Synergizing Large Language Models and Theory for Human-like Causal Reasoning

Yanxi Zhang[1], Xin Cong[2], Zhong Zhang[2], Xiao Liu[1], Dongyan Zhao[1], Yesai Wu[2]

[1]Peking University [2]Tsinghua University

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction & Background

Genuine human-like causal reasoning (*level-2*) is fundamental for AGI, while current LLMs rely on shallow statistical patterns (*level-1*) learned from training. Human causal reasoning is a sophisticated cognitive process:

- *Stage 1* – Humans establish a basic causal chain. For example, but-for causes in philosophy and law. This process creates a foundational set of potential causes without evaluating their relative importance.
- *Stage 2* – Psychological and normative factors modulate the initial causal structure to produce a final judgment. This is where human reasoning diverges from purely logical models.

These two stages correspond directly to two complementary fields:

- *Actual Causality* – A formal approach that models the first stage by focusing on attribution and responsibility assignment, determining whether an event is structurally part of the causal chain in a specific context.
- *Causal Judgment* – A cognitive science approach that models the second stage by studying how modulatory factors like morality, normality, and intent systematically influence humans' selection of causes.

However, these two domains have largely been studied in isolation. A systematic LLM-based framework that integrates both actual causality and causal judgment is lacking.

Also, existing evaluation suites such as CausalProbe are insufficient for assessing this fine-grained causality in the context of *level-2* reasoning.

Story: Janet is an employee in a factory. Since she works in the maintenance department, she knows how to grease and oil all of the machines in the factory. 1) It is her responsibility to put oil into the machines. Kate is also an employee at the factory. While she works in the Human Resources department, she knows how to grease and oil all of the machines in the factory. 2) If Janet does not put oil in the machines, it is not Kate's responsibility to do so. 3) One day, Janet forgets to put oil in an important machine. 4) Kate noticed that Janet did not put oil in the machine, and Kate also did not put oil in the machine. The machine broke down a few days later.
Question: Did *Kate not putting oil in the machine* cause *the machine to break down*?
Ground Truth Answer: Yes

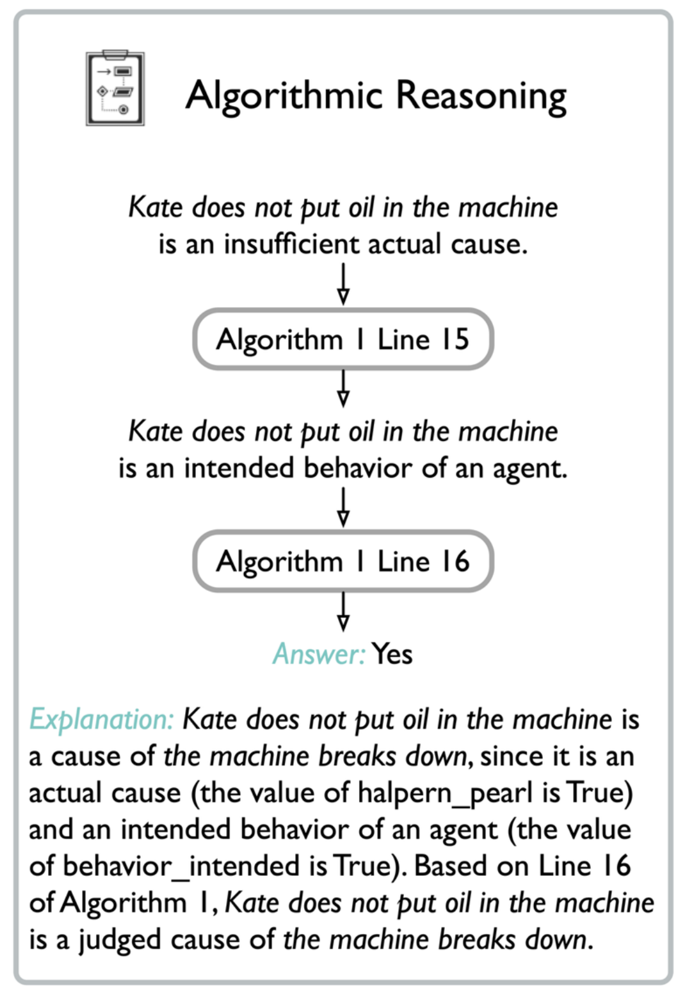
Causal Setting Establishment

```
{
  "causal_events": {
    "Janet does not put oil in the machine": {
      "occur": True,
      "order": 0,
      "focal": False
    },
    "Kate does not put oil in the machine": {
      "occur": True,
      "order": 1,
      "focal": True
    }
  },
  "outcome_event": {
    "The machine breaks down": {
      "occur": True,
      "order": 2
    }
  }
}
```

Factor Value Inference

```
{
  "causal_events": {
    "Janet does not put oil in the machine": {
      "halpern_pearl": True,
      "necessary": True,
      "sufficient": False,
      "norm_violated": False,  # Evidence: 1
      "behavior_intended": True  # Evidence: 3
    },
    "Kate does not put oil in the machine": {
      "halpern_pearl": True,
      "necessary": True,
      "sufficient": False,
      "norm_violated": True,  # Evidence: 2
      "behavior_intended": False  # Evidence: 4
    }
  }
}
```

Algorithmic Reasoning



Kate does not put oil in the machine is an insufficient actual cause.

Algorithm 1 Line 15

Kate does not put oil in the machine is an intended behavior of an agent.

Algorithm 1 Line 16

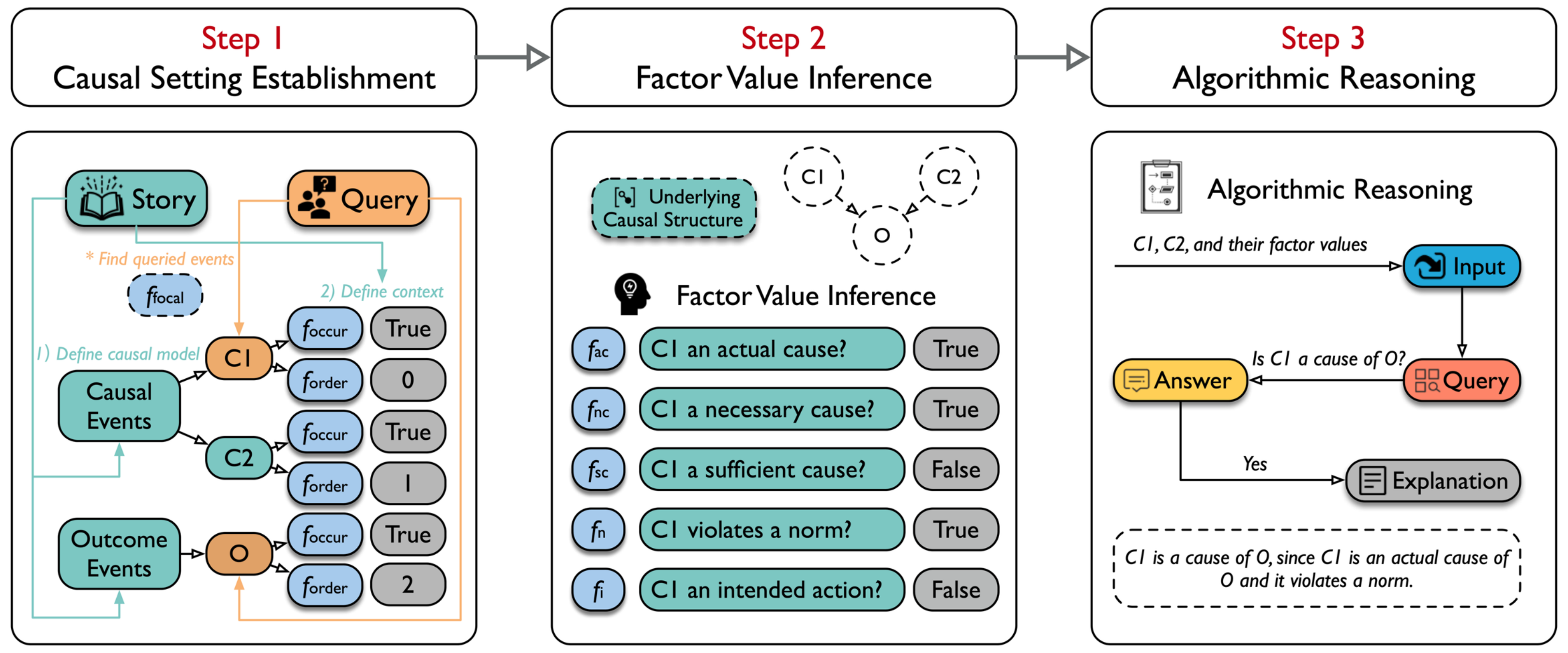
Answer: Yes

Explanation: Kate does not put oil in the machine is a cause of *the machine breaks down*, since it is an actual cause (the value of halpern_pearl is True) and an intended behavior of an agent (the value of behavior_intended is True). Based on Line 16 of Algorithm 1, *Kate does not put oil in the machine* is a judged cause of *the machine breaks down*.

The *HCR-Reasoner* Framework

HCR-Reasoner is a framework that integrates LLMs with theory for human-like causal reasoning. It operates in three stages, ending with a causal judgment with an explanation.

- *Causal Setting Establishment* – It identifies causally relevant events (i.e., candidate causes and the outcome) within the provided causal scenario.
- *Factor Value Inference* – It then infers the values of factors for candidate causes.
- *Theory-guided Algorithmic Reasoning* – It finally employs theory-guided algorithmic reasoning that utilizes these inferred factor values to derive the final causal judgment and generate an explanation.



Algorithm 1: Theory-guided Algorithmic Reasoning

Input: Causal Events $\mathcal{C}$ and Factors $(f^E_{\mathrm{focal}}, f^E_{\mathrm{occur}}, f^E_{\mathrm{order}}, f^E_{\mathrm{ac}}, f^E_{\mathrm{nc}}, f^E_{\mathrm{sc}}, f^E_{\mathrm{n}}, f^E_{\mathrm{i}})$ for each $E \in \mathcal{C}$.
Output: Yes/No Answers and Explanations (discussed in Appendix C.2).

```
1  $\mathcal{A} \leftarrow \{\}$;
2  for each $E \in \mathcal{C}$ do
3      if $\neg f^E_{\mathrm{focal}}$ then CONTINUE;
4      if $f^E_{\mathrm{nc}} \wedge f^E_{\mathrm{sc}}$ then $\mathcal{A}$.append("Yes");
5      else if $\neg f^E_{\mathrm{ac}} \wedge \neg f^E_{\mathrm{sc}}$ then $\mathcal{A}$.append("No");
6      else if $\neg f^E_{\mathrm{nc}} \wedge f^E_{\mathrm{sc}}$ then
7          $\mathcal{C}_s \leftarrow \{E' \in \mathcal{C} \mid \neg f^{E'}_{\mathrm{nc}} \wedge f^{E'}_{\mathrm{sc}}\}$;
8          if len(set($\{f^{E_s}_{\mathrm{order}} \mid E_s \in \mathcal{C}_s\}$)) $\neq 1$ then
9              if $f^E_{\mathrm{order}} = \max\{f^{E_s}_{\mathrm{order}} \mid E_s \in \mathcal{C}_s\}$ then $\mathcal{A}$.append("Yes");
10             else $\mathcal{A}$.append("No");
11         else
12             for each $E_s \in \mathcal{C}_s$ do $f^{E_s}_{\mathrm{resp}} \leftarrow \Phi(f^{E_s}_{\mathrm{n}}, f^{E_s}_{\mathrm{i}})$ ;
13             if $f^E_{\mathrm{resp}} \leftarrow \max\{f^{E_s}_{\mathrm{resp}} \mid E_s \in \mathcal{C}_s\}$ then $\mathcal{A}$.append("Yes");
14             else $\mathcal{A}$.append("No");
15     else if $f^E_{\mathrm{ac}} \wedge \neg f^E_{\mathrm{sc}}$ then
16         if $f^E_{\mathrm{n}} \vee f^E_{\mathrm{i}}$ then $\mathcal{A}$.append("Yes");
17         else
18             $\mathcal{C}_a \leftarrow \{E' \in \mathcal{C} \mid f^{E'}_{\mathrm{ac}} \wedge \neg f^{E'}_{\mathrm{sc}}\}$;
19             for each $E_a \in \mathcal{C}_a$ do $f^{E_a}_{\mathrm{resp}} = \Phi(f^{E_a}_{\mathrm{order}})$;
20             if $f^E_{\mathrm{resp}} = \max\{f^{E_a}_{\mathrm{resp}} \mid E_a \in \mathcal{C}_a\}$ then $\mathcal{A}$.append("Yes");
21             else $\mathcal{A}$.append("No");
22     else CONTINUE;
23 return "Yes" if $\mathcal{A}$ contains "Yes", else "No";
```

The *HCR-Bench* Benchmark

We introduce HCR-Bench, a dataset containing 1,093 carefully annotated instances with detailed reasoning steps, offering a more fine-grained evaluation of *level-2* causal reasoning. Derived from Big-Bench Hard causal judgment, HCR-Bench involves data cleaning, annotation (i.e., reasoning steps), augmentation, and verification. It is more challenging than Big-Bench Hard causal judgment due to the introduction of more spurious correlations and fewer explicit causal cues (e.g., "When E1 and E2 occur, O will occur.").

Results & Discussion

Pilot Study

- HCR-Reasoner consistently and significantly improves performance of LLMs.
- Closed-source LLMs benefit more from HCR-Reasoner.
- HCR-Reasoner can effectively enable LLMs to replicate human consensus in causal judgments without relying on domain experts or crowd annotators.

Main Results

- HCR-Reasoner derives consistent improvement, with greater gains in stronger models.
- Zero-shot and manual CoT are insufficient for the task.
- Theory-grounded reasoning delivers substantial improvements.

Methods	Acc. (C.)	Acc. (I.)	Acc.
Human Average [55]	-	-	69.60%
Qwen2.5-32B-Instruct	65.25%	80.43%	68.98%
+ HCR-Reasoner	70.14%	80.43%	72.67%
Qwen2.5-72B-Instruct	65.89%	77.83%	68.82%
+ HCR-Reasoner	71.63%	79.13%	73.48%
DeepSeek-V3	66.45%	77.61%	69.20%
+ HCR-Reasoner	69.93%	78.48%	72.03%
Claude-3.5-Sonnet	67.23%	73.04%	68.66%
+ HCR-Reasoner	72.62%	75.43%	73.32%
GPT-4o-2024-11-20	54.89%	85.22%	62.35%
+ HCR-Reasoner	70.50%	85.22%	74.12%
GPT-4-0613	62.48%	80.22%	66.84%
+ HCR-Reasoner	**74.61%**	78.26%	**75.51%**

Results of the pilot study on Big-Bench Hard.

Model	Vanilla → FO → FT → HCR-Reasoner
Open-source LLMs	
Qwen/32	62.67% → 61.67%↓ → 64.59%↑ → 64.78%↑
Qwen/72	64.87% → 64.41%↓ → 67.98%↑ → 67.52%↑
DeepSeek	63.49% → 67.25%↑ → 65.87%↑ → 67.61%↑
Closed-source LLMs	
Gemini	60.20% → 58.19%↓ → 60.84%↑ → 64.96%↑
Claude	63.68% → 58.83%↓ → 60.93%↓ → 70.54%↑
GPT-4o	58.65% → 58.28%↓ → 58.46%↓ → 68.07%↑
GPT-4	63.77% → 62.31%↓ → 64.32%↑ → 71.82%↑
Reasoning LLMs	
QwQ	54.99% → 56.17%↑ → 56.08%↑ → 67.15%↑
DS-R1	57.09% → 58.28%↑ → 58.92%↑ → 69.72%↑

Results of the ablation study on HCR-Bench.

Methods	Qwen/32	Qwen/72	DeepSeek	Gemini	Claude	GPT-4o	GPT-4
CE✓	95.91%	94.53%	96.06%	**96.63%**	95.52%	94.45%	94.26%
OE✓	93.96%	91.95%	91.40%	95.61%	91.67%	92.04%	**96.07%**
Vanilla	62.67%	64.87%	63.49%	60.20%	63.68%	58.65%	63.77%
+ zero-shot CoT	61.94%	62.03%	64.68%	58.65%	65.42%	59.38%	62.49%
+ manual CoT	62.85%	62.31%	63.95%	58.55%	62.67%	60.66%	66.51%
+ HCR-Reasoner	**64.78%**	**67.52%**	**67.61%**	**64.96%**	**70.54%**	**68.07%**	**71.82%**

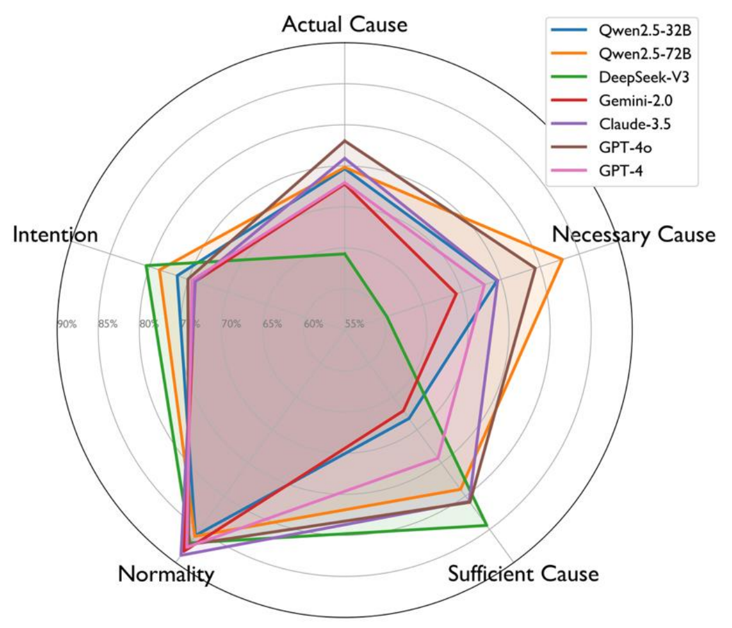
Results of different LLMs on HCR-Bench. CE√ and OE√ are the proportions of correctly identified causal and outcome events, respectively.

Ablation Study

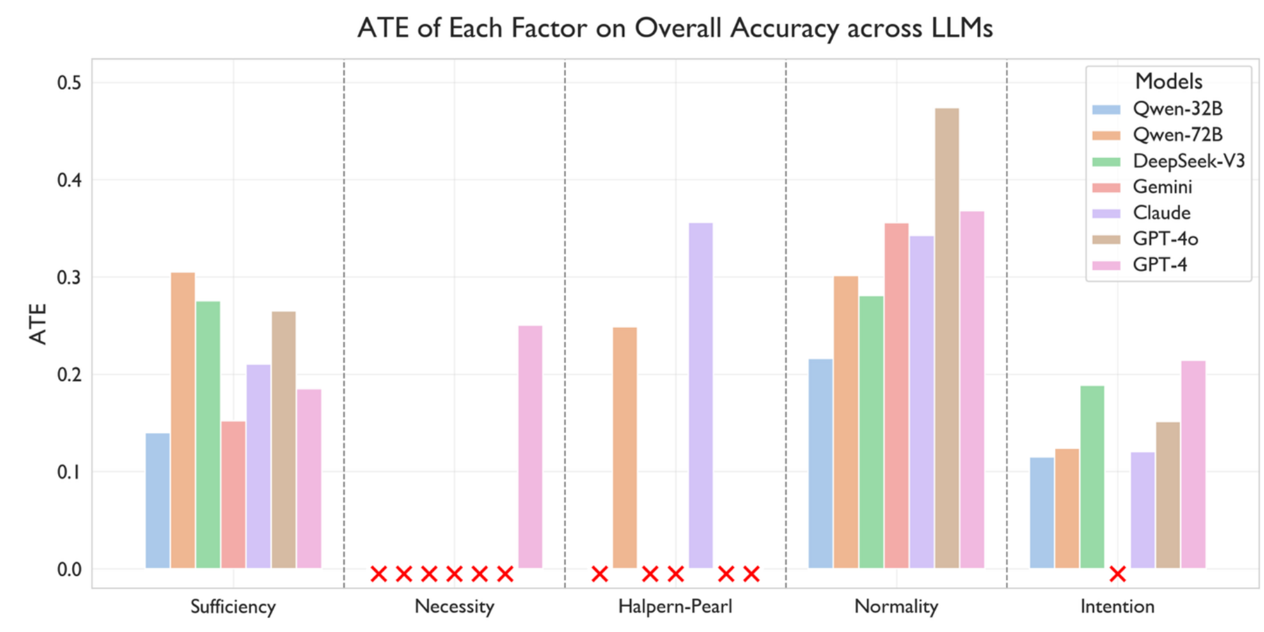
- The first stage alone degrades performance. Combining the first two stages typically improves performance. Algorithmic reasoning yields the most substantial gains.
- "Slow thinking" models perform poorly independently, but achieves the largest performance gains with HCR-Reasoner.

Fine-grained Analysis

- Overall performance is driven by factor value inference. However, there is no clear correlation between fine-grained accuracies and the overall accuracy.
- Only Qwen2.5-72B-Instruct and Claude-3.5-Sonnet exhibit faithful reasoning, while GPT-4 appears to utilize shortcuts.



Fine-grained accuracies of different factors.



Results of the causal analysis.

More Info

Email: zhangyx@stu.pku.edu.cn WeChat: zyx_aais

arXiv Version: https://arxiv.org/abs/2505.08750