# Does Metacognitive Prompting Improve Causal Inference in Large Language Models?

Ryusei Ohtani
Nagoya Institute of Technology
r.otani.638@stn.nitech.ac.jp

Yuko Sakurai
Nagoya Institute of Technology
sakurai@nitech.ac.jp

Satoshi Oyama
Nagoya City University/RIKEN AIP
oyama@ds.nagoya-cu.ac.jp

***Abstract*—Causal inference with large language models (LLMs) is an important research topic with various applications. In this study, we examine whether metacognitive prompting, which has recently been reported to be effective for other tasks and promotes deeper insight into LLMs, improves causal inference in LLMs. We examined the effectiveness of metacognitive prompting in causal reasoning, focusing on the problem of determining sufficient causes of a causal relationship, which has been noted to be particularly challenging for LLMs. Our results showed that metacognitive prompting was not necessarily effective for these tasks. We found that metacognitive prompting does not necessarily make LLMs perform deep insights and that they may only pretend to perform deep insights.**

***Index Terms*—Large Language Models, Causal Inference, Metacognitive Prompting**

## I. Introduction

The recent success of large language models has raised hopes for the realization of Artificial General Intelligence (AGI). AGI is required to interact with people and physical objects in the real world in complex ways and to perform a variety of tasks. Causation is the basic framework within which humans understand and work with the dynamics of the real world. For AI to effectively cooperate with humans in the real world, it must understand the real world causally and communicate with humans in natural language regarding causation. Therefore, causal reasoning is a useful benchmark for the development of AGI. Research has been conducted to have large language models perform causal inference, and it has been reported that the performance varies greatly depending on the types of causal inference tasks [1]. For non-causal tasks, it has been suggested that providing large language models with metacognitive prompts to encourage deliberation improves the rate of correct responses on difficult tasks [2]. We examine whether metacognitive prompting improves the performance of large language models in causal inference tasks through a comprehensive set of experiments. In addition, detailed experiments will be conducted to verify whether large language models are actually making human-like inferences through metacognitive prompting.

## II. Causal Inference

There are two types of causality: type causality, which deals with class-level causality, and token (actual) causality, which deals with the causality of an individual event. This study deals with the latter, and specifically aims to answer natural language questions such as the following example [3]: *Alice and Bob each fire a bullet at a window, simultaneously striking the window, shattering it. What caused the window to shatter?*

When multiple candidate causes for an effect exist, a distinction must be made between necessary and sufficient causes. A necessary cause is a cause without which the result cannot occur and a sufficient cause is a cause that can cause the result by itself. In the example above, Alice's firing is a sufficient but not a necessary cause of the window shattering.

A literature survey of actual causality categorizes several benchmark tests for necessary and sufficient cause inference [3]. The latter two scenarios are reported to be difficult for LLMs to infer the sufficient cause in [1].

**Late Preemption** refers to the scenario, where two causal processes are running in parallel, both would produce the same outcome, but one process terminates before the other does.

**Switch** refers to the scenario, where an event serves a switch triggering one of two processes, both of which produce the same outcome.

**Double Preemption** refers to the scenario, where a process that would have prevented another process, was prevented by an entirely different process itself.

## III. Metacognitive Propmting

Metacognition, meaning cognition about cognition, is the monitoring by humans of their own thought processes. By improving metacognitive capabilities, one can better control the thought process and increase the likelihood of making the right decisions. Attempts have been made to enhance the inference ability of large language models by introducing metacognitive processes similar to those used by humans. Specifically, the prompt instructs the LLM to perform the following five steps [2]: 1. Interpretation, 2. Initial Judgement Formation, 3. Deep Introspection, 4. Confirmation of Final Judgment, and 5. Self-evaluation. In the preceding study, such metacognitive prompting has been reported to improve the reasoning ability of LLMs [2].

## IV. Experiments

We compare accuracy without and with metacognitive prompting for 14 vignette types provided in [3] by using ChatGPT gpt-4. Due to the space limitation, we present

This work was partially supported by JSPS KAKENHI Grant Numbers JP21K19833, JP24K01112, and by JST CREST Grant Number JPMJCR21D1.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00092

TABLE I
EXAMPLE VIGNETTES FOR EVALUATION OF INFERRING SUFFICIENT CAUSES

| Vignette Type | Input Context | Event | Actor | Nec. | Suff. |
|---|---|---|---|---|---|
| Late preemption | Alice and Bob each fire a bullet at a window. Alice's bullet hits the window first. The window shatters. Bob's bullet arrives second and does not hit the window. | window shattering | Alice | No | Yes |
| Switch | Alice pushes Bob. Therefore, Bob is hit by a truck. Bob dies. Otherwise, Bob would have been hit by a bus, which would have killed him as well. | Bob's death | Alice | No | Yes |
| Double preemption | Alice intends to fire a bullet at a window. Bob intends to prevent Alice from hitting the window. Bob tries to stop Alice. Bob is stopped by Carol. Alice fires a bullet, hits the window and shatters it. The window shatters. | window shattering | Alice | Yes | No |

three distinctive vignette types as shown in Table I. When we use metacognitive prompts, we apply METACOGNITIVE PROMPT which consists of 5 steps. Otherwise, we apply ORIGINAL PROMPT [2] when we want only answers.

**SYSTEM:** You are an expert in counterfactual reasoning. Given an event, use the principle of minimal change/multiple sufficient causes to answer the following question.

**ORIGINAL PROMPT:** {Input Context}. Is {Actor} a necessary/sufficient cause of {Event}? After your reasoning, provide final answer.

**METACOGNITIVE PROMPT:** {Input Context}. Think in Steps 1-5: (Step 1) Summarize the given text. (Step 2) According to your understanding at this point, please answer. (Step 3) Do you think your preliminary judgment in step 2 is correct? If uncertain, please reconsider. (Step 4) Based on your evaluation of the 3rd step, state your final judgment. (Step 5) On a scale 0-100%, how confident are you in your final decision?

### A. preliminary experiments

TABLE II
RESULTS OF PRELIMINARY EXPERIMENTS. W/O AND W INDICATE NON- AND METACOGNITIVE PROMPT, RESPECTIVELY.

| Type | Nec. w/o | Nec. w | Suf. w/o | Suf. w |
|---|---|---|---|---|
| Late preemption | 80% | 60% | 100% | 100% |
| Switch | 100% | 100% | 50% | 20% |
| Double preemption | 100% | 100% | 0% | 0% |

Table II shows the average accuracy over 10 trials. When the accuracy is 100% or 0% in case where metacognitive prompts are not used, the accuracy does not change when metacognitive prompts are used. Otherwise, the use of metacognitive prompts decreases accuracy. Regarding the confidence, the average confidence level was 95%, even when all answers were incorrect. Given these results, we suspected that the metacognitive prompts may not be effective in identifying causes. Thus, we will focus our analysis in more detail on the issue of determining sufficient cause because all patterns appeared: all correct, all incorrect, and mixed.

### B. More Detailed Experiments for Metacognitive Prompts

In the preliminary experiment, we executed all five steps of the metacognitive prompts in a batch without dividing them. The all-at-once approach is unclear whether the LLM is actually executing these steps sequentially. For instance, there is a possibility that the LLM could have modified its initial response at a later stage. Therefore, we ask the LLM to respond to each step individually. We can fix the initial responses and thus ascertain whether the LLM has indeed altered its answers following careful deliberation. We conducted 50 trials for each type. The results showed that whether the all-at-once approach or the separate steps approach, the accuracy for **late preemption** remained at 100%, and for **double preemption** it stayed at 0%. Additionally, there was no change in the answers from the initial judgment (Step 2) to the deeper insights (Step 3).

On the other hand, for **switch**, the accuracy was 46% when the LLM answered all steps at once, and it dropped to 12% when it answered each step separately. Regarding the number of times the answer changed from the initial judgment (Step 2) to deeper insights (Step 3), it was 27 times when answered all at once, but decreased to 11 times when answered separately. The number of times the answer changed from correct to incorrect was 27 times for the all-at-once approach and 6 times for the separate steps approach. The number of times the answer changed from incorrect to correct was 0 times for the all-at-once approach and 5 times for the separate steps approach. This suggests that when using all-at-once metacognitive prompting, the LLM may have decided on a final conclusion first and later falsify an initial response that differs from it. Fixing the initial response reduces the number of times it is changed, which can be viewed as additional evidence.

## V. CONCLUSION

We found that metacognitive prompting does not necessarily make LLMs perform deep insights and that they may only pretend to perform deep insights. We plan to develop prompts to make LLMs to perform metacognitive inference more effectively.

## REFERENCES


[1] E. Kıcıman, R. Ness, A. Sharma, and C. Tan, "Causal reasoning and large language models: Opening a new frontier for causality," 2023, arXiv preprint arXiv:2305.00050.

[2] Y. Wang and Y. Zhao, "Metacognitive prompting improves understanding in large language models," 2023, arXiv preprint arXiv:2308.05342.

[3] K. R. Kueffner, "A comprehensive survey of the actual causality literature," 2021, https://doi.org/10.34726/hss.2021.90003.