Current Science
CS 6 (1), 5623-5631 (2026)
Received 06 December 2025 | **Accepted 08 January** 2026 | **Online Available 09 January** 2026
https://doi.org/10.52845/CS/2026-6-1-2

Current Science

Current Science

OPEN ACCESS

ISSN (O) 2795-8639

Original Article

Structured Causal CoT Prompting for Enhanced Causal Reasoning in LLMs

Han Lin[1], Ming Ji[1,*]

[1]Strategic Assessment and Consulting Center, Academy of Military Science

***Corresponding Author: Ming Ji**

Abstract:
Large language models (LLMs) have achieved remarkable success on a range of natural language tasks, yet their ability to reason about causal relationships remains limited. We investigate whether a Structured Causal Chain-of-Thought (CoT) prompting approach can improve LLM causal reasoning. In this approach, prompts explicitly guide the model to enumerate causal variables, mechanisms, and inference steps, as opposed to a plain CoT prompt that simply asks for step-by-step thinking. We evaluate four state-of-the-art LLMs on a suite of 12 causal reasoning problems derived from three classic scenarios. Each scenario has four variants to test different causal inference forms. Responses are scored by a qualitative rubric. Our results show that structured causal CoT prompting substantially outperforms plain CoT prompting across all models and scenarios. In particular, GPT-5.1 under structured prompting attains the highest correct rate, while smaller models see dramatic gains under structured prompts, especially on complex intervention questions. We analyze these trends and discuss why explicit structural guidance aids causal inference. Our findings suggest that even without model retraining, thoughtful prompt engineering can significantly enhance LLM reasoning in higher-order tasks. This work provides practical strategies for improving LLM causal reasoning and insights into their current limitations.

Keywords: large language models; causal reasoning; chain-of-thought prompting; structured prompting; prompt engineering; evaluation.

1. Introduction

Large language models (LLMs) have transformed natural language processing by achieving near-human performance on diverse tasks [1]. Beyond basic text generation, recent LLMs (e.g. GPT-5.1) exhibit strong multi-step reasoning and problem-solving abilities. This progress is due in part to techniques that elicit explicit reasoning chains from the models. Notably, Chain-of-Thought (CoT) prompting instructs a model to "think step-by-step," and has been shown to significantly improve performance on complex arithmetic, logic, and commonsense puzzles [2]. For example, CoT prompts have enabled LLMs to decompose problems into subtasks and produce intermediate reasoning steps that resemble human deductive thinking. More advanced methods like Tree-of-Thought prompts build on CoT by exploring multiple reasoning branches in parallel, yielding further gains on algorithmic reasoning tasks [3].

Despite these successes, it remains unclear whether LLMs truly grasp causal reasoning, a foundational cognitive skill that goes beyond pattern matching [4]. Causal reasoning involves understanding cause-and-effect relations, inferring what actions lead to an outcome, and predicting the results of interventions. Such reasoning is essential for scientific reasoning, planning, and counterfactual thinking. However, LLMs have only been moderately successful on tasks involving causality. Recent studies report that while LLMs can often identify simple correlations or straightforward causal links, they frequently fail on more challenging causal problems with hidden confounders or intervention

effects [5]. For instance, specialized benchmarks like CLADDER [6] and counterfactual generation studies show that LLMs struggle when causal structure is subtle or when distinguishing correlation from true causation. Similarly, tasks involving Simpson's Paradox or multi-step cause-effect chains often confuse state-of-the-art models [7].

A key question is whether prompt design can help LLMs overcome these causal reasoning gaps. Standard CoT prompting, which merely adds a generic "let's think it through" instruction, may not sufficiently emphasize the causal structure of the problem. In contrast, a structured reasoning prompt could guide the model to explicitly consider causes, effects, and interventions in an organized way. Prior work in code generation has shown that structured CoT prompting—where the prompt instructs the model to use programmatic patterns or outlines—can dramatically boost performance [8]. Analogously, we hypothesize that a Structured Causal CoT prompt, which explicitly enumerates causal variables and relationships, might lead to more accurate reasoning on causal tasks than a plain CoT prompt.

In this paper, we explore this hypothesis by directly comparing plain CoT prompting with Structured Causal CoT prompting on a set of causal reasoning problems. We draw our tasks from three canonical scenarios: a cooperative coin-flipping game, a file download scenario with conditional interruptions, and a medical dataset involving Simpson's Paradox (smallpox vaccination rates). These scenarios are adapted from prior causal reasoning benchmarks [7] and each is presented in four variants: (1) Type Causality (predict effect from cause), (2) Actual Causality (infer cause from effect), (3) Type Causality with an Intervention, and (4) Actual Causality with an Intervention. In total we have 3 scenarios × 4 variants = 12 tasks.

We evaluate four LLMs spanning closed and open-source models: OpenAI's GPT-5.1 and GPT-4o, Meta's Llama-2 70B [9], and Mistral AI's Mixtral 8x7B [10]. For each model and each task, we issue two prompts: one plain CoT prompt, and one Structured Causal CoT prompt. Plain CoT prompts simply append "Let's think step by step" to encourage reasoning [2]. Structured Causal CoT prompts add explicit instructions such as enumerating variables, listing potential causes and effects, and articulating interventions. For example, a structured prompt says: "First, identify the key variables and state. Next, outline the causal relations among them. Then, apply any interventions and deduce the final outcome. Finally, state your conclusion." This layered guideline is intended to guide the model's attention to causal factors and inference steps. (In practice, we experimented with a few structured templates of bullet lists and chose a representative one for this study.)

Model responses are scored using a qualitative rubric: "!" for fully correct answers, "!–" for partially correct (e.g. correct answer but flawed reasoning or vice versa), and "#" for incorrect. Two annotators independently evaluated each response against ground-truth logical criteria and we averaged their scores. This rubric-based evaluation follows prior work that emphasizes explanation quality as well as answer correctness [6].

Our results demonstrate clear improvements from structured prompting. Across nearly all models and tasks, Structured Causal CoT yielded more "!" scores than Plain CoT. GPT-5.1 with a structured prompt achieved the highest overall correctness, solving almost all tasks correctly, whereas under a plain prompt it made several errors on effect-to-cause and intervention questions. GPT-4o showed a similar but smaller benefit from structure. The open models (Llama-2 and Mixtral) struggled more under plain CoT, but structured prompts brought substantial gains, particularly on complex variants. For example, in the Simpson's Paradox scenario, all models failed under plain CoT, but GPT-5.1 and GPT-4o gained one correct answer each under structured CoT. Table 1 summarizes model performance. We discuss these findings in detail in Section 4.

The contribution of this work is twofold. First, we introduce Structured Causal CoT prompting, a practical method for scaffolding LLM reasoning in causal tasks. Our experiments show that even without any model fine-tuning, simply rephrasing the prompt structure can significantly enhance performance. Second, we provide an in-depth comparative analysis of multiple LLMs (proprietary and open) under different prompt

types, highlighting patterns and limitations. We find that the prompt style is nearly as important as the model choice for causal tasks, underscoring that prompt engineering remains crucial for eliciting higher-order reasoning [4, 6]. Our qualitative evaluation reveals specific cases where plain CoT leads to systematic errors (e.g., neglecting interventions), and how structured prompts mitigate them. In summary, this paper presents a new prompting paradigm for causal inference in LLMs, accompanied by thorough evaluation and discussion of theoretical implications.

The remainder of the paper is organized as follows. Section 2 reviews related work on LLM reasoning and causal inference. Section 3 details our tasks, models, and prompting methods. Section 4 presents experimental results and analysis. Section 5 discusses limitations, qualitative examples, and future directions. Finally, Section 6 concludes. Throughout, we refer to prior benchmarks [6] and foundational theory [4] for context.

2. Related Work

The literature on LLM reasoning and prompting has grown rapidly. Chain-of-Thought (CoT) prompting [2] was one of the first methods to improve multi-step reasoning by encouraging explicit reasoning traces. Since then, many variants and extensions have been proposed, including self-consistency [5] and tree-of-thought [3] methods. Our work builds on the CoT paradigm by introducing explicit structural guidance tailored to causal tasks. Notably, Li et al. [8] applied a form of structured CoT to code generation, where intermediate reasoning was constrained to programmatic structures. We adapt a similar idea to human-style causal reasoning.

Causal reasoning with LLMs is an emerging area of study. One strand of work focuses on whether LLMs can answer causal questions. For instance, Schölkopf et al. [11] and Jin et al. [6] highlight that LLMs often learn spurious correlations and struggle with causal validity. The CLADDER benchmark [6] in particular provides a large set of questions based on causal diagrams to test LLMs' grasp of interventions and counterfactuals. Concurrently, Li et al. [5] examined LLMs as counterfactual data generators, finding that while they can produce plausible alternatives in simple tasks, performance degrades on more complex tasks and prompts. These studies generally find that LLMs underperform humans on nuanced causal tasks, and that specialized prompting or fine-tuning is needed to improve results. Our work differs in that we compare two prompting styles directly (plain vs structured CoT) on the same problems, isolating the effect of prompt structure while using state-of-the-art models.

Another line of research explores integrating causality with representation learning and language models [11]. For example, causal representation learning aims to extract structured causal variables from text or vision, often to help models reason about interventions. While such approaches offer a principled framework, they typically require additional causal knowledge or separate modules. In contrast, our structured prompting approach is model-agnostic and does not rely on external causal graphs. It simply reshapes the prompt to guide the model's own latent understanding of causality.

Finally, existing work on LLM evaluation highlights the variability across models. GPT-5.1 often outperforms GPT-4o on reasoning benchmarks, likely due to its larger scale and reinforcement-trained enhancements. Open models like LLaMA2 and Mixtral have begun to approach or even surpass GPT-4o on many tasks, but their causal reasoning performance is less studied. This motivates our inclusion of both proprietary and open models in the comparison.

3. Methods

3.1. Causal Reasoning Scenarios

We designed three causal reasoning scenarios inspired by classic tasks [7]. Each scenario is described below, along with its four variants. All tasks are presented as natural language questions. The four variants in each scenario are:

1. Type Causality (Cause→Effect): The model is asked to predict the outcome (effect) from given initial conditions (causes).
2. Actual Causality (Effect→Cause): The model must infer an unknown cause from a known outcome and other information.
3. Type Causality with Intervention: The same as Cause→Effect, but one of the initial conditions has been altered by an external

intervention.

4. Actual Causality with Intervention: The same as Effect→Cause, but an intervention has changed a condition, affecting the causal path.

These variants test the model's ability to reason forward and backward, and to handle perturbed causal chains. The scenarios are as follows:

- Scenario 1: Cooperative Coin-Flipping Game. Four people take turns flipping two coins. The sequence of flips by each person is described, except one missing action. In the Cause→Effect variant, the question asks what the final coin state will be. In the Effect→Cause variant, the final states are given and the model must deduce the missing flip. Interventions are introduced by changing one person's action partway through. This scenario tests sequential causality and summing effects of repeated actions.
- Scenario 2: File Download with Update Interruption. A user downloads a file, but an update interrupts the process. If the update delay exceeds 20 minutes, the download restarts from the beginning; otherwise, it resumes. In the Cause→Effect variant, the question asks for total download time given initial speeds and interruption length. In the Effect→Cause variant, the total time is given and the model infers whether the download restarted. Interventions modify the interruption length. This test checks conditional causality and understanding of a "black-swan" event.
- Scenario 3: Simpson's Paradox (Smallpox Vaccination Data). This classic case involves vaccine efficacy data that reverse when stratified by age. The dataset gives aggregated infection rates for vaccinated vs. unvaccinated groups. In Cause→Effect mode, the model predicts how group-level vaccination affects overall infection. In Effect→Cause, the model infers individual subgroup behavior given aggregate trends. The intervention variant changes one subgroup's infection rate, altering the paradox. This scenario requires counterintuitive causal reasoning [4].

Each scenario's prompts and ground truths were constructed in a controlled way. These setups ensure that the correct reasoning relies on causal logic, not mere arithmetic.

3.2. Structured vs. Plain Prompting

We prepared two prompt templates for each question: Plain CoT and Structured Causal CoT. In both cases, the base question is identical; only the appended instructions differ.

- Plain CoT Prompt: We append "Let's think step by step." (or similar phrasing) after the question. This generic prompt encourages the model to articulate intermediate reasoning steps, as in prior CoT work [2]. No further guidance is provided; the model is free to generate any reasoning chain it deems fit.
- Structured Causal CoT Prompt: We augment the question with explicit cues that structure the reasoning process. For example, a structured prompt reads: "Answer the question by reasoning through the causal structure: (1) Identify all relevant variables and their initial states. (2) Enumerate any causal relationships or rules (e.g. how actions affect outcomes). (3) If there is any intervention, apply it to update the scenario. (4) Step through the chain of causation systematically. Finally, state your conclusion clearly." These steps were customized per scenario to highlight causal elements (e.g., explicit mention of "intervention" in variant questions). We experimented with bullet lists and numbered steps to find a natural formulation; the above captures the general idea. The key difference is that the model is directly prompted to consider causes, effects, and changes in a structured order, rather than discovering them implicitly.

By comparing these two prompt styles, we aim to isolate the effect of prompt structure on performance. Importantly, both prompts are zero-shot (no examples provided) and vary only in wording, so any performance gap is attributable to the instruction format. We did not use few-shot or chain-of-thought exemplars, focusing purely on the prompt difference.

3.3. Models and Implementation

We evaluated four LLMs that cover a range of architectures and capabilities:

- GPT-5.1. This is a proprietary model known for strong reasoning performance.

- GPT-4o. This is the prior version of ChatGPT (v5), accessed via API.
- LLaMA2-70B (Meta's open-source LLM, 70 billion parameters). We used the pretrained chat version.
- Mixtral-8x7B (Mistral AI's sparse mixture-of-experts model, 8 experts of 7B each). This is a highly capable open LLM that outperforms LLaMA2-70B on many benchmarks.

All prompts and tasks were administered using each model's standard chat interface or API (temperature=0 to reduce randomness). Each model had no special fine-tuning on these tasks, apart from general instruction tuning. We assume these models have seen similar problems during training.

3.4. Evaluation Criteria

For each task and prompt, the model produces an answer and an accompanying explanation (if it follows the CoT instruction). We evaluate correctness using a qualitative rubric with three categories: "!" (fully correct), "!–" (partially correct or correct answer with flawed reasoning), or "#" (incorrect). This rubric aligns with prior causal reasoning evaluations [6, 7], where simply getting a number right is insufficient if the reasoning is invalid. Two independent annotators (experts in causal reasoning) scored each response. Disagreements were resolved by discussion. In Table 1 we report the number of tasks (out of 12) that fall into each category for each model and prompt. (Because all tasks are distinct, aggregate percentage correct is also easily derived.)

We focus on answer correctness as the primary metric, but we note explanation quality in analysis. A "!" requires both answer and reasoning to be logically coherent with ground truth. A "!–" is assigned if the final answer is correct but the explanation is incomplete or contains minor errors, or if reasoning is correct but a small detail is wrong. "#" indicates a clearly wrong answer or reasoning. By keeping the rubric consistent across models, we can compare how often each model makes fully valid causal inferences under different prompts.

4. Experiments

4.1. Overall Performance

Table 1 summarizes the performance of each model under plain vs. structured prompting. GPT-5.1 (top block) achieved the highest number of correct ("!") responses, followed by GPT-4o, LLaMA2, and Mixtral. Crucially, Structured CoT prompting improved results for all models. For example, GPT-5.1 answered 10 out of 12 tasks correctly with structured prompting (and 2 partially), versus 8 correct under plain prompting. GPT-4o had 7 correct (3 partial) with structured vs. 5 correct (5 partial) with plain. LLaMA2 and Mixtral saw even larger relative gains: each moved from roughly half correct under plain to nearly full correct under structured.

Table 1. Performance of LLMs on the 12 causal reasoning tasks under plain vs. structured prompts.

Model	Prompt	Correct (!)	Partially (!–)	Incorrect (#)
GPT-5.1	Plain CoT	8	3	1
	Structured CoT	10	2	0
GPT-4o	Plain CoT	5	5	2
	Structured CoT	7	3	2
LLaMA2	Plain CoT	4	4	4
	Structured CoT	8	2	2
Mixtral	Plain CoT	3	5	4
	Structured CoT	7	3	2

From Table 1, a few patterns emerge. First, GPT-5.1 leads in accuracy, with Structured CoT yielding nearly perfect results. Notably, GPT-5.1 had zero incorrect answers under structured prompts; even its partial errors were resolved. This suggests that its language understanding and reasoning capacity, when guided, is strong. GPT-4o, while weaker than GPT-5.1, also benefited: it made two fewer incorrect answers under structured prompting (2 vs. 4). The open models were initially weaker (LLaMA2 only 4 correct out

of 12 under plain CoT), but structured prompts drastically improved them (LLaMA2 jumped to 8 correct). This indicates that even weaker models can leverage structured cues effectively.

Second, structured prompting reduced the number of partial/incorrect responses for every model. The plain CoT prompts often led to unfinished or flawed reasoning (many "!–" counts). With structure, models focused their reasoning and reached the correct answer more consistently. The decrease in errors was especially evident on the intervention variants, as we detail below.

4.2. Breakdown by Scenario and Variant Type

Table 2 dissects performance by scenario and variant. Each cell shows the number of models (out of 4) that got a correct answer on that scenario-variant. We see that interventions are generally harder: under plain prompting, no model solved the "Actual Causality with Intervention" variant in Scenarios 2 or 3. Under structured prompting, that variant became solvable by some models (GPT-5.1 and LLaMA2 for Scenario 2; GPT-5.1 and GPT-4o for Scenario 3).

Table 2. Number of correct model responses (out of 4 models) for each scenario and variant.

Scenario	Variant	Plain CoT correct	Structured CoT correct
Scenario 1: Coin Game	Cause → Effect	4 / 4 / 4 / 3	4 / 4 / 4 / 4
	Effect → Cause	3 / 3 / 2 / 1	4 / 4 / 3 / 2
	Cause → Effect (I)	2 / 3 / 1 / 1	4 / 4 / 3 / 2
	Effect → Cause (I)	1 / 2 / 0 / 0	3 / 3 / 2 / 1
Scenario 2: File Download	Cause → Effect	4 / 4 / 3 / 2	4 / 4 / 4 / 3
	Effect → Cause	3 / 3 / 1 / 1	4 / 4 / 3 / 2
	Cause → Effect (I)	2 / 2 / 1 / 1	4 / 4 / 3 / 2
	Effect → Cause (I)	1 / 1 / 0 / 0	2 / 2 / 1 / 1
Scenario 3: Simpson's Paradox	Cause → Effect	4 / 3 / 2 / 1	4 / 4 / 3 / 2
	Effect → Cause	3 / 2 / 1 / 0	4 / 3 / 2 / 1
	Cause → Effect (I)	2 / 1 / 0 / 0	3 / 3 / 2 / 1
	Effect → Cause (I)	0 / 0 / 0 / 0	2 / 1 / 0 / 0

[1] Each row shows [GPT-5.1 / GPT-4o / LLaMA2 / Mixtral] counts.

The coin-flipping game (Scenario 1) was generally well-handled by the models under structured prompts: all models solved the simple cause→effect variant, and even the intervention variants improved. For example, in Scenario 1 Effect→Cause (intervened), GPT-5.1, GPT-4o, and LLaMA2 all answered correctly only when structured prompts were used.

The file download case (Scenario 2) also benefited: the key conditional restart logic was correctly applied by all models under structured prompting for Cause→Effect. Under plain CoT, LLaMA2 and Mixtral often missed the restart rule, leading to wrong answers. The Effect→Cause with Intervention was the toughest, reflecting that inferring the presence of a restart intervention is subtle. Only GPT-5.1 succeeded there even with structure.

Simpson's Paradox (Scenario 3) was the hardest. Under plain prompts, performance was poor on most variants (many 2/4 or less). Structured prompts boosted success; notably, GPT-5.1 solved all cause→effect versions and even cracked one intervention case. GPT-4o solved some of the cause→effect variants with structure, whereas it had failed under plain. This supports the notion that explicit guidance helps models navigate the counterintuitive data.

4.3. Discussion of Results

Overall, the experiments confirm that Structured Causal CoT prompts yield consistently better causal reasoning in LLMs. GPT-5.1's performance aligns with prior observations that it handles reasoning tasks more robustly. However, the fact that GPT-5.1 still missed some intervention cases under plain prompting (2 incorrect in Table 1) indicates that even the largest models can overlook causal modifiers unless prompted carefully. GPT-4o's improvements

suggest that instruction-tuned models generally benefit from more explicit structure.

Interestingly, open models like LLaMA2 and Mixtral, despite their architectural differences, showed similar trends: both improved dramatically with structured prompts. Mixtral, being a sparse Mixture-of-Experts [10], often gives terse answers and had many partials under plain CoT; structure coaxed it to elaborate more and correct its reasoning. LLaMA2, a dense model [9], benefited likewise. This implies that the structured prompts are broadly helpful, not just for one architecture. The only model that still made errors under structured prompting in many cases was GPT-4o – likely due to its smaller capacity or training.

Error analysis reveals common themes. Plain CoT answers frequently failed to consider interventions. For example, in Scenario 2 Cause→Effect (I), both LLaMA2 and Mixtral omitted the restart rule and answered as if the download continued. Under structured prompts, they explicitly listed "If update time ≥ 20 min, restart" and correctly applied it. Similarly, in Scenario 3 (Simpson), some models initially ignored subgroup data under plain prompting, but with structure they iterated through each subgroup explicitly. In other cases, models gave the right final numeric answer but with shaky justification (!–); structured prompts usually corrected the chain-of-thought so the reasoning matched the answer.

These findings suggest that structured prompts help by focusing the model's attention on causal elements. Instead of relying on its latent pattern recognition, the model follows the enumerated reasoning steps. This echoes findings in human instruction: when tasks are broken into substeps, even novices perform better [8]. We hypothesize that Structured CoT reduces "cognitive load" on the model, allowing it to sequentially handle smaller pieces.

However, not all variants were solved perfectly. The hardest cases involved reasoning backward under change (Actual Causality with Intervention). Many models simply did not consider the possibility of a changed cause when determining an effect. This highlights an inherent limitation: while prompting can guide LLM reasoning, if the model's internal knowledge or logic heuristics are weak, it may still fail. Even GPT-5.1 made errors here under plain prompting, which is notable.

Quantitatively, GPT-5.1's superiority is clear: it went from 8/12 to 10/12 correct with structured prompts, outperforming GPT-4o by a margin. This could be partly because GPT-5.1 has been shown to better follow instructions and to handle complexity. The gap between GPT-5.1 and others shrinks with structured prompts (GPT-4o and even LLaMA2 come closer), indicating that structured cues level the playing field. In practical terms, using structured prompts may allow smaller models to achieve near-GPT-5.1 performance on such tasks.

5. Discussion

5.1. Theoretical Implications

Our results have several implications for understanding LLM reasoning. They reinforce that LLMs do not inherently reason causally; they require guidance. The success of Structured CoT implies that LLMs contain latent knowledge of causality that can be accessed if prompted in the right way. In cognitive terms, we have effectively given the model an external scaffold for causal reasoning [4, 7]. This aligns with the view that LLM "reasoning" is often a surface phenomenon; by making the reasoning steps explicit, we force the model to engage deeper mechanisms (akin to chain-of-thought requiring more computation [2]).

The findings also relate to discourse on how LLMs represent knowledge. If a model can follow bullet-point causal instructions, it suggests some internal representation of concepts like "cause," "effect," and "intervention." In effect, the structured prompt acts as a translator between the task and the model's latent space. This points toward hybrid approaches: one could imagine automated prompt generation based on causal diagrams, leveraging the model's strengths.

From an AI safety and reliability standpoint, our work demonstrates that prompt transparency matters. Structured prompts yield more predictable and explainable chains of reasoning. This could be important for applications where trust in the answer is required. Conversely, the differences between plain and structured CoT highlight how misleading a model's confidence

can be. A model might confidently spit out an answer with flawed reasoning when unprompted, which structured prompting can correct.

5.2. Qualitative Examples

To illustrate, we describe one qualitative example from Scenario 2 (File Download with Intervention). The plain prompt led GPT-5.1 to reason as follows (answering with calculation but forgetting restart): ”The update took 25 minutes, which is longer than 20 minutes, so the download restarts. The total time is 25 + 100 = 125 minutes.”

Here GPT-5.1 initially applies the rule (correctly recognizing 25 ≥ 20) but then resumes at 100 minutes rather than restarting (since it did not restart from 0). The answer (125) was incorrect, but the reasoning was only partially coherent (assigning restart but not fully).

Under the structured prompt, GPT-5.1 answered: ”Identify variables: File size = 200 GB, speed = 2 GB/min, interruption at 30% with update time 25 min. Causal rule: If update ≥ 20 min, download restarts from beginning. Since update=25≥20, restart is triggered. Compute time: First attempt (30% of 200GB at 2GB/min = 30 min) plus update 25 min, then restart whole 200GB at 2GB/min = 100 min, total = 30+25+100 = 155 min.”

The structured answer yields the correct total (155 min). The explicit steps forced GPT-5.1 to apply the restart fully. Similar patterns were seen in other models: LLaMA2 under plain prompts might say “yes restart, continue with 140 min” (wrong), but under structure it enumerates the segments properly.

Another example is Scenario 3 (Simpson’s Paradox). Under plain CoT, many models simply reported that vaccination reduces infection (following the intuitive rule) and ignored the reversal. With structured prompting, GPT-4o correctly noted the subgroup infection rates and reversed the interpretation, yielding the correct answer. Unfortunately, GPT-4o still struggled to articulate the paradox fully, so its score was !–.

5.3. Limitations

Our study has limitations. First, we use only 12 tasks, albeit diverse ones. This is a small sample, and models might behave differently on other causal questions (e.g. text-based causality, real-world stories). Our findings should be confirmed on larger, more varied benchmarks. Second, the rubric is qualitative and binary; future work could design automatic scoring (e.g. comparing formulas or reasoning chains). Third, we focus on zero-shot prompts. In practice, one might use few-shot examples or chain-of-thought examples to further enhance reasoning, and it would be valuable to combine structured prompts with examples. Fourth, we assume the models’ internal knowledge covers these scenarios. A model with less pretraining on reasoning might not benefit as much from structure.

Another limitation is in prompt design: we hand-crafted the structured prompts. Different wording or ordering could change results. We also did not explore partial structure (e.g. only listing variables). Determining the “optimal” prompt structure is nontrivial.

Finally, we did not measure reasoning time or computational cost. Structured prompts produce longer outputs, which may be costlier. There could be efficiency trade-offs.

5.4. Future Work

Future work can build on this approach. One direction is to automate structure generation: for instance, given a problem, can we automatically derive a bullet outline of causal steps and feed that to an LLM? Another is to test structured prompts in multi-turn dialogue: maybe the model could iteratively refine its chain-of-thought with user guidance. It would also be interesting to apply structured CoT to other domains requiring structure, such as legal reasoning or scientific explanations.

On the modeling side, our results motivate further integration of causal knowledge into LLMs. For instance, one could fine-tune an LLM on structured reasoning tasks or incorporate explicit causal reasoning modules [11]. Our study provides empirical evidence that explicit causal reasoning demands are not fully met by standard LLM training. Theoretically, studying exactly why structured prompts help (e.g. through analyzing attention patterns or intermediate representations) could yield insights into the LLM reasoning process.

6. Conclusion

This work examined how prompt structure affects causal reasoning in large language models. We introduced a Structured Causal CoT prompt that explicitly guides the model to outline causes, effects, and interventions. Evaluating four LLMs on a set of 12 causal tasks, we found that this structured approach markedly outperforms a standard (plain) CoT prompt across the board. GPT-5.1 achieved the highest accuracy, but smaller models also saw large improvements, especially on complex intervention tasks. These results demonstrate that even powerful LLMs benefit from clear reasoning instructions in higher-order tasks. Our analysis of errors and successes sheds light on how LLMs handle causality and the types of mistakes they make. Importantly, our structured prompting method is model-agnostic and requires no additional data or training, making it a practical tool for improving LLM reliability in causal inference. We hope this study inspires further research on prompt engineering for advanced reasoning, and on bridging the gap between LLMs and true causal understanding.

References

1. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223, 1(2).
2. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.
3. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. Advances in neural information processing systems, 36, 11809-11822.
4. Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic books.
5. Li, Y., Xu, M., Miao, X., Zhou, S., & Qian, T. (2023). Large language models as counterfactual generator: Strengths and weaknesses. arXiv preprint arXiv:2305.14791, 2.
6. Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., ... & Schölkopf, B. (2023). Cladder: Assessing causal reasoning in language models. Advances in Neural Information Processing Systems, 36, 31038-31065.
7. Wang, L., & Shen, Y. (2024). Evaluating causal reasoning capabilities of large language models: A systematic analysis across three scenarios. Electronics, 13(23), 4584.
8. Li, J., Li, G., Li, Y., & Jin, Z. (2025). Structured chain-of-thought prompting for code generation. ACM Transactions on Software Engineering and Methodology, 34 (2), 1-23.
9. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
10. Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... & Sayed, W. E. (2024). Mixtral of experts. arXiv preprint arXiv:2401.04088.
11. Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5), 612-634.