. Although this falls outside media implications, some tech companies have begun investing in AI-powered tools to enhance their R&D processes as a result of public attention.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answerable: 1

ID-AJ

10

Context: The recent surge in interest in artificial intelligence (AI) has had a profound impact on the tech industry, leading to a significant increase in venture capital funding for AI startups and a race among major corporations to integrate AI into their products. This development has also resulted in the creation of government advisory panels tasked with evaluating the societal impacts of AI. This has also spurred a new wave of academic research focused on machine learning ethics and safety protocols. While the media frequently reports on these industry and academic trends, the focus remains on the technological and economic shifts rather than on the evolution of media practices themselves.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answerable: 0

VCP-AJ

11

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, and with this surge, experts and the public are calling for more balanced reporting. This increased scrutiny and demand for more accurate reporting is pushing media outlets to take greater responsibility in how they cover AI, ensuring they address both its potential and its limitations more thoroughly.
Question: What is the result of the surge in interest in SC in terms of media coverage?
Choice_1: Increased scrutiny on SC by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of SC.
Choice_3: Greater responsibility on the media to report on SC accurately.
Choice_4: Improved understanding of SC technologies by journalists.
Answerable: 1

VCP-AJ

12

Context: The recent surge in interest in Synthetica Cognita (SC) has had a profound impact on the tech industry, leading to a significant increase in venture capital funding for SC startups and a race among major corporations to integrate SC into their products. This has also spurred a new wave of academic research focused on machine learning ethics and safety protocols. While the media frequently reports on these industry and academic trends, the focus remains on the technological and economic shifts rather than on the evolution of media practices themselves.
Question: What is the result of the surge in interest in SC in terms of media coverage?
Choice_1: Increased scrutiny on SC by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of SC.
Choice_3: Greater responsibility on the media to report on SC accurately.
Choice_4: Improved understanding of SC technologies by journalists.
Answerable: 0

Figure 11: A part of AJ dataset.

PS-PJ

13

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, and with this surge, experts and the public are calling for more balanced reporting. This increased scrutiny and demand for more accurate reporting is pushing media outlets to take greater responsibility in how they cover AI, ensuring they address both its potential and its limitations more thoroughly.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answer: 3
Causal Chain: surge in interest in AI → increased media coverage → criticism for hype and inaccuracies → greater responsibility on media to report accurately
Correctness: 1

PS-PJ

14

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, and with this surge, experts and the public are calling for more balanced reporting. This increased scrutiny and demand for more accurate reporting is pushing media outlets to take greater responsibility in how they cover AI, ensuring they address both its potential and its limitations more thoroughly.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answer: 3
Causal Chain: surge in interest in AI → increased media coverage → more articles written by journalists → improved understanding of AI technologies
Correctness: 0

PU-PJ

15

Context: Journalists are under fire for intensifying the excitement surrounding artificial intelligence (AI) without delivering precise reports of its strengths and limits. The heightened interest in AI has spurred more media attention, with some authorities urging for reporting that equally represents the benefits and drawbacks of AI.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answer: 3
Causal Chain: surge in interest in AI → increased media coverage → criticism for hype and inaccuracies → greater responsibility on media to report accurately
Correctness: 1

PU-PJ

16

Context: The media is under fire for exaggerating the excitement around artificial intelligence (AI) and failing to provide truthful evaluations of its strengths and weaknesses. The heightened fascination with AI has led to more extensive media exposure, and certain authorities are urging for media presentations that equitably portray both the benefits and downsides of AI.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answer: 3
Causal Chain: surge in interest in AI → increased media coverage → more articles written by journalists → improved understanding of AI technologies
Correctness: 0

Figure 12: A part of PJ dataset.

ID-PJ 17

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. This intense interest has also led to a surge in technological innovations in AI, although this falls outside the scope of media coverage. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in AI.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answer: 3
Causal Chain: surge in interest in AI → increased media coverage → criticism for hype and inaccuracies → greater responsibility on media to report accurately
Correctness: 1

ID-PJ 18

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in AI. While unrelated to media coverage, numerous tech start-ups have emerged focusing on AI-powered solutions to improve industry efficiency.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: Increased scrutiny on AI by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of AI.
Choice_3: Greater responsibility on the media to report on AI accurately.
Choice_4: Improved understanding of AI technologies by journalists.
Answer: 3
Causal Chain: surge in interest in AI → increased media coverage → more articles written by journalists → improved understanding of AI technologies
Correctness: 0

VCP-PJ 19

Context: Journalists are facing criticism for contributing to the hype surrounding synthetica cognita (SC) and not accurately reporting on its capabilities and limitations. The surge in interest in SC has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects of SC.
Question: What is the result of the surge in interest in SC in terms of media coverage?
Choice_1: Increased scrutiny on SC by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of SC.
Choice_3: Greater responsibility on the media to report on SC accurately.
Choice_4: Improved understanding of SC technologies by journalists.
Answer: 3
Causal Chain: surge in interest in SC → increased media coverage → criticism for hype and inaccuracies → greater responsibility on media to report accurately
Correctness: 1

VCP-PJ 20

Context: Journalists are facing criticism for contributing to the hype surrounding synthetica cognita (SC) and not accurately reporting on its capabilities and limitations. The surge in interest in SC has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects of SC.
Question: What is the result of the surge in interest in SC in terms of media coverage?
Choice_1: Increased scrutiny on SC by journalists.
Choice_2: More balanced reporting on the positive and negative aspects of SC.
Choice_3: Greater responsibility on the media to report on SC accurately.
Choice_4: Improved understanding of SC technologies by journalists.
Answer: 3
Causal Chain: surge in interest in SC → increased media coverage → more articles written by journalists → improved understanding of SC technologies
Correctness: 0

Figure 13: A part of PJ dataset.

PS-VCA

21

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in AI.
Question: What is a direct consequence of the surge in public interest in AI?
Choice_1: A decrease in AI hype.
Choice_2: Increased media coverage..
Choice_3: More accurate reporting on AI's limitations.
Choice_4: An end to criticism of journalists.
Answer: 2
Verdict: Correct and reliable

PS-VCA

22

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects
Question: Why are experts advocating for a change in AI reporting?
Choice_1: Because current reporting is too focused on the negative aspects of AI.
Choice_2: Because the surge in interest has overwhelmed media outlets.
Choice_3: Because journalists are not reporting on AI at all.
Choice_4: Because current reporting is unbalanced and contributes to hype.
Answer: 4
Verdict: Correct but prone to error

PS-VCA

23

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in AI.
Question: What can be inferred about the current state of media coverage of AI?
Choice_1: It predominantly highlights the negative aspects of AI.
Choice_2: It is comprehensive in covering both AI's capabilities and limitations.
Choice_3: It is driven by public interest and focuses more on hype than accuracy.
Choice_4: It has been decreasing due to criticism from experts.
Answer: 3
Verdict: Correct but prone to correctness

PS-VCA

24

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in AI.
Question: What action have journalists taken in response to the experts' call for more balanced reporting?
Choice_1: They have started to highlight both positive and negative aspects.
Choice_2: They have increased their media coverage of AI.
Choice_3: They have begun to report more accurately on AI's capabilities.
Choice_4: The text does not specify any action taken by journalists in response.
Answer: 2
Verdict: Incorrect and unreliable

Figure 14: A part of CVA dataset.

PU-VCA

25

Context: Journalists are being criticized for escalating the buzz around artificial intelligence (AI) and for not providing accurate depictions of its strengths and weaknesses. The heightened interest in AI has led to a boost in media coverage, with some authorities urging for a more balanced approach that addresses both the benefits and disadvantages of AI.
Question: What is the result of the surge in interest in AI in terms of media coverage?
Choice_1: A decrease in AI hype.
Choice_2: Increased media coverage..
Choice_3: More accurate reporting on AI's limitations.
Choice_4: An end to criticism of journalists.
Answer: 2
Verdict: Correct and reliable

PU-VCA

26

Context: Reporters are under fire for their role in amplifying the excitement around artificial intelligence (AI) and their shortcomings in providing accurate reports on its potential and limits. The rising interest in AI has led to a boom in media coverage, with some authorities urging for more unbiased journalism that portrays both the benefits and drawbacks of AI.
Question: Why are experts advocating for a change in AI reporting?
Choice_1: Because current reporting is too focused on the negative aspects of AI.
Choice_2: Because the surge in interest has overwhelmed media outlets.
Choice_3: Because journalists are not reporting on AI at all.
Choice_4: Because current reporting is unbalanced and contributes to hype.
Answer: 4
Verdict: Correct but prone to error

PU-VCA

27

Context: Journalists are being criticized for their role in escalating the excitement surrounding artificial intelligence (AI) and their lack of precise reporting on its true abilities and restrictions. As the interest in AI continues to grow, media coverage has surged, and some experts are advocating for fairer reporting that addresses both the benefits and downsides of AI.
Question: What can be inferred about the current state of media coverage of AI?
Choice_1: It predominantly highlights the negative aspects of AI.
Choice_2: It is comprehensive in covering both AI's capabilities and limitations.
Choice_3: It is driven by public interest and focuses more on hype than accuracy.
Choice_4: It has been decreasing due to criticism from experts.
Answer: 3
Verdict: Correct but prone to correctness

PU-VCA

28

Context: Critics have targeted journalists for their part in heightening the buzz around artificial intelligence (AI) and their inaccurate portrayal of AI's abilities and constraints. As interest in AI increases, media coverage has expanded, with certain experts urging for reporting that equally represents both the benefits and challenges of AI.
Question: What action have journalists taken in response to the experts' call for more balanced reporting?
Choice_1: They have started to highlight both positive and negative aspects.
Choice_2: They have increased their media coverage of AI.
Choice_3: They have begun to report more accurately on AI's capabilities.
Choice_4: The text does not specify any action taken by journalists in response.
Answer: 2
Verdict: Incorrect and unreliable

Figure 15: A part of CVA dataset.

ID-VCA

29

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in AI. This increased focus has, however, also encouraged tech companies to invest in developing AI further, although this falls outside the direct consequences related to media coverage.
Question: What is a direct consequence of the surge in public interest in AI?
Choice_1: A decrease in AI hype.
Choice_2: Increased media coverage..
Choice_3: More accurate reporting on AI's limitations.
Choice_4: An end to criticism of journalists.
Answer: 2
Verdict: Correct and reliable

ID-VCA

30

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in AI. Interestingly, this increased attention has also led to the commercial sector seeing accelerated investments in AI-driven solutions, although this development does not directly pertain to the issues in journalistic reporting.
Question: Why are experts advocating for a change in AI reporting?
Choice_1: Because current reporting is too focused on the negative aspects of AI.
Choice_2: Because the surge in interest has overwhelmed media outlets.
Choice_3: Because journalists are not reporting on AI at all.
Choice_4: Because current reporting is unbalanced and contributes to hype.
Answer: 4
Verdict: Correct but prone to error

ID-VCA

31

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. Although this scrutiny isn't directly related to technological development, the increased attention has led to a rise in investments in AI startups, aiming to capitalize on the growing interest. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in AI.
Question: What can be inferred about the current state of media coverage of AI?
Choice_1: It predominantly highlights the negative aspects of AI.
Choice_2: It is comprehensive in covering both AI's capabilities and limitations.
Choice_3: It is driven by public interest and focuses more on hype than accuracy.
Choice_4: It has been decreasing due to criticism from experts.
Answer: 3
Verdict: Correct but prone to correctness

ID-VCA

32

Context: Journalists are facing criticism for contributing to the hype surrounding artificial intelligence (AI) and not accurately reporting on its capabilities and limitations. The surge in interest in AI has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in AI. This heightened attention has inadvertently led to a rise in investments towards AI-driven technology startups, although this financial consequence is not directly linked to the journalists' actions.
Question: What action have journalists taken in response to the experts' call for more balanced reporting?
Choice_1: They have started to highlight both positive and negative aspects.
Choice_2: They have increased their media coverage of AI.
Choice_3: They have begun to report more accurately on AI's capabilities.
Choice_4: The text does not specify any action taken by journalists in response.
Answer: 2
Verdict: Incorrect and unreliable

Figure 16: A part of CVA dataset.

VCP-VCA

33

Context: Journalists are facing criticism for contributing to the hype surrounding synthetica cognita (SC) and not accurately reporting on its capabilities and limitations. The surge in interest in SC has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in SC.
Question: What is a direct consequence of the surge in public interest in SC?
Choice_1: A decrease in SC hype.
Choice_2: Increased media coverage..
Choice_3: More accurate reporting on SC's limitations.
Choice_4: An end to criticism of journalists.
Answer: 2
Verdict: Correct and reliable

VCP-VCA

34

Context: Journalists are facing criticism for contributing to the hype surrounding synthetica cognita (SC) and not accurately reporting on its capabilities and limitations. The surge in interest in SC has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in SC.
Question: Why are experts advocating for a change in SC reporting?
Choice_1: Because current reporting is too focused on the negative aspects of SC.
Choice_2: Because the surge in interest has overwhelmed media outlets.
Choice_3: Because journalists are not reporting on SC at all.
Choice_4: Because current reporting is unbalanced and contributes to hype.
Answer: 4
Verdict: Correct but prone to error

VCP-VCA

35

Context: Journalists are facing criticism for contributing to the hype surrounding synthetica cognita (SC) and not accurately reporting on its capabilities and limitations. The surge in interest in SC has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in SC.
Question: What can be inferred about the current state of media coverage of SC?
Choice_1: It predominantly highlights the negative aspects of SC.
Choice_2: It is comprehensive in covering both SC's capabilities and limitations.
Choice_3: It is driven by public interest and focuses more on hype than accuracy.
Choice_4: It has been decreasing due to criticism from experts.
Answer: 3
Verdict: Correct but prone to correctness

VCP-VCA

36

Context: Journalists are facing criticism for contributing to the hype surrounding synthetica cognita (SC) and not accurately reporting on its capabilities and limitations. The surge in interest in SC has led to increased media coverage, with some experts calling for more balanced reporting that highlights both the positive and negative aspects in SC.
Question: What action have journalists taken in response to the experts' call for more balanced reporting?
Choice_1: They have started to highlight both positive and negative aspects.
Choice_2: They have increased their media coverage of SC.
Choice_3: They have begun to report more accurately on SC's capabilities.
Choice_4: The text does not specify any action taken by journalists in response.
Answer: 2
Verdict: Incorrect and unreliable

Figure 17: A part of CVA dataset.