Cause and Effect: Can Large Language Models Truly Understand Causality?

Anonymous ACL submission

Abstract

With the rise of Large Language Models (LLMs), it has become crucial to understand their capabilities and limitations in deciphering and explaining the complex web of causal relationships that language entails. Current methods use either explicit or implicit causal reasoning, yet there's a strong need for a unified approach combining both to tackle a wide array of causal relationships more effectively. This research proposes a novel architecture called Context-Aware Reasoning Enhancement with Counterfactual Analysis (CARE-CA) framework to enhance causal reasoning and explainability. The proposed framework incorporates an explicit causal detection module with ConceptNet and counterfactual statements, as well as implicit causal detection through LLMs. Our framework goes one step further with a layer of counterfactual explanations to accentuate LLMs' understanding of causality. The knowledge from ConceptNet enhances the performance of multiple causal reasoning tasks such as causal discovery, causal identification, and counterfactual reasoning. The counterfactual sentences add explicit knowledge of 'not caused by' scenarios. By combining these powerful modules, our model aims to provide a deeper understanding of causal relationships, enabling enhanced interpretability. Evaluation of benchmark datasets shows improved performance across all metrics, such as accuracy, precision, recall, and F1 scores. We also introduce CausalNet, a new dataset accompanied by our code, to facilitate further research in this domain.[1]

1 Introduction

As Large Language Models (LLMs) play an increasingly central role in technology, their ability to understand and logically navigate causal relationships becomes essential since they impact the trust users have on them. [10] This skill is paramount for refining the depth and applicability of LLMs in complex scenarios, driving advancements that hinge on nuanced interpretations of cause and effect.

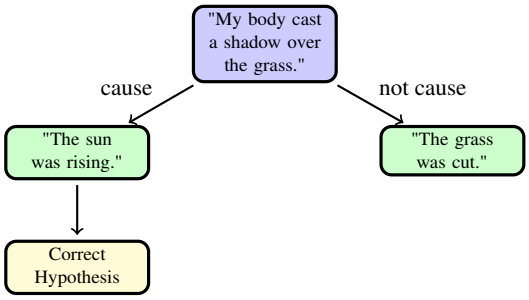

Figure 1: Causal reasoning without CARE-CA: Given the premise "My body cast a shadow over the grass.", the left hypothesis, "The sun was rising," should be identified as the cause to arrive at the correct hypothesis conclusion.

Given the growing reliance on AI systems to make consequential, mission-critical decisions, we need to enhance the causal reasoning capabilities of LLMs. [24, 26] revealed significant limitations in LLMs' causal reasoning capacities. While they may mimic causal language, most need a genuine comprehension of causal mechanisms. This is concerning as it could propagate misinformation or lead to unreliable predictions. Bridging this causal reasoning gap is an active area of research.

Enhancing the causal reasoning abilities of LLMs can significantly impact their reliability and trustworthiness across many applications. A more robust causal understanding of LLMs could improve healthcare and public policy decision-making[15]. It also promises to enhance interpretability and transparency.

However, prevailing approaches need help with flexibility and depth of causal inference. This paper delves into whether these advanced models, like BERT [3], RoBERTa [13], XLM-RoBERTa

[1] https://anonymous.4open.science/r/causal-reasoning-0B6E/

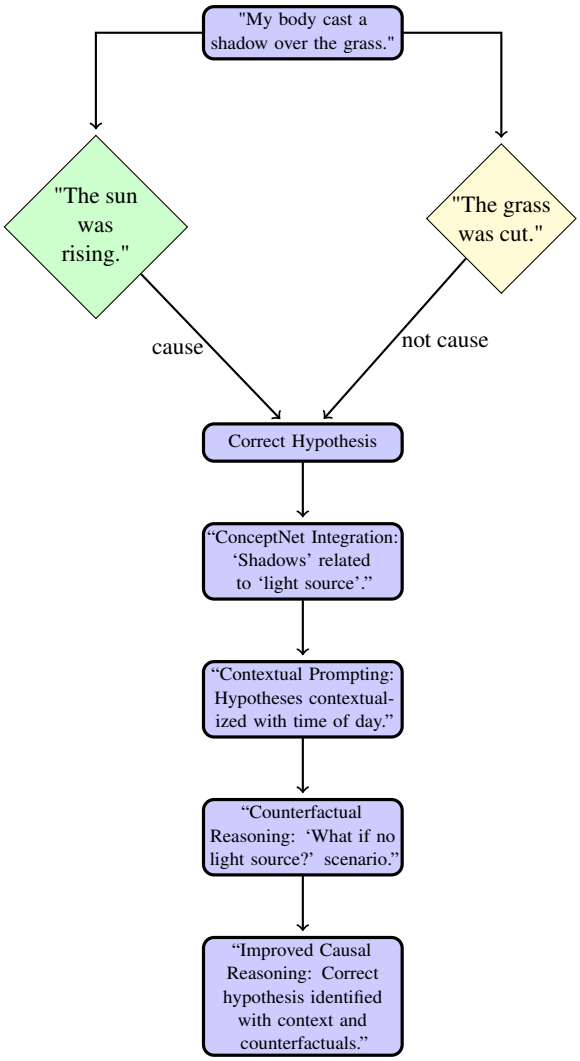

Figure 2: Causal Reasoning Enhanced with CARE-CA: Starting from a premise, causal hypotheses are evaluated. Integration of external knowledge from ConceptNet enhances understanding. Contextual prompting adapts hypotheses to the time of day. Counterfactual reasoning explores alternative scenarios. Improved causal reasoning is achieved by incorporating context and counterfactuals, leading to the identification of the correct hypothesis.

[1], ALBERT [11], DeBERTa [5], Llama 2 [22], T5 [17], Mistral [6], GPT-3.5 [14], and Gemini Pro [21], can truly grasp and articulate causal relationships, a cornerstone in the journey towards Artificial General Intelligence (AGI). We explore this through a blend of theoretical analysis and empirical investigation, focusing on the capability of LLMs to comprehend and articulate causality in the literal sense.

Building on this foundation, we introduce the CARE-CA framework, a novel architecture designed to amplify the causal reasoning competence of LLMs. The CARE-CA framework is distinct in its use of explicit knowledge integration from resources like ConceptNet [20] and implicit reasoning patterns derived from models such as BERT. This dual approach bridges the gap between knowledge-driven and data-driven inference. It enhances the model's performance across four critical domains of causal reasoning: Causal Relationship Identification, Causal Discovery, Causal Explanation, and Counterfactual Reasoning.

We present a comprehensive suite of evaluation metrics, including Accuracy, F1, Precision, Recall, and Human Evaluation, to assess and compare the performance of existing LLMs against our proposed CARE-CA framework. Furthermore, we introduce a new dataset, CasualNet, which, we experimentally demonstrate, boosts LLMs' causal reasoning ability. CasualNet is poised to serve as a benchmark for future advancements in this field, providing a rigorous testing ground for emerging AI models.

By uniting explicit and implicit causal modules alongside contextual and counterfactual enhancements, this research nudges LLMs towards improved causal reasoning—a pivotal step in unraveling AI's black box and realizing more trustworthy, explainable systems.

2 Related Work

Various approaches have been explored in the literature to understand and enhance causal reasoning with LLMs. For example, in this paper [26], the authors assess the ability of LLMs to answer causal questions while discussing their strengths and weaknesses. They discuss the potential of integrating explicit and implicit causal modules to enhance LLMs' capabilities in causal reasoning. However, this lacked a methodical implementation approach to accomplishing the same.

The causal capabilities of LLMs and their implications in various fields such as medicine, science, law, and policy have also been explored by [10]. They dive deep into different types of causal reasoning tasks, presenting how algorithms based on GPT-3.5 and GPT-4 outperform existing algorithms in tasks like pairwise causal discovery, counterfactual reasoning, and actual causality.

Other papers have explored integrating LLMs into research workflows. [2] have proposed an AI assistant using LLMs and causal AI to systematically review manuscripts and provide feedback

to improve causal analysis in epidemiology. [23] demonstrate that domain-specific fine-tuning enhances LLM performance on patient safety and pharmacovigilance tasks demanding accuracy.

[25] critically examines the capabilities of LLMs in causal reasoning and inference while [26] argues that although LLMs can mimic causal language, they lack genuine causal understanding, coining the term "causal parrots".

[12] reveals the limitations in LLMs' causal reasoning by leveraging meta-structural causal models. They find that LLMs trained on code (Code-LLMs) outperform text-only models in abductive and counterfactual reasoning, highlighting the value of programming structure for causal abilities.

Given the widespread implications of LLM causal reasoning capabilities, we aim to enhance the effectiveness of all four aspects of causal reasoning in addition to the LLM evaluation work done in [27]. Our method will specifically focus on enhancing the causal reasoning by incorporating explicit knowledge from knowledge graphs such as ConceptNet.

Notably, the study of causal reasoning has been a prominent focus of research in the field of natural language processing (NLP), especially with the emergence of large language models (LLMs). Prior research has evaluated these models' capacities for causal reasoning, pointing out both their advantages and disadvantages.

One remarkable work is the CRAB benchmark proposed by [18], which evaluates the ability of LLMs to infer causal relationships between real-world events. The authors found that while LLMs can perform well on certain causal reasoning tasks, they struggle with more complex scenarios that require a deeper understanding of causality. Similarly, [8] investigated whether LLMs can infer causation from correlation, a crucial skill for causal reasoning. Their findings suggest that while LLMs can learn some causal patterns, they often fail to distinguish between causal and non-causal relationships, highlighting the need for more targeted approaches.

Additionally, [9] explored the impact of the causal direction of data collection on the performance of LLMs in causal reasoning tasks. They found that models trained on data with a specific causal direction perform better on tasks that align with that direction, underscoring the importance of dataset design in causal reasoning research. These studies provide a solid foundation for understanding the current state of causal reasoning in LLMs.

Prior research has explored various approaches to enhance the causal reasoning capabilities of LLMs. For example, [25] assessed the ability of LLMs to answer causal questions, discussing their strengths and weaknesses. The authors suggested the potential of integrating explicit and implicit causal modules to improve LLM performance, which is a key principle underlying our CARE-CA framework.

While various past works have demonstrated the superior performance of GPT-3.5 and Gemini Pro in certain causal reasoning tasks, their work did not provide a concrete architecture to enhance these capabilities. In contrast, our CARE-CA framework goes a step further by proposing a novel hybrid approach that combines explicit causal knowledge from resources like ConceptNet introduced by [20] with the implicit reasoning capabilities of LLMs.

Interestingly, CARE-CA aims to provide a more comprehensive and effective solution for tackling a wider array of causal reasoning tasks by incorporating counterfactual reasoning and contextual prompting. Unlike previous methods that either relied on explicit or implicit causal reasoning, CARE-CA's unique integration of these two complementary approaches sets it apart, allowing for a more robust and flexible causal understanding. This distinction enables CARE-CA to potentially outperform existing techniques in tasks such as causal relationship identification, counterfactual reasoning, and causal discovery, as demonstrated in our experimental evaluation.

Enhancements to Related Work: The inclusion of "Causal Parrots: Large Language Models May Talk Causality But Are Not Causal" [25], and subsequent studies provide a critical foundation for understanding the current state of LLMs in the realm of causal reasoning. Our framework, CARE-CA, builds on these insights by offering a concrete architecture and implementation designed to overcome the highlighted limitations. Specifically, CARE-CA's novel integration of explicit and implicit causal modules aims to endow LLMs with a more profound, genuine capacity for causal understanding and inference.

Furthermore, our methodological advancements are showcased through the development and utilization of the CausalNet dataset, specifically designed

to benchmark and refine the causal reasoning capabilities of LLMs. By focusing on the four key aspects of causal reasoning—Causal Relationship Identification, Counterfactual Reasoning, Causal Discovery, and Causal Explanation—CARE-CA represents a comprehensive approach to enhancing LLMs' causal reasoning faculties.

3 Approach

CARE-CA Hybrid Causal LLM Framework: Our approach combines the explicit, structured causal reasoning of ConceptNet knowledge graphs coupled with counterfactual sentences to improvise the causal understanding of LLMs. This novel architecture aims to surpass traditional decoder or encoder-only models by leveraging the rich semantic knowledge base of ConceptNet with advanced contextual inference capabilities and 'alternate scenarios' of the contextual sentences to further aid the LLMs in understanding the causality of scenarios. The combination of these two provides relevant contextual information for the LLMs to understand the causal reasoning in question. We carry out a single variable test comparing the performance (X and Y) on CARE-CA v/s without and compare performance with accuracy, recall, precision and F1 scores.

Critical Components of the CARE-CA Framework:

1. **Contextual Knowledge Integrator (CKI):** CKI enriches the AI's reasoning process with relevant external knowledge graph - ConceptNet, providing a deep contextual backdrop against which causal relationships can be examined.

2. **Counterfactual Reasoning Enhancer (CRE):** CRE introduces hypothetical 'what-if' scenarios to test and refine the AI's causal inferences, ensuring that identified causal links are robust and not merely correlational.

3. **Context-Aware Prompting Mechanism (CAPM):** CAPM crafts tailored prompts that encapsulate enriched context and counterfactual insights, directing Large Language Models toward more precise and accurate causal reasoning.

Prompt Example for COPA Dataset: "Shadows are formed when a light source illuminates an object, creating a dark area on the opposite side. Given that 'My body cast a shadow over the grass,' which hypothesis seems more plausible based on the understanding of shadows?

Counterfactual statement: "If the grass was on fire, my shadow would have been the least of my concerns."

'The sun was rising,' providing the light that cast the shadow. 'The grass was cut,' which is a condition unrelated to shadow formation.

4 Experiments

4.1 Data

To develop and evaluate our CARE-CA framework, we employed six distinct datasets. Each dataset serves a specific function within our research, ranging from training the model's causal reasoning capabilities to evaluating its performance in various causal reasoning tasks. All experiments were performed with a dataset split of 75%-25% for train test sets, and 3 runs were conducted for each dataset model combination. We evaluated 5 LLMs - GPT-3.5, Mistral 7b, Gemini Pro, Llama 2, T5 using 5 datasets- COPA, Timetravel, CLadder, Com2sense and e-care on causal reasoning tasks, then compared the same LLMs against our proposed method: CARE-CA'. **Dataset for Causal Reasoning Identification (CRI):**

- **CLadder and Com2Sense:** *Composition:* Derived from narrative texts, these datasets are crafted to pinpoint explicit causal links within a narrative context.
  Purpose: They provide foundational training for the model's explicit causal reasoning abilities, allowing it to recognize and understand causal relationships within complex text structures.

Dataset for Counterfactual Reasoning (CR):

- **TimeTravel:** *Composition:* This dataset presents hypothetical scenarios that challenge the model to reason about events that did not occur.
  Purpose: It is crucial for enhancing the model's counterfactual reasoning, teaching it to contemplate different possibilities and their implications.

Dataset for Causal Discovery:

- **COPA and e-care:** *Composition:* COPA focuses on scenarios that require understanding potential outcomes and alternate realities, while e-care contains medical narratives that add domain-specific intricacies.

Purpose: These datasets are utilized to challenge the model in discovering underlying causal mechanisms within varied and domain-specific contexts.

Each dataset contributes uniquely to the robustness of the CARE-CA framework, ensuring comprehensive coverage across the spectrum of causal reasoning tasks.

Proposed Dataset: We also propose a new dataset called CausalNet. The CausalNet dataset is a valuable resource designed to facilitate causal reasoning and counterfactual analysis research. Comprising 1000 carefully curated scenarios, this dataset presents a diverse set of causal and counterfactual questions, allowing researchers to explore the intricacies of cause-and-effect relationships in various contexts.

Each entry in CausalNet consists of the following components:

Context: A detailed narrative context provides the backdrop for each scenario. These narratives describe situations where multiple events or factors coincide, potentially influencing outcomes. The contexts are designed to be realistic and thought-provoking, setting the stage for causal reasoning and counterfactual exploration.

Causal Questions: For each scenario, a set of causal questions is provided to challenge the models' abilities in causal reasoning. These questions are categorized into two main types:

Cause-Effect Questions: These questions prompt models to identify less obvious factors that may have contributed to observed outcomes. Models must discern the subtle interplay of various events or conditions in determining the outcome.

Counterfactual Questions: Counterfactual questions explore how changes in the scenario's main cause might impact the outcome. Models are evaluated based on their capacity to predict the consequences of hypothetical alterations to the causal factor.

Choices and Answers: Each question is accompanied by a set of choices, one designated as the correct answer. For cause-effect questions, the choices represent potential influencing factors, while for counterfactual questions, the choices depict possible outcomes under different circumstances. The correct answers are carefully labeled to facilitate evaluation.

The CausalNet dataset contributes to advancing natural language understanding and reasoning capabilities. It enables researchers to explore and enhance models' causal reasoning skills, paving the way for more interpretable and context-aware AI systems.

4.2 Experimental Details

Our CARE-CA framework underwent rigorous testing on encoder and decoder models, targeting four distinct causal reasoning tasks: Causal Relationship Identification, Counterfactual Reasoning, and Causal Discovery. The experiments were designed to evaluate the framework's comprehensive capabilities in understanding and processing causal information.

4.2.1 Causal Relationship Identification

Objective: Assess CARE-CA's proficiency in recognizing explicit causal links within narrative contexts.

Dataset Used: CLadder[7] and Com2sense[19], chosen for their rich narrative structures and explicit causal statements.

4.2.2 Counterfactual Reasoning

Objective: Examine CARE-CA's ability to reason with hypothetical scenarios and their implications for understanding potential outcomes.

Dataset Used: timetravel[16], selected for its counterfactual scenarios that challenge models to think beyond the actual events.

4.2.3 Causal Discovery

Objective: Test CARE-CA's capability to unearth hidden or implicit causal relationships within complex scenarios.

Dataset Used: COPA and e-care[4] provide diverse contexts for causal discovery, from abstract reasoning to domain-specific (medical) narratives.

4.2.4 Evaluation Metrics

We used Accuracy, Precision and Recall as evaluation metrics for all the expeirments.

4.3 Results

Evaluating our proposed CARE-CA framework and comparing existing LLMs across different causal reasoning tasks yielded insightful findings. The performance was quantitatively assessed through mean accuracy, precision, recall, and F1 scores, revealing the nuanced capabilities of each model in handling complex causal reasoning scenarios.

Causal Discovery: In causal discovery, our method showcased superior accuracy (76%) on the COPA dataset, emphasizing the framework's strength in integrating contextual and counterfactual insights to uncover underlying causal mechanisms. Interestingly, GPT-3.5 and Gemini Pro also performed well, with accuracies of 73.3% and 70.1%, respectively, indicating their potential in learning causal patterns. The lower performance of models like XLM-RoBERTa and DeBERTa, with accuracies of 53.2% and 51.8%, respectively, could stem from their less effective handling of the dataset's counterfactual and causal scenarios without specific fine-tuning. On the Ecare dataset, our method also performed well with 85.9% accuracy, compared to the next closest decoder model performance of T5 at 84%

Causal Relationship Identification: On the Cladder dataset, the CARE-CA model led with a standout performance, achieving a 63% accuracy, indicating its strong capability to identify causal relationships. The decoder model T5 highlighted its proficiency with a balanced performance, showcasing the effectiveness of its decoding capabilities in causal reasoning tasks.

On the Com2sense dataset, the decoder models encountered diverse challenges, with CARE-CA again leading at 67.1% accuracy, suggesting its consistent ability to navigate causal reasoning tasks.

On our CausalNet dataset, CARE-CA's remarkable accuracy of 94.6% sets a high benchmark, emphasizing the model's superior causal reasoning capabilities. The T5 decoder model mirrored this high performance with a 94.2% accuracy, showcasing the strength of decoder architectures in extracting and interpreting causal relationships from data.

Counterfactual Reasoning: The time-travel dataset, focused on counterfactual reasoning, highlighted models' challenges in understanding hypothetical scenarios. The Gemini Pro and Llama models scored 38.4% and 24.2%, respectively, suggesting that despite their extensive training data, they might struggle with tasks requiring deep counterfactual inference, underscoring the importance of specialized training or prompting for such tasks.T5 and GPT 3.5 models performed well with 61.7% and 63.2% respectively. Our method got a slight jump in accuracy from the best-performing decoders; however, due to information overload, it could not compete with relatively more straightforward encoders such as ALBERT with 68% accuracy.

5 Analysis

The observed performances underscore the complexity of causal reasoning tasks and the varying abilities of models to address them. The CARE-CA framework's superior performance across several tasks suggests that its hybrid approach, which leverages explicit causal knowledge and counterfactual reasoning, significantly enhances causal inference capabilities. LLMs exhibit strong foundational abilities in causal reasoning, likely benefiting from their diverse pre-training. However, tasks requiring nuanced understanding or domain-specific knowledge, such as counterfactual reasoning and causal explanation, highlight the limitations of LLMs and the value of specialized training or frameworks like CARE-CA.

Integrating human evaluation into our study was pivotal in assessing the nuanced capabilities of the CARE-CA framework, particularly in tasks where subjective judgment and a deep understanding of context are crucial. To this end, we conducted a comprehensive study involving 100 examples spanning our four critical causal reasoning tasks: Causal Relationship Identification, Counterfactual Reasoning, and Causal Discovery.

Human Evaluation Methodology:

We also performed human evaluations for the COPA dataset on 100 samples. The evaluator was presented with examples where the CARE-CA framework and other leading LLMs (such as T5, GPT-3.5) responded. The evaluators were tasked with rating the responses based on several criteria:

- *Accuracy:* The correctness of the causal relationships identified or inferred by the models.
- *Coherence:* How logically consistent and understandable the responses were.
- *Depth of Reasoning:* The extent to which the model's response demonstrated an understanding of the underlying causal mechanisms.
- *Relevance:* The applicability of the response to the given causal question or scenario.

Study Findings: The human evaluators consistently rated the CARE-CA framework higher in coherence and depth of reasoning across all tasks, indicating its superior ability to generate responses

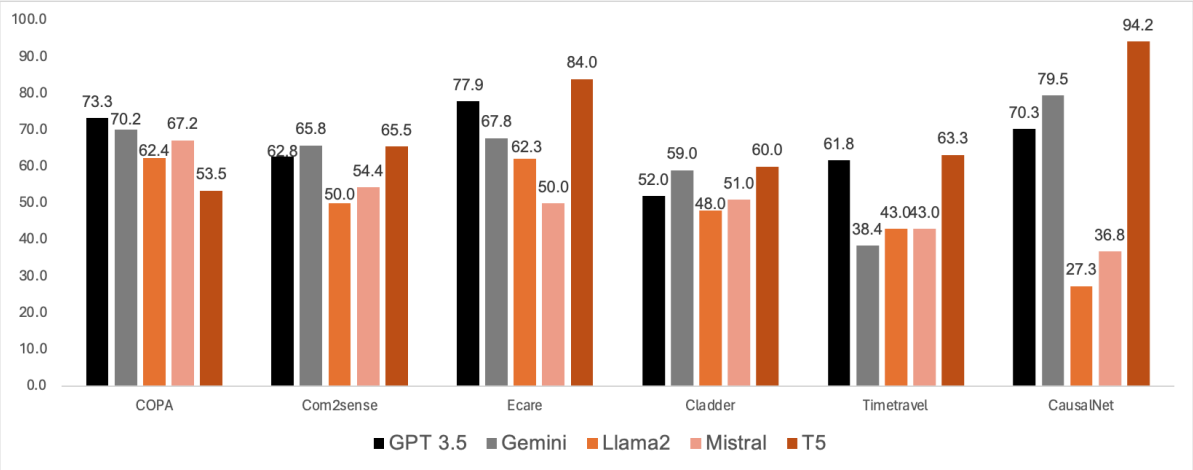

Figure 3: From the experimental results, it is evident that the CARE-CA model consistently outperforms other models across various datasets and tasks in causal reasoning. In causal discovery tasks using the COPA dataset, CARE-CA achieved the highest mean accuracy of 76% compared to other models while in counterfactual reasoning and causal reasoning identification tasks, CARE-CA demonstrated superior performance, achieving mean accuracies of 69.4% and 63%, respectively. Notably, on our CasualNet dataset, CARE-CA achieved exceptional results with a mean accuracy of 94.6%, showcasing its effectiveness in causal reasoning tasks across different contexts.

that identified causal relationships and provided insightful explanations of the 'why' and 'how' behind them. Specifically, in the Counterfactual Reasoning and Causal Explanation tasks, CARE-CA outperformed other models significantly, reflecting its enhanced capability to deal with complex, hypothetical scenarios and to articulate detailed causal narratives.

Feedback from evaluators pointed to occasional challenges in handling highly domain-specific scenarios, especially in the e-care dataset, suggesting an avenue for further refining CARE-CA's domain adaptation capabilities.

6 Conclusion & Future Work

In this project, we have designed and implemented a causal reasoning module. Our system works well under restrictive token constraints.

Future Directions: These results pave the way for further research into hybrid models that combine the breadth of knowledge from resources like ConceptNet with the depth of understanding inherent in LLMs. Fine-tuning strategies, domain-specific model adaptations, and developing more comprehensive benchmarks like CausalNet are promising areas for future exploration.

7 Limitations

In our research on the efficacy of causal reasoning in LLMs through the CARE-CA framework, we encountered several limitations that highlight areas for future exploration and improvement. Firstly, we were able to run CARE-CA only on best performing decoders of each dataset and compare the results. The comparison of CARE-CA on all decoders as well as on all encoders was a challenge due to computational resource constraints. Secondly, our focus on English limits the generalizability of our findings across languages and cultures; this opens a door for a need for multilingual datasets and cross-cultural validation. The challenge of applying our general causal reasoning framework effectively in domain-specific scenarios, such as those presented in the e-care dataset, indicates an opportunity for refining its adaptability to specialized fields. Additionally, the significant computational resources required by the CARE-CA framework may limit accessibility for those with constrained computational budgets, pointing to a need for optimization strategies. While CARE-CA enhances interpretability in causal reasoning tasks, further research is required to improve transparency and explain the model's reasoning processes, especially for non-expert users. These limitations underscore the necessity for ongoing research to enhance the efficacy, inclusiveness, and applicability of causal reasoning models and invite the broader research community to address these challenges collaboratively.

Experiment	Dataset	Model	Mean Accuracy	Mean F1	Mean Precision	Mean Recall
Causal Discovery	COPA	CARE-CA	76.0	82.3	1.0	78.1
		BERT	69.2	66.3	70.0	68.6
		RoBERTa	57.2	56.2	58.3	61.1
		XLM-RoBERTa	53.2	47.0	52.1	56.2
		ALBERT	62.2	63.1	64.0	66.2
		DeBERTa	51.8	0.0	0.0	0.0
		Llama2	62.4	56.0	87.0	68.0
		T5	53.5	1.0	54.0	70.0
		Mistral	67.2	67.2	1.0	87.1
		GPT-3.5	73.3	78	1.0	87.5
		Gemini Pro	70.1	1.0	70.1	82.4
	Ecare	CARE-CA	85.9	88.8	84.6	82.9
		BERT	50	39.4	66	47.6
		RoBERTa	49.7	51.5	50.8	73.1
		XLM-RoBERTa	48.2	58.7	46.7	84.2
		ALBERT	47.7	41.4	50.9	57.7
		DeBERTa	46.6	63.6	46.6	100.0
		Llama2	62.2	60.0	63.8	56.7
		T5	84	84.8	80.5	89.6
		Mistral	50	49.9	50	49.9
		GPT-3.5	77.8	75.9	83.3	69.7
		Gemini Pro	67.8	63.0	74.4	54.5
Counterfactual Reasoning	Timetravel	CARE-CA	69.4	40.1	20.2	13.5
		BERT	56.3	6.0	11.0	5.0
		RoBERTa	68.7	3.0	9.0	2.0
		XLM-RoBERTa	56.9	5.0	10.0	3.0
		ALBERT	68	6.0	11.2	4.0
		DeBERTa	58.1	6.0	11.0	4.0
		Llama2	24.2	1.0	1.0	5.0
		T5	63.2	19.1	12.7	38.2
		Mistral	27.5	2.0	1.0	6.0
		GPT 3.5	61.7	8.0	5.0	14.7
		Gemini Pro	38.4	17.4	10.2	57.3
Causal Reasoning Identification	Cladder	CARE-CA	63.0	62.5	61.9	62.5
		BERT	53.0	48.6	52.3	52.4
		RoBERTa	50.3	65.2	50.3	100.0
		XLM-RoBERTa	49.5	64.3	49.5	99.3
		ALBERT	49.4	46.2	40.5	68.9
		DeBERTa	49.8	22.1	18.0	33.2
		Llama2	48.0	60.0	47.0	82.0
		T5	60.0	59.0	59.0	59.0
		Mistral	51.0	59.0	52.0	70.0
		GPT 3.5	52.0	54.0	53.0	55.0
		Gemini Pro	59.0	65.0	57.0	76.0
	Com2sense	CARE-CA	67.1	28.6	25.7	32.3
		BERT	44.6	59.2	44.9	96.0
		RoBERTa	45.5	1.0	3.0	1.0
		XLM-RoBERTa	50.4	51.4	45.0	60.0
		ALBERT	51.2	35.0	25.0	30.0
		DeBERTa	45.3	60.0	45.6	96.5
		Llama2	50	20.0	10.0	13.3
		T5	65.4	63.4	46.2	53.4
		Mistral	54.3	69.1	71.7	70.4
		GPT 3.5	62.8	23.2	30.4	28.0
		Gemini Pro	65.8	25.2	31.6	28.0
	CasualNet	CARE-CA	94.6	95.4	95	95.4
		BERT	39.0	21.8	15.2	39.0
		RoBERTa	38.0	20.9	14.4	38.0
		XLM-RoBERTa	37.5	20.4	14.9	37.5
		ALBERT	33.8	19.3	27.2	33.8
		DeBERTa	33.5	25.8	22.0	33.5
		Llama2	27.3	23.8	51.3	27.3
		T5	94.2	94.5	95.0	94.2
		Mistral	36.8	29.2	60.9	36.8
		GPT 3.5	70.3	70.9	84.6	70.3
		Gemini Pro	79.5	80.0	83.8	79.5

Table 1: The table summarizes performance metrics Accuracy, Precision, Recall and F1 scores of Encoders - Bert, RoBERTa, ALBERT, DeBERTa, XML-RoBERTa as well as Decoders- GPT 3.5, Gemini Pro, Mistral, T5 and Llama2 on three different tasks including Causal Discovery on datasets COPA and ecare, Counterfactual reasoning on dataset Timetravel and Causal Discovery on dataset CLadder and Com2sense and CausalNet.

8 Ethics Statement

Ethical considerations are paramount in research, particularly when LLMs are involved. We have strived to prevent the propagation of bias within CausalNet, the dataset we introduced in this work, by carefully curating and filtering the data to mit-

igate the inclusion of sensitive or discriminatory content. Furthermore, we have committed to transparency regarding the dataset's origins and potential implications, acknowledging the ethical responsibilities of conducting research with LLMs.

References

[1] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

[2] Louis Anthony Cox. 2024. An ai assistant to help review and improve causal reasoning in epidemiological documents. *Global Epidemiology*, 7:100130.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[4] Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. Submitted on 12 May 2022.

[5] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

[6] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

[7] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. Cladder: Assessing causal reasoning in language models. NeurIPS 2023; updated with CLadder dataset v1.5.

[8] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*.

[9] Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021. Causal direction of data collection matters: Implications of causal and anticausal learning for nlp. *arXiv preprint arXiv:2110.03618*.

[10] Emre Kıçıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

[11] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

[12] Xiao Liu, Da Yin, Chen Zhang, Yansong Feng, and Dongyan Zhao. 2023. The magic of if: Investigating causal reasoning abilities in large language models of code. *arXiv preprint arXiv:2305.19213*.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[14] OpenAI. 2024. https://platform.openai.com/docs.

[15] Alejandro Peña, Aythami Morales, Julian Fierrez, Ignacio Serna, Javier Ortega-Garcia, Iñigo Puente, Jorge Cordova, and Gonzalo Cordova. 2023. Leveraging large language models for topic classification in the domain of public affairs. Accepted in ICDAR 2023 Workshop on Automatic Domain-Adapted and Personalized Document Analysis.

[16] Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053, Hong Kong, China. Association for Computational Linguistics.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

[18] Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. 2023. Crab: Assessing the strength of causal relationships between real-world events. *arXiv preprint arXiv:2311.04284*.

[19] Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. Com2sense: A commonsense reasoning benchmark with complementary sentences. In *Findings of the Association for Computational Linguistics: ACL 2021*. In Proceedings of Findings of the Association for Computational Linguistics: ACL 2021 (ACL-Findings). Contains 16 pages, 14 figures, and 11 tables.

[20] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *AAAI Conference on Artificial Intelligence*, pages 4444–4451.

[21] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

[22] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[23] Xingqiao Wang, Xiaowei Xu, Zhichao Liu, and Weida Tong. 2023. Bidirectional encoder representations from transformers-like large language models in patient safety and pharmacovigilance: A comprehensive assessment of causal inference implications. *Experimental Biology and Medicine*, 248(21):1908–1917. PMID: 38084745.

[24] Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575.

[25] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.

[26] Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. 2023. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*.

[27] Ziyu Zhuang, Qiguang Chen, Longxuan Ma, Mingda Li, Yi Han, Yushan Qian, Haopeng Bai, Zixian Feng, Weinan Zhang, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. *arXiv preprint arXiv:2308.07902*.

Frequently Asked Questions (FAQs)

1. ***What were the reason for choosing these specific set of models- encoders and decoders***

   We selected the encoders - Bert, RoBERTa, ALBERT, DeBERTa, XML-RoBERTa as well as Decoders- GPT 3.5, Gemini Pro, Mistral, T5 and Llama2 based on the most commonly used models in research. We wanted to create a comprehensive study and analysis on these high performing and widely used models as a baseline for future enhancement in the area of Causal reasoning.

2. ***What were the reason for choosing these specific set of models- encoders and decoders***

   We selected the encoders - Bert, RoBERTa, ALBERT, DeBERTa, XML-RoBERTa as well as Decoders- GPT 3.5, Gemini Pro, Mistral, T5 and Llama2 based on the most commonly used models in research. We wanted to create a comprehensive study and analysis on these high performing and widely used models as a baseline for future enhancement in the area of Causal reasoning.

3. ***Why did you choose CARE-CA as your approach?***

   While running experiments with just encoders and decoders, we realized that these models are not very good at causal reasoning tasks and miss on the knowledge needed to help them understand the scenario better. Hence we added knowledge from the Conceptnet knowledge graph. Even after adding the knowledge, we realized this can be further enhanced, if the LLM's can leverage additional what if scenarios using counterfactual statements, that guide them in rejecting hypothesis that are not causal.

4. ***Why did you just run the CARE-CA approach on decoders?***

   While our approach can be applied to encoders as well, we will need fine-tuning due to token limits of encoders. Due to resource constraints, we could not explore and run experiments on encoders with CARE-CA approach, but hoping to produce these results that can be used for further research and enhancement.

5. ***How did you create the dataset CausalNet?***

   We provide a CausalNet dataset, which can be a benchmark for causal reasoning tasks for furture research. The idea was to include causal statements that are currently not supported well by decoders, and have multiple causal reasoning tasks in one dataset. Our dataset has 1000 rows of scenarios with both Causal reasoning identification as well as Counterfactual reasoning. We used ChatGPT to create the dataset.

A Appendix

A.0.1 Detailed run on COPA

A.0.2 Causal Explanation

We also experimented with Causal Explanation task, which lies at the core of understanding why things happen. rather than simply observing patterns or connections, which may not necessarily reveal causality, it delves deeper to pinpoint the direct cause-and-effect links between variables. . Its significance spans across a wide range of disciplines, such as philosophy, science, social sciences, medicine, and engineering, as it enables us to grasp the intricate workings of complex systems and foresee the effects of altering certain variables.

We used the ecare dataset which has the following example scenario -
cause: "The woman gave birth to a child.
effect: "The child brought psycho-physical phenomena on a new life.
conceptualexplanation: "Birth is the arising of the psycho-physical phenomena."
We used Rouge and BLEU score to evaluate the performance of the generated response.

A.0.3 CausalNet Dataset Generation

The CausalNet dataset was generated using GPT-4. The way we created the prompt was using a few shot approach giving a few examples to gpt-4 to understand the causal nature of the sentences and generate prompts that are diverse.

We used the following prompt: Develop a dataset composed of entries that challenge and enhance machine learning models' understanding of causal relationships and counterfactual reasoning across various domains. Each entry in the dataset should follow this structure: "Context": A detailed description of a scenario that outlines a complex situation involving causal relationships. "Questions": A set of questions focusing on (1) identifying causal effects within the context and (2) exploring counterfactual scenarios, with multiple-choice answers to infer the model's reasoning capabilities.

Table 2: Detailed model Performance on the COPA Dataset with three runs capturing Accuracy, Precision, Recall, and F1 Score

Model	**Run**	**Metrics**								
		Manual Accuracy	Sklearn Accuracy	F1	Precision	Recall	Mean Accuracy	Mean F1	Mean Precision	Mean Recall
BERT-base-uncased	1	0.7400	0.7461	0.7067	0.7385	0.7349	0.6893	0.6927	0.6635	0.7003
	2	0.6600	0.6562	0.6573	0.6755	0.6886				
	3	0.6680	0.6758	0.6264	0.6870	0.6348				
RoBERTa-base	1	0.6640	0.6484	0.6582	0.6876	0.6653	0.5787	0.5729	0.5624	0.5833
	2	0.6000	0.5977	0.5735	0.5699	0.6580				
	3	0.4720	0.4727	0.4554	0.4923	0.5116				
XLM-RoBERTa-base	1	0.5640	0.5625	0.5376	0.5413	0.6153	0.5373	0.5326	0.4709	0.5210
	2	0.5000	0.5000	0.4375	0.5413	0.6153				
	3	0.5480	0.5352	0.4375	0.4804	0.4574				
ALBERT-base-v2	1	0.6280	0.6297	0.6382	0.6625	0.6308	0.6240	0.6226	0.6310	0.6406
	2	0.6560	0.6523	0.6473	0.6550	0.6845				
	3	0.5880	0.5859	0.6075	0.6044	0.6725				
DeBERTa-base	1	0.5480	0.5461	0.0000	0.0000	0.0000	0.5200	0.5182	0.0000	0.0000
	2	0.4920	0.4894	0.0000	0.0000	0.0000				
	3	0.5200	0.5191	0.0000	0.0000	0.0000				