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

S. Ashwani, **K. Hegde**, N. R. Munnuru, D. S. Sengar, M. Jindal, K. C. R. Kathala, D. Banga, V. Jain, and A. Chadha

Nov 7, 2024

AI Trustworthiness and Risk Assessment for Challenged Contexts (ATRACC)
AAAI 2024 Fall Symposium
Arlington, VA

Outline

Introduction

Motivation

- Large Language Models (LLMs) are increasingly taking the space once occupied by *search*
- If LLMs are to make the jump to playing a key role in high stakes decision making, understanding causality is crucial. It is also useful for:
  - Refining LLMs' depth and applicability
  - Enhancing trust
  - Improving interpretability
  - Advancing towards Artificial General Intelligence (AGI) (why not?!)
- Current LLMs may mimic causal language without true comprehension

Goals and Contributions

- **Research Goal**: Develop a framework to enhance LLMs' causal reasoning ability.
- **Contributions**:
  - **CARE-CA Framework**: A novel architecture that incorporates explicit and implicit causal reasoning.
  - **CausalNet Dataset**: A new dataset for benchmarking causal reasoning tasks in LLMs.

CARE-CA Framework

CARE-CA: Overview

- It stands for Context-Aware Reasoning Enhancement with Counterfactual Analysis
- Combines explicit and implicit causal reasoning
- Key components:
  - **Contextual Knowledge Integrator (CKI)**: Uses ConceptNet for external knowledge to understand causal relationships.
  - **Counterfactual Reasoning Enhancer (CRE)**: Introduces "what-if" scenarios to confirm causal relationships.
  - **Context-Aware Prompting Mechanism (CAPM)**: Enriches prompts to guide LLMs towards accurate causal reasoning.

Goal: Achieve accurate and comprehensive causal understanding.

CARE-CA: Architecture

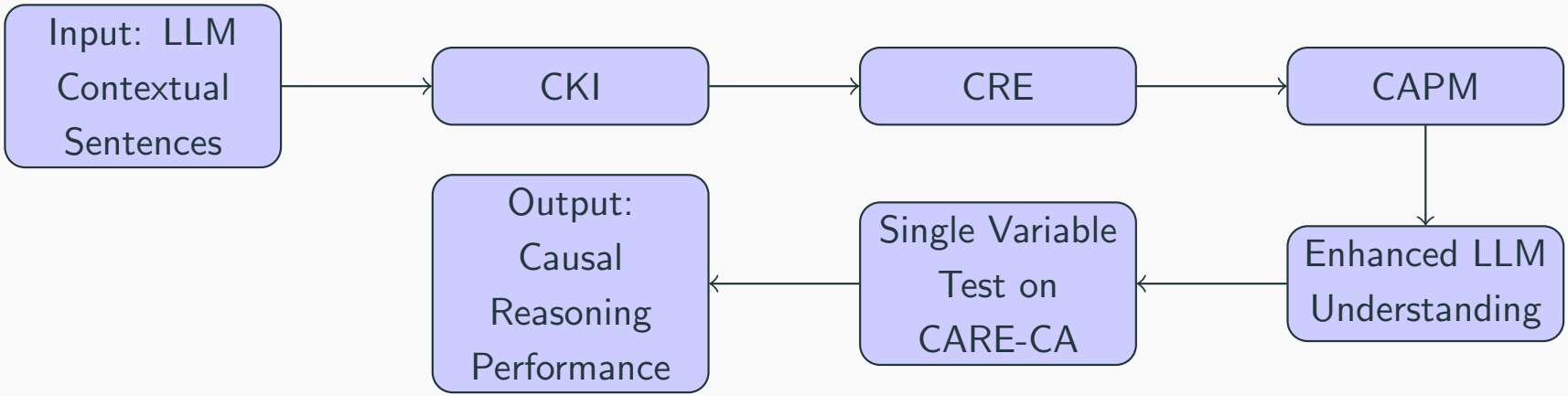

Figure 1: CARE-CA: Architecture

CARE-CA: Example

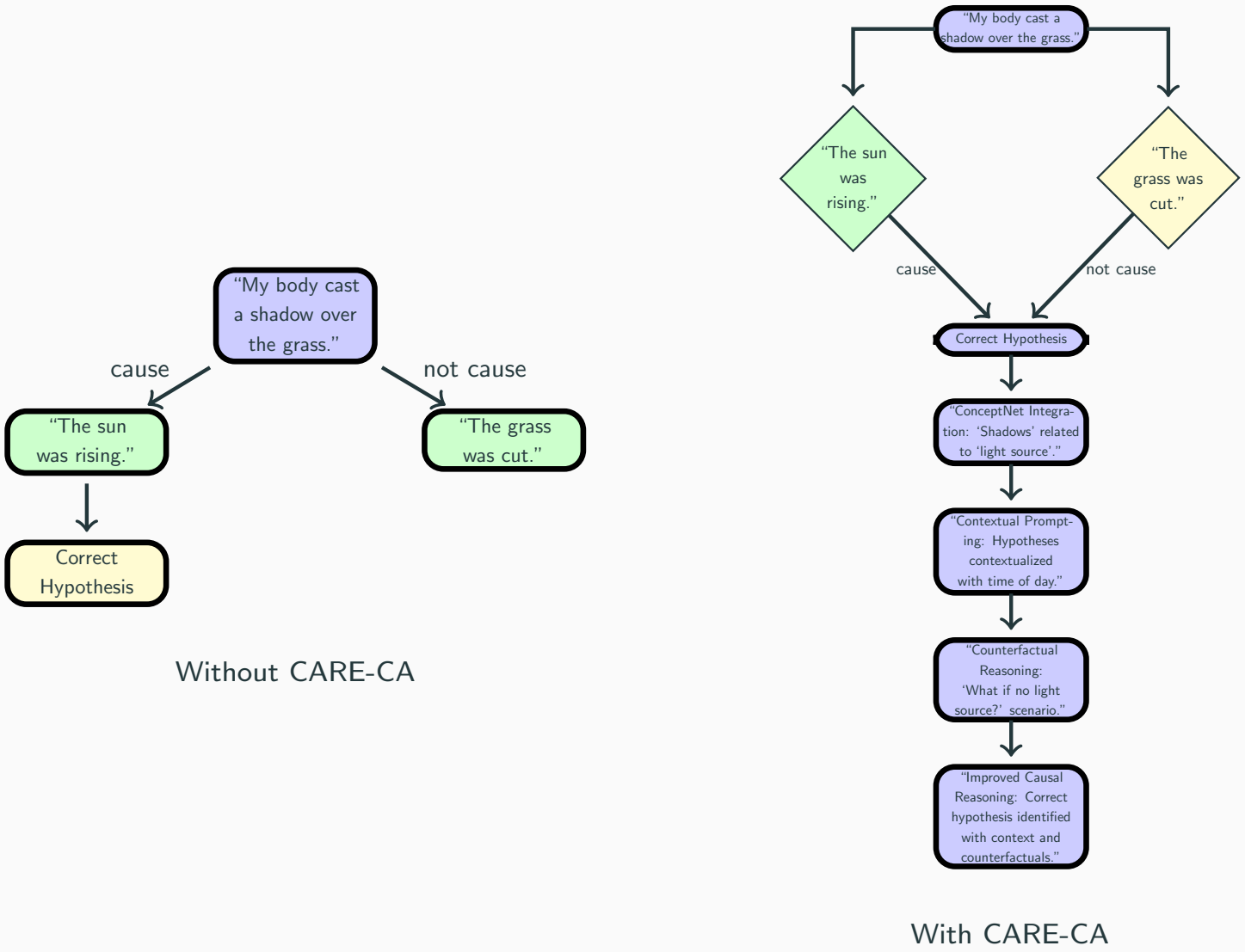

Figure 2: CARE-CA: Before and After

Datasets and Evaluation

Datasets

- **Existing datasets**:
  - COPA: Causal discovery
  - e-care: Domain-specific causal reasoning
  - TimeTravel: Counterfactual reasoning
  - CLadder and Com2Sense: Causal relationship identification
- **Introduced CausalNet dataset**:
  - 1000 scenarios testing causal and counterfactual reasoning.
  - Example entry with detailed narrative and causal questions.
- **Metrics**: Accuracy, Precision, Recall, and F1.

Results

Performance Comparison

- CARE-CA model excels in causal reasoning across datasets and tasks.
- On CausalNet, it achieves 94.6% mean accuracy, demonstrating superior performance in diverse causal contexts.

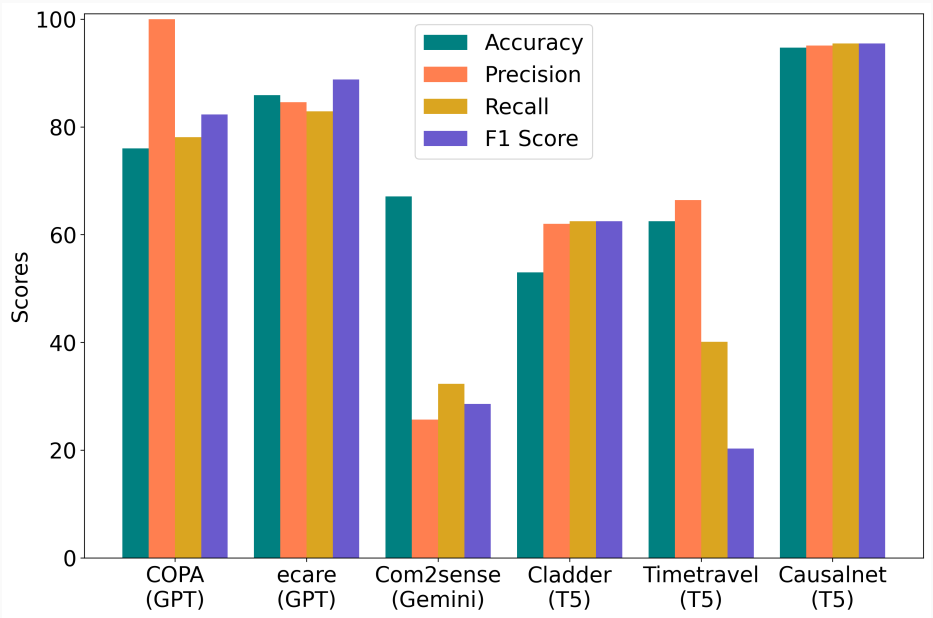

Figure 3: CARE-CA Performance

Performance Comparison

- CausalNet dataset enhances performance across all models.
- T5 shows highest improvement with 94.2% accuracy. Results demonstrate CausalNet's effectiveness in boosting causal reasoning capabilities.

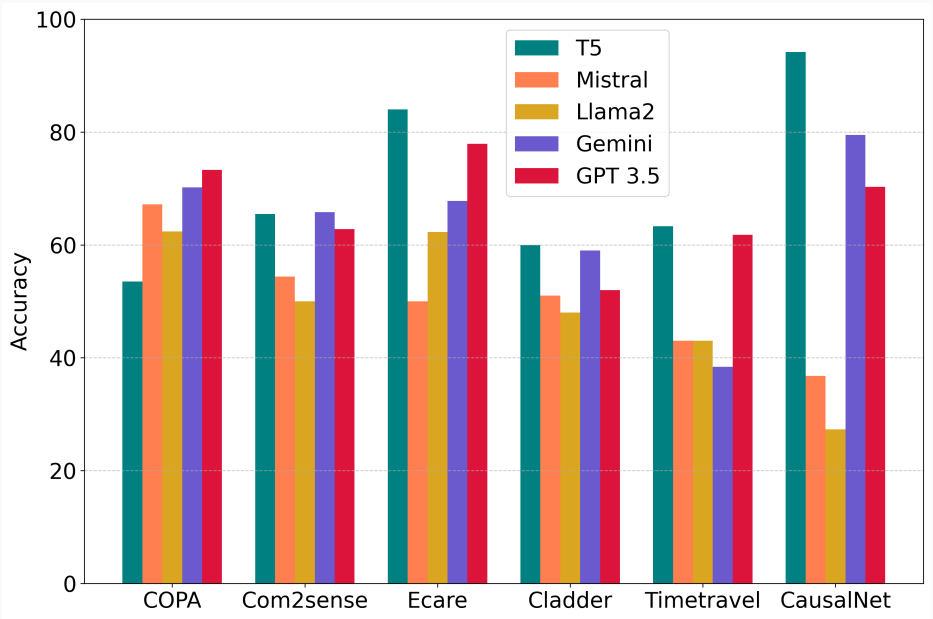

Figure 4: CausalNet Performance

Key Findings

- CARE-CA outperformed traditional LLMs across tasks
- Exceptional performance on CausalNet (94.6% accuracy)
- Improved performance in:
  - Causal discovery
  - Causal relationship identification
  - Counterfactual reasoning
- Demonstrated robustness across diverse causal contexts

Conclusion and Future Work

Conclusion

- CARE-CA significantly enhances causal reasoning in LLMs
- Successfully bridges data-driven and knowledge-driven causal inference
- CausalNet provides a new benchmark for causal reasoning evaluation
- Paves the way for more interpretable and reliable AI systems

Future Directions

- Explore hybrid models combining breadth and depth of knowledge
- Develop fine-tuning strategies for domain-specific adaptations
- Expand multilingual capabilities of CARE-CA
- Optimize framework for diverse domains and complex scenarios
- Further research on transparency and explainability

Thank You

Questions?