6-18-2024 1:00 PM

An Exploration of Causal Cognition in Large Language Models

Vicky Chang,

Supervisor: Daley, Mark J., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Neuroscience

Abstract

Causal cognition, how beings perceive and reason about cause and effect, is crucial not only for survival and adaptation in biological entities but also for the development of causal artificial intelligence. Large language models (LLMs) have recently taken center stage due to their remarkable capabilities, demonstrating human-like reasoning in their generative responses. This thesis explores how LLMs perform on causal reasoning questions and how modifying information in the prompt affect their reasoning. Using 1392 causal inference questions from the CLADDER dataset, LLM responses were assessed for accuracy. With simple prompting, LLMs performed more accurately on intervention queries compared to association or counterfactual queries. Chain-of-Thought (CoT) prompting was also explored with formal reasoning steps included in the prompts. Contrary to expectations, LLMs achieved higher accuracy with simple prompts rather than CoT-enhanced prompts, suggesting that the framework for accurate causal cognition in LLMs differs from that of human cognition.

Keywords

Causality, Cognition, Causal Cognition, Artificial Intelligence, Large Language Models, Causal Inference, Chain-of-thought prompting.

Summary for Lay Audience

Have you ever wondered how the latest, highly popular generative artificial intelligence (AI) models like ChatGPT manage to understand and respond to your questions? At the heart of both human and artificial "thinking" is the ability to understand causes and their effects. This is becoming increasingly important as models such as ChatGPT continue to improve and evolve. This study explores how well these models, known as large language models, grasp questions that require them to think about causes and effects, and whether changing how we pose these questions can influence their response.

We studied these models using 1392 questions designed to test their ability to think about various cause-and-effect scenarios. These questions covered scenarios that required thinking about connections between different events, taking action to change events, and imagining hypothetical events. We found that these models are better at answering questions about direct actions and their immediate results than questions about connections or hypothetical scenarios.

Interestingly, when we tried to help the models by breaking down the questions into step-by-step instructions, a method we thought would improve their answers, it did not help as much as we expected. In fact, simpler, more direct questions without extra guidance led to better answers. This discovery suggests that the way these generative models "think" about causes and effects might be quite different from how humans do, offering insights into both the potential and limitations of artificial intelligence in understanding complex reasoning.

Co-Authorship Statement

This thesis comprises of collaborative work involving myself and my supervisor, Dr. Mark Daley. This project was jointly conceptualized by Dr. Mark Daley and myself to develop the research questions and design the framework for the experiments to address these questions. I was responsible for conducting the experiments, including setting up the experiment, collecting the data, and compiling the results. The results were interpreted together by Dr. Mark Daley and myself. We jointly analyzed the data, discussed the implications of the findings, and refined the final interpretations presented in this thesis.

Acknowledgments

To my family, Margaret, Jeffrey, Ryan, my partner, David, and my corgi, Miya, thank you for being there for me every step of the way. I am so grateful to have you as my support system.

To the friends I made in the Neuroscience program. Alma, Raphael, Kunj, Hannah, Lian, Gabby, Ashlyn, thank you for the coffee chats after seminar, it was endless fun sharing everything going on in our lives, research or otherwise. I am lucky to have met such great friends who continue to motivate and inspire me to be a better person and a better scientist.

To the Neuroscience program committee and staff. Lara Staecker, Susan Simpson, Dr. Brian Corneil, and Dr. Ali Khan, thank you for your hard work in helping me navigate through the milestones to complete this MSc degree.

To the members of the Computational Convergence Lab. Rory Campbell, I have enjoyed each one of our Friday chats and I have learned so much from you. Maitray Patel and Gemma Postill, thank you for inspiring me and supporting me whenever I had questions about the future. Jacob Morra, thank you for your willingness to help me with anything related to neuroscience or artificial intelligence, you are a wonderful mentor. Carson Johnson, I am very glad you joined the lab, thank you for the awesome conversations about philosophy and for your very helpful feedback on my introduction.

To my supervisor, Dr. Mark Daley. Thank you for being an incredible role model throughout the years. Every chat with you leaves me inspired to learn more about such a wide variety of topics. Thank you for dedicating your very valuable time to encourage me and guide me through the challenges of interdisciplinary research. I am extremely grateful for the opportunity to be your student and to be part of your lab.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

AI	Artificial Intelligence
ATE	Average Treatment Effect
CLADDER	Causal Ladder
CNN	Convolutional Neural Network
CoT	Chain-of-Thought
DAG	Directed Acyclic Graph
GPT	Generative Pre-trained Transformer
HANS	Heuristic Analysis for Natural Language Inference Systems
LLM	Large Language Model
LSTM	Long Short-Term Memory
MCTS	Monte Carlo Tree Search
MLP	Multi-Layer Perceptron
NDE	Natural Direct Effect
NIE	Natural Indirect Effect
PTSD	Post-Traumatic Stress Disorder
RNN	Recurrent Neural Network

Chapter 1

1 Introduction: Cognition, Causal Cognition, and Why We Should Care About This Regarding Large Language Models

Cognition encompasses the high-level mental processes that are essential for understanding both natural and artificial intelligence systems, including memory, reasoning, problem-solving, and decision-making (Sternberg, 1999). These cognitive functions are foundational for adapting to and interacting with the environment (Bringsjord & Bringsjord, 2012). As artificial intelligence (AI) integrates more deeply into daily human life, it prompts a parallel and extended investigation into cognitive process in machines, exploring their capability to emulate complex human cognitive functions. This thesis builds upon these discussions, starting with historical perspectives from animal behavior studies and extend through studies that have demonstrated the capacity for AI to perform cognitive functions. It aims to bridge the gap between the disciplines of cognitive neuroscience, computer science, and statistics—specifically, causal inference. By focusing on causality within cognitive processes and how this intersects with AI, this work ventures into a relatively unexplored area, positing that at this intersection of cognition, causality, and AI lies a new frontier for discovery and understanding.

1.1 Cognition in Humans, Animals, and Machines

1.1.1 Foundations for Cognitive Neuroscience

The exploration of cognitive processes traces its origins to animal behavior studies, notably those of Ivan Pavlov and B.F. Skinner, who demonstrated the foundational principles like conditioned responses and associative learning (Pavlov, 1927; Skinner, 1938). This groundwork was extended by evolutionary psychologists who posited that many cognitive mechanisms evolved to address recurrent survival challenges faced by early humans (Tooby & Cosmides, 1992). Recent comparative studies have broadened our understanding, illustrating how cognition has evolved under varying pressures across species, providing insight into the universality and diversity of cognitive strategies (Shettleworth, 2010).

Cognitive neuroscience, merging neurobiology and cognitive science, has provided insights into the neural substrates underpinning thought processes. Landmark studies by researchers such as Elizabeth Spelke and Stanislas Dehaene have identified core cognitive systems involved in spatial reasoning, numerical cognition, and theory of mind, underscoring the foundational presence of these systems across human cultures and even other species (Dehaene et al., 1997; Spelke, 2000). These discoveries increasingly inform AI development, particularly in modeling cognitive processes within computational frameworks. For example, Spelke's insights into spatial reasoning are being utilized to improve AI navigation systems in robots and autonomous vehicles, mimicking human-like spatial navigation abilities (Jefferies & Yeap, 2008). Dehaene's insights into human numerical processing mechanisms have informed the development of AI to emulate these cognitive functions, potentially enhancing their application in numerically driven domains like finance and big data analytics (Dehaene & Cohen, 2007). Lastly, understanding theory of mind – which involves understanding that others have beliefs, desires, intentions, and perspectives different from one's own – informs AI development in socially assistive robotics, which aims to develop robots that assist people with special needs through social interactions (Scassellati et al., 2012). The convergence between cognitive neuroscience and AI development illustrates pathways for cognition studies to inspire future directions in AI and vice versa (Figure 1), deepening our understanding of human and artificial brain and mind (Hassabis et al., 2017).

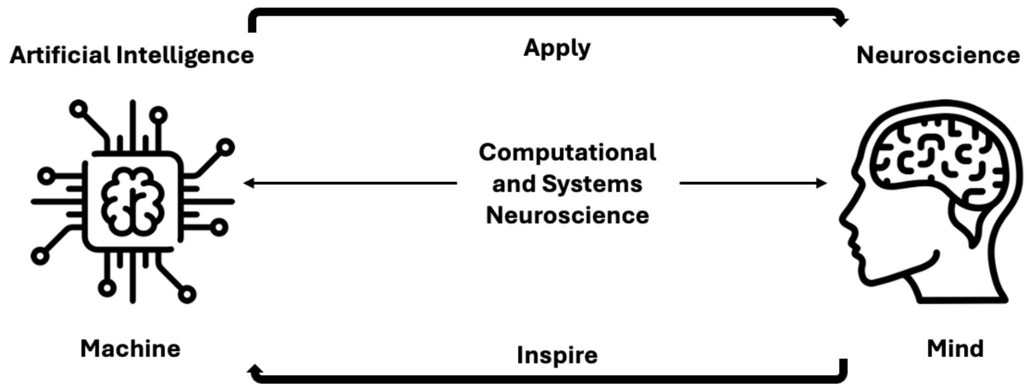


Figure 1. The bi-directional influence of neuroscience and artificial intelligence adapted from Li et al. (2022)

One example of this bi-directional influence is the relationship between the human visual system and neural network architectures. The human visual system has long inspired the development of convolutional neural networks (CNNs), which are now a cornerstone in the field of artificial intelligence. From the field of neuroscience, studies have shown that the visual cortex processes information through a hierarchical structure, with simple features like edges detected in early layers and more complex patterns recognized in subsequent layers. This understanding lead to the design of CNNs, where multiple layers mimic this hierarchical processing, enabling these networks to excel in image recognition tasks (Fukushima, 1980; Lecun et al., 1998). Conversely, advancements in CNNs have provided new insights into how biological vision systems might work. The discovery of how CNNs can perform object recognition tasks has influenced theories about the neural mechanisms underlying visual perception in the brain (Yamins & DiCarlo, 2016). Overall, this symbiotic relationship exemplifies how neuroscience and artificial intelligence continue to drive each other forward, leading to more sophisticated models and deeper understanding in both fields.

1.1.2 Examples of Cognition in Artificial Intelligence, and the Conversation Around AI Consciousness

Artificial Intelligence (AI) is a branch of computer science dedicated to creating machines capable of performing tasks that typically require human intelligence. For example, these tasks might range from navigating to playing strategy games to financial forecasting. Other algorithms can be implemented to accomplish these tasks, but with AI the goal is to automate these processes, usually by implementing algorithms that are capable of learning from data, thereby minimizing human input and enhancing efficiency (Russell & Norvig, 2016). The capabilities of AI systems have grown exponentially, with models now performing, and often out-performing, complex tasks that require cognitive abilities traditionally thought to be unique to humans. Examples of this have arisen throughout recent years. DeepMind’s AlphaGo and AlphaZero not only demonstrated mastery of the games of Go and Chess but have also led to innovative strategies that have profoundly influenced the games’ strategic paradigms. AlphaGo’s victory over world champion Lee Sedol and AlphaZero’s adaptability across various games demonstrate AI’s capacity for complex strategy thinking, challenging the uniqueness of human cognitive flexibility (Silver et al., 2017). In the realm of visual recognition, AI models like those developed from the ImageNet Challenge have achieved and surpassed human-level accuracy in identifying and classifying objects within complex visual scenes (Krizhevsky et al., 2012). Lastly, the development of OpenAI’s Generative Pre-trained Transformer (GPT) 3 demonstrates a significant leap in generative AI through showcasing an ability to produce contextually relevant and coherent text across a wide range of topics (Brown et al., 2020). This advancement pushes the boundaries of AI and machine learning and provides a platform to explore the computational modeling of human language, an area of keen interest in cognitive neuroscience.

With the rapid advancements in artificial intelligence, particularly in large language models (LLMs), we must consider the broader implications of using AI. The human-like response from LLMs that we encounter in the use of everyday applications makes them a candidate for cognition research and also brings to the forefront the debate about AI consciousness. The challenge remains in defining the precise attributes that contribute to consciousness in intelligent systems.

Thomas Nagel’s famous article “What is it like to be a bat?” posits that a being is conscious or has subjective experience if there’s something it’s like to be that being (Nagel, 1974). While we cannot fully grasp a bat’s subjective experience, there is a general consensus that bats possess consciousness and subjective experience. In contrast, an inanimate object like a wooden table is not considered to have any subjective experience—there is nothing it is like to be a table. The debate over what qualifies as conscious also brings attention to the lack of a universally accepted definition of consciousness. Evidence of consciousness can still be gathered, for instance through human verbal reports or interpretations of animal behavior (Pennartz et al., 2019). A recent study by Butlin et al. (2023) offers a review of what consciousness could signify for AI, touching upon the prevailing theories of consciousness, the computations linked to conscious process, and the traits that might suggest consciousness in AI systems. While this thesis does not attempt to resolve the complex issue of AI consciousness, it provides a relevant backdrop for the investigation into a related yet distinct domain: the causal reasoning capabilities of AI systems, specifically of LLMs. Causal reasoning is closely linked to cognitive experiences and a potential step toward consciousness, along with other facets like sensory, affective, and agentive experiences. This work aims to explore causal reasoning as a distinct but related area within the broader context of consciousness. Contributing towards the larger question of "What it's like to be an LLM?" this thesis aims to explore the sub-area of how LLMs reason about causality. An LLMs approach to understanding causal relationships may fundamentally differ from human processes. By examining how AI models like LLMs handle causality, this thesis aims to uncover unique insights into both the capabilities and limitations of AI in mimicking human-like cognitive functions.

1.1.3 Introduction to Large Language Models and the Transformer Architecture

Narrowing in on the intersection between cognitive neuroscience and artificial intelligence, this thesis seeks to explore one element of cognition in LLMs. LLMs are a type of AI system designed to understand, generate, and interact using human language. This type of system assigns probabilities to sequences of text, using these probabilities to generate new text based on given prompts. LLMs stand out due to their architecture and training on broad datasets, enabling them to produce text that closely mimics human writing. Currently, LLMs predominantly use the transformer architecture, noted for its efficiency and effectiveness in handling sequence-based data without requiring the sequential data processing inherent in previous models such as recurrent neural networks (Vaswani et al., 2017).

Since the introduction of the transformer architecture by Vaswani et al. (2017), many variations of transformer-based language models have emerged. This thesis will specifically examine autoregressive, decoder-only models such as GPT-3 by OpenAI. Unlike the original architecture, which featured an encoder-decoder structure designed for language translation, many modern models streamline this configuration by omitting the encoder component. These decoder-only models start with a token embedding, followed by a series of residual blocks, followed by a token unembedding. Each residual block consists of an attention layer, followed by a multi-layer perceptron (MLP) layer (Figure 2).

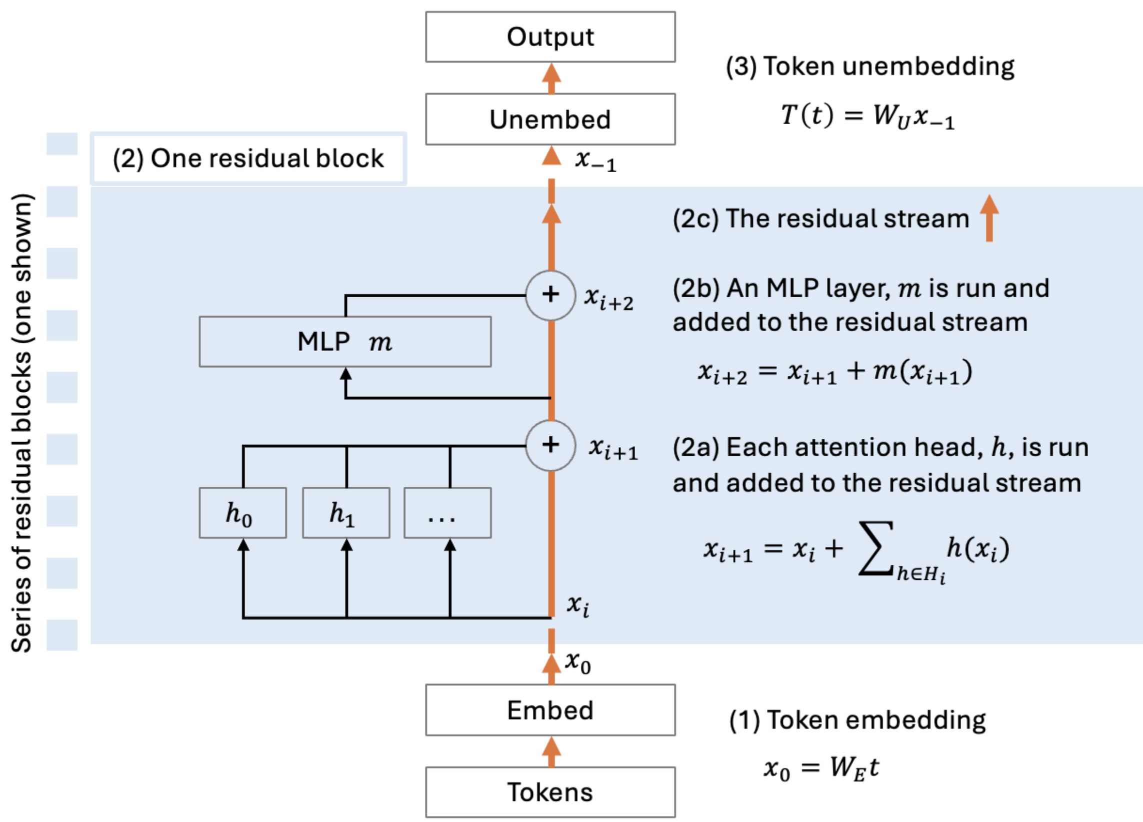


Figure 2. High-level architecture of a decoder-only transformer model, adapted from Elhage et al. (2021).

(1) Token embedding: Each word or subword (token) of the input text is converted into a vector representation (an embedding, x_0). This is a numerical form that encapsulates its semantic and syntactic attributes (Lin et al., 2017). Embeddings serve as the initial input for the subsequent layers of the transformer (Vaswani et al., 2017).

(2) Residual blocks: Each block refines the embeddings by adding more context or adjusting the representation based on the input's (x_i) interactions within the text. Each residual block contains an attention head (h) and a multi-layer perceptron (m) layer, these both write to the residual stream (Elhage et al., 2021).

a. Attention layer (h): This layer determines the importance or 'attention' that should be given to other tokens when processing a specific token. The mechanism works by scoring how much each token in the sequence should influence another, allowing the model to focus more on relevant tokens and less on irrelevant ones. Within the attention layer, there are multiple attention heads ($h_0, h_1, \ldots$), which work in parallel, and are trained (tuning the parameters or weights to best fit an objective function) to each produce their own attention scores independently. By doing so, the model can capture various aspects of context, as each head might focus on different types of relationships between tokens (Vaswani et al., 2017).

b. Multi-layer perceptron, MLP (m): This is a small feedforward neural network applied independently to each position. The MLP layer transforms the attention vectors to vectors that are used in subsequent layers. However, there has not been much success in understanding MLP layers in a transformer architecture, and research teams at companies such as Anthropic are still tackling this problem (Elhage et al., 2021).

c. Residual stream: The residual stream functions as a pathway that carries information across different layers without transformation. Each layer first "reads" its input by performing a linear projection of the residual stream, processes it, and then "writes" its output back by adding another linear projection to the stream. The residual stream is a sum of the output of all the previous layers and the original embedding and maintains an additive linear structure. The residual stream is a high-dimensional vector space, ranging from hundreds to tens of thousands of dimensions depending on the model size. Such dimensionality allows layers to encode and retrieve information from different vector subspaces, and information within these subspaces to be preserved across layers unless explicitly modified or overwritten, thereby serving as a form of "memory" or "bandwidth" within the network (Elhage et al., 2021).

(3) Token unembedding: After passing through the series of residual blocks, the high-dimensional token vectors are converted back to a more interpretable form, usually as scores or probabilities, $T(t)$, over a vocabulary for text generation (Vaswani et al., 2017).

In the training phase of LLMs, the primary task involves adjusting the model's parameters through a systematic process to enhance prediction accuracy and minimize errors (Goodfellow et al., 2016). Training begins by inputting a vast dataset, typically composed of texts from diverse sources such as books, articles, and websites, which is preprocessed into a format suitable for the model, like tokenization (Vaswani et al., 2017). During the forward pass, the model generates predictions based on initial parameter values, which are compared against actual outputs to calculate loss using a predetermined loss function (LeCun et al., 2015). This loss quantifies the discrepancy between predictions and true data, guiding the backpropagation process where gradients for each parameter are computed, indicating how adjustments should be made to reduce the error. Then, parameters—including weights, biases, embedding matrices, and attention mechanism coefficients—are updated with optimization algorithms to minimize the loss function. This cycle repeats across multiple epochs, allowing the model to process the entire dataset repeatedly, refining its parameters to enhance performance (Goodfellow et al., 2016; LeCun et al., 2015). Specifically, weights and biases in neural layers, embedding matrices for token vectorization, and parameters governing attention mechanisms are finely tuned (Vaswani et al., 2017). Through these training steps, LLMs learn to generate outputs that closely mimic human-like text. The size of these models is often indicated by the number of parameters they contain. To summarize, each parameter is a learnable weight in the neural network that contributes to the model's ability to discern and generate appropriate textual responses based on the data it has learned during training.

1.1.4 The Emergence and Impact of Large Language Models for Cognitive Research

The rise in popularity of LLMs marks a significant milestone in ongoing research within the fields of computational linguistics and cognitive science. Despite their rapid integration into mainstream technology—exemplified by tools such as ChatGPT—it is crucial to recognize that these models are the culmination of extensive interdisciplinary research spanning several decades. This section will explore the evolution and cognitive capabilities of LLMs, emphasizing their growing influence and potential in understanding and replicating complex cognitive functions.

The origins of LLMs can be traced back to the early endeavors of AI in the 1950s, where the foundation was laid with rule-based systems aimed at translating text and simulating basic human linguistic abilities. This period set the stage for the intricate relationship between language processing technologies and cognitive theories (Newell & Simon, 1972). In the late 20th century, with the advent of the internet and explosion of digital data, statistical methods gained prominence. Such methods were employed to leverage large text corpora and develop probabilistic models, steering the field from rule-based to data-driven paradigms (Darema, 2004). The 2000s, with the introduction of neural networks and renaissance of deep learning, brought a new dimension to AI research. Technologies like recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) enhanced the ability of machines to handle the sequential and contextual nature of language (Hochreiter & Schmidhuber, 1997). Next, a significant breakthrough came with the introduction of the transformer model which was previously described (Vaswani et al., 2017).

Attention is not only an important aspect of human cognition, but also an important aspect of LLMs. The transformer model uses self-attention mechanisms, diverging from the traditional sequential text processing typical of RNNs by enabling parallel data processing (Vaswani et al., 2017). Compared with previous models, self-attention mechanisms in LLMs align well with certain cognitive processes observed in human thought, such as the simultaneous consideration of multiple contexts within memory and reasoning tasks. Additionally, transformer models can perform tasks that were previously deemed uniquely human. For instance, OpenAI's GPT series can generate code, comprehend multiple languages, and analyze extensive passages of text simultaneously. Despite these similarities, it is not implied that transformer-based models think or process information exactly as humans do. Nor is the possibility that these models engage in a form of 'thinking' dismissed altogether. Rather, the depths and limits of artificial cognition and its parallels to human thought processes continues to be an active area of investigation within cognitive science and AI research. Geoffrey Hinton, a Turing award winner who has been dubbed the "Godfather of AI" said in a recent interview:

> I am very confident that they think. So suppose I'm talking to a chatbot, and I suddenly realize it's telling me all sorts of things I don't want to know. Like it's telling me it's writing out responses about someone called Beyonce, who I'm not interested in because I'm an old white male, and I suddenly realized it thinks I'm a teenage girl. Now when I use the word 'thinks' there, I think that's exactly the same sense of 'thinks' as when I say 'you think something.' (Hinton, 2023)

Reasoning and thinking are foundational cognitive processes that have been the subject of extensive philosophical and scientific inquiry. Reasoning is typically defined as the process of drawing conclusions or inferences from premises or evidence. It involves logical steps to arrive at new information or decisions, often categorized into deductive, inductive, and abductive reasoning. Thinking, on the other hand, is a broader term that encompasses reasoning but also includes other cognitive activities such as imagining, remembering, and problem-solving. Various theories attempt to delineate these concepts. The dual-process theory, for instance, distinguishes between System 1—fast, automatic, and often subconscious processes—and System 2—slow, deliberate, and conscious processes (Kahneman, 2011). This theory suggests that humans have two distinct modes of thinking, which can sometimes lead to different outcomes or biases in reasoning. Another significant approach is the theory of mental models. According to this theory, thinking involves constructing and manipulating mental representations of real or hypothetical situations (Johnson-Laird, 1995). These mental models allow individuals to simulate different scenarios and reason through complex problems by envisioning the outcomes of various actions or events. The computational theory of mind, which we will adopt for this thesis, equates thinking to computational processes in the brain. This theory posits that cognitive functions, including reasoning and thinking, can be understood as information processing activities carried out by neural circuits. Just as a computer processes data through algorithms, the brain processes information through neural networks (Pinker, 1999). This perspective aligns with the development of artificial intelligence, and the bi-directional influence explored in Figure 1, where neural network models may simulate human cognitive processes, providing a bridge between biological and artificial systems. Just as there is a diversity of approaches to understanding cognitive processes, there is a diversity of approaches towards understanding thinking and reasoning. As emphasized in a discussion on AI consciousness, different frameworks can offer unique insights but often fail to converge on a singular definition. By adopting the computational theory of mind, this thesis will explore thinking and reasoning as processes that can be modeled and understood through the lens of information processing (Butlin et al., 2023). The aim of this approach is to allow for the integration of insights from both neuroscience and artificial intelligence, while acknowledging the ongoing debate and lack of consensus in defining these cognitive functions.

From a research perspective, LLMs offer a unique platform to test hypotheses about language processing and cognitive functions as mentioned in the abstract. For instance, recent studies have shown that LLMs have approximate aspects of human syntactic processing and semantic prediction, shedding light on how neural networks can be leveraged to model complex cognitive tasks (Butlin et al., 2023). An important area in AI and cognition research is causal cognition. As AI models continue to evolve, methodologies and theoretical models to better understand and enhance the causal reasoning is gaining importance—and these ideas will be expanded upon in the following section. To explore whether current LLMs can reason about causality, first we must define what it means to reason about causality and understand the foundational frameworks that have been developed to understand the science of causality.

1.2 Causality, Causal Cognition, and Causal Inference

A man steps on a rake. The handle swings up. A bird flutters. A window shatters. A curtain flaps.

If you were to perceive these events, you would probably not recount them in such a simple, disjointed way. You might instead tell a story that enriches the bare facts with connections and contexts, like this:

A man absentmindedly stepped on a rake, causing the handle to swing up abruptly. This sudden motion startled a nearby bird, which took flight and, in its panic, collided with a window. The impact shattered the glass, and the resulting vibration caused a curtain inside to flap wildly.

This richer, connected narrative exemplifies how humans employ causal cognition to interpret sequences of events, implying a chain of cause-and-effect that helps make sense of what might otherwise appear as disjointed happenings (A. Bender & Beller, 2019). This narrative construction goes beyond merely listing events by integrating them with causal links – “caused,” “startled,” and “resulted in.” Language that describes causality is pervasive and we use it without a second thought, but this example highlights the assumptions we naturally make about the physical world—for instance, assuming the laws of physics governing how objects in motion cause others to move or break. We naturally attribute agency and causality, here considering the man’s accident as the trigger for a series of reactions that culminate in the flapping curtain, thus framing a sequence of events as a comprehensible story.

Causal cognition involves the processes engaged in understanding and reasoning about cause-effect relationships, which is crucial across various cognitive domains (A. Bender, 2020; Sloman, 2005). This capacity allows individuals, including humans and other animals (Kummer, 1996), to interpret causal sequences from daily experiences, such as understanding that stepping on a rake might cause the handle to swing up unexpectantly in our story. The ability to understand causal mechanisms facilitates critical functions such as diagnosis, prediction, and intervention, and is so advantageous that it has been considered an evolutionary driving force (Lombard & Gärdenfors, 2017). Moreover, these mechanisms have been fundamental to the development of organized society (Pearl & Mackenzie, 2018). To structure what we know about causal cognition in the literature, we first need to define what is meant by "understanding causality." The conceptualization of causality has historically been muddled by ambiguous terminology and a lack of a standardized framework, which has significantly impeded systematic study and clear communication in the field (Pearl & Mackenzie, 2018).

Causality has been a subject of philosophical inquiry for centuries, with fundamental works by Hume (1779) exploring the nature of cause and effect through a philosophical lens, emphasizing empirical observation and the notion of constant conjunction—where causes are regularly followed by their effects. This line of inquiry was later expanded by Giere & Salmon (1988), who introduced the idea of causal fields as a way to frame conditions under which causes bring about effects, and G. H. von Wright (2004), who distinguished between different types of causation such as necessary, sufficient, and contributory causes in his analytical approach. However, it was not until the work of Judea Pearl and colleagues that a more formalized and robust mathematical framework for understanding causality began to take shape, laying the groundwork for the science of causality.

The science of causality seeks to discern not just the relationships connecting one variable to another but the directional influences that these variables exert upon each other. The question of "Why" is foundational across scientific disciplines, profoundly affecting our understanding of phenomena from natural sciences to human behavior. While causal cognition involves thinking about and understanding cause-effect relationships, causal inference is the statistical approach bridging the observation of patterns, usually in data, to understanding cause and effect relationships. Causal inference provides important formal definitions and frameworks that can be applied to study causal cognition. Typically, causal inference tools allow researchers to clearly delineate and manipulate causal relationships in observational data, thus addressing the limitations of previous methods that were heavily reliant on controlled experiments (Pearl, 2000). This is particularly important in fields such as epidemiology and economics, where experiments may be unethical or unfeasible, thus formal approaches to isolate variable effects while controlling for confounders leads to powerful insights about causality that would otherwise not be possible simply with observational studies. The following section will provide a review the foundational ideas that make this possible, starting with Judea Pearl's conceptualization of the "Ladder of Causation." Then, we will review the gaps in the literature about causal cognition in LLMs, and we can borrow tools from causal inference to explore, using a formal framework, causal reasoning abilities of LLMs in this interdisciplinary study.

1.2.1 The Science of Causality and Pearl's Ladder of Causation

In "The Book of Why," Pearl outlines a conceptual framework for understanding cause-and-effect relationships through what he calls the "Ladder of Causation." This framework illuminates our understanding of causal reasoning across different species by delineating three distinct levels: association, intervention, and counterfactuals (Pearl & Mackenzie, 2018) (Figure 3). These levels represent a spectrum of cognitive abilities observed in both humans and non-human animals.

At the foundational level, association pertains to observing correlations, a capability observed in many species. For instance, research indicates that rats can discern simple associations between stimuli and rewards, a basic form of causal understanding (Blaisdell et al., 2006).

Ascending to the intervention level, we encounter capabilities typically associated with more complex cognitive processes. Intervention advances beyond observation to manipulating one variable and seeing effects on another. For example, primates, including chimpanzees and rhesus monkeys, have demonstrated the ability to manipulate their environments purposefully to assess outcomes, suggesting a grasp of basic causal mechanisms (Kaminski et al., 2008; Kummer & Goodall, 2003).

The top of the ladder, counterfactuals, involves understanding hypotheticals and reflecting on alternative outcomes, abilities that appear predominantly human. While studies such as those by Byrne (2005) illustrate that some primates can engage in tactical deception—a rudimentary form of counterfactual thinking—only humans have the capacity to fully contemplate complex hypothetical situations and their potential alternative histories, a critical skill in advanced problem-solving and moral reasoning (Weisberg & Gopnik, 2013).

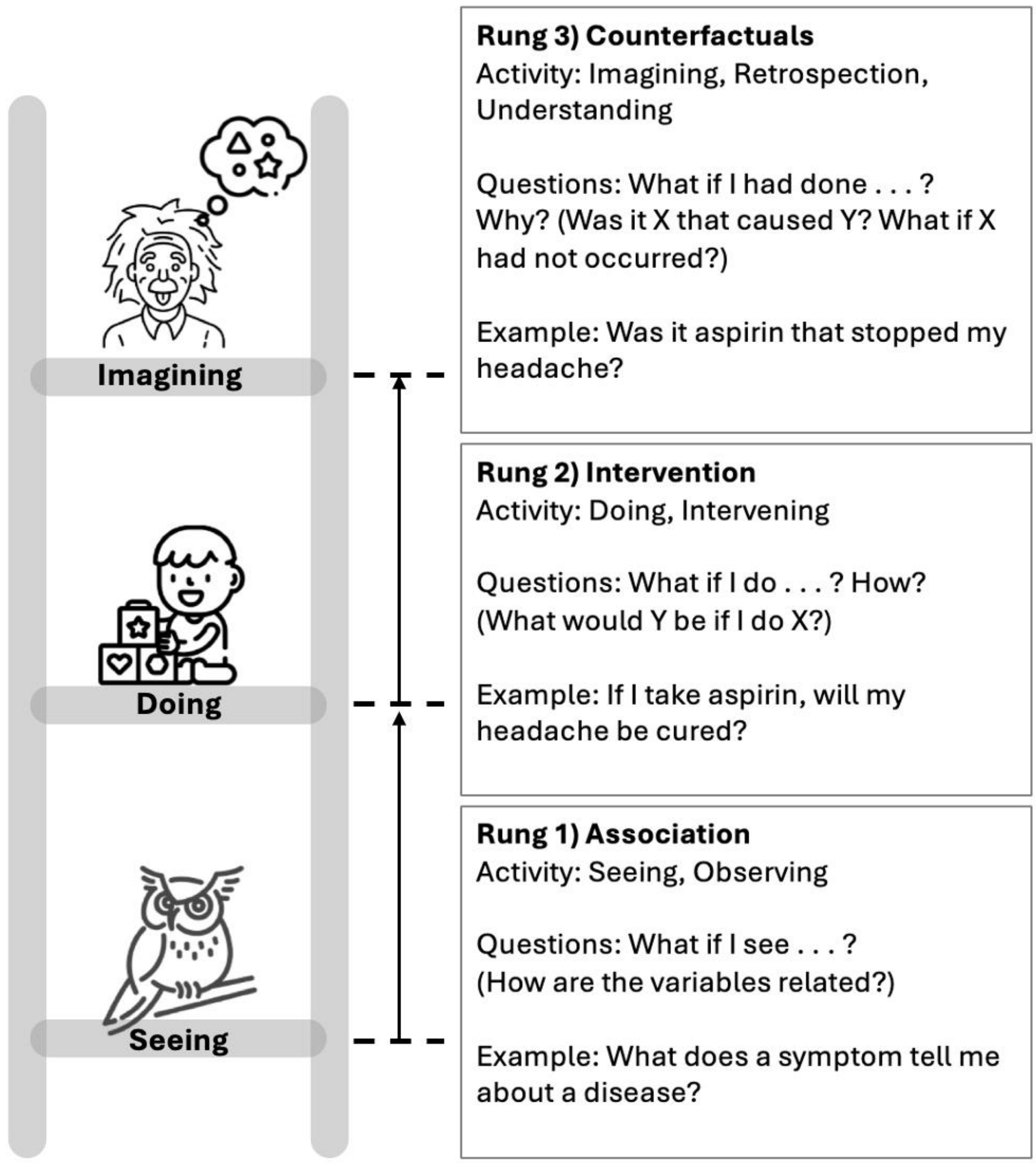


Figure 3. Pearl's Ladder of Causation. A conceptual framework delineating three distinct levels of causal reasoning: association, intervention, and counterfactuals. Each rung on the ladder represents a deeper understanding of causality. Adapted from Pearl & Mackenzie (2018).

Each of these levels requires increasingly complex analytical approaches, from basic statistical correlations to complex models that can simulate potential interventions and hypotheticals (Pearl, 2000). Expanding upon causal cognition in machines, understanding where algorithms fall on Pearl's Ladder of Causation provides significant insights into their capabilities and limitations. Most contemporary AI algorithms, particularly those in machine learning and deep learning domains, operate primarily at the first rung—association. These algorithms excel at identifying patterns and correlations in vast datasets but do not inherently understand the causal mechanisms behind these correlations (Schölkopf et al., 2021). Some advanced models, especially in reinforcement learning, arguably venture into the second rung—intervention. Such models can simulate the effects of different actions in controlled environments. These learning strategies and decision-making processes mirror rung 2 causal reasoning (Sutton & Barto, 2018). However, reaching the third rung—counterfactual reasoning—remains a significant challenge. Current AI systems lack the nuanced cognitive abilities to consider hypothetical alternatives to observed events, which are crucial for full causal understanding (Pearl & Mackenzie, 2018). This gap highlights the ongoing need in AI research to develop algorithms that not only predict, but also understand and reason about causality at a deeper, more human-like level.

To bridge the gap highlighted in AI's ability to predict versus understand and reason about causality, it is imperative to touch upon two foundational pillars in the science of causality: causal discovery and causal estimation. These components provide additional granularity to delineate the capabilities of LLMs in performing human-like causal reasoning.

1.2.2 Causal Discovery

Causal discovery is a field within the science of causality that focuses on identifying causal relationships from observational data. This field uses algorithms to construct models that can predict causal networks, which detail the interconnections and directional influences among variables. These networks are often represented using directed acyclic graphs (DAGs), which are graphs composed of nodes representing variables and directed edges that signify causal relationships, without any cycles—meaning there is no path from any node back to itself.

A directed acyclic graph (DAG) not only visually represents the causal structure, but also serves as a mathematical model that helps specify the relationships and dependencies among the variables it includes. The introduction of DAGs to represent causal relationships traces back to Sewall Wright's path analysis in the 1920s (Figure 4), initially used to study genetic influences among guinea pigs, which effectively laid the groundwork for modern causal diagrams (S. Wright, 1920)

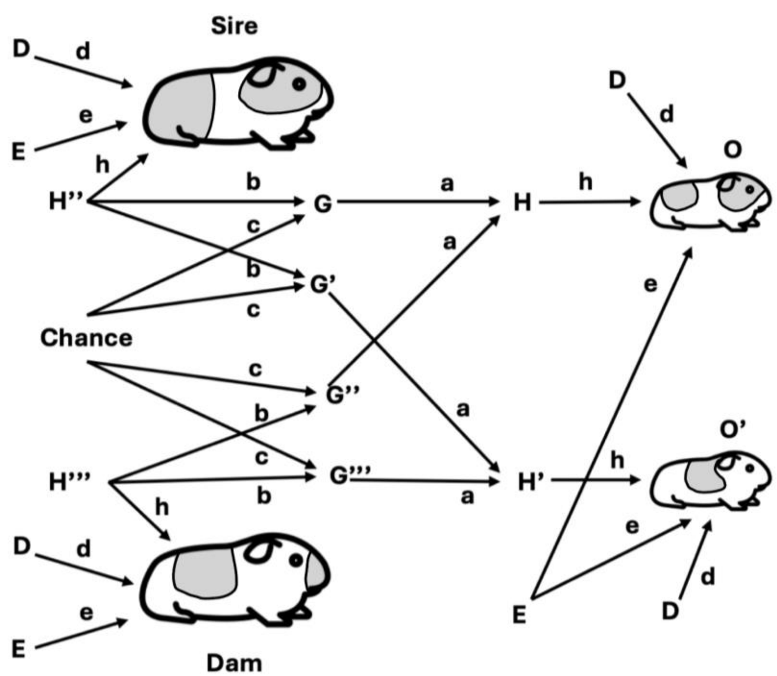


Figure 4. Sewall Wright's Path Analysis Diagram of Guinea Pig Coat Colour Genetics. Using DAGs causal relationships among genetic and phenotypic variables are depicted. Each node represents a variable (e.g. gene frequency, phenotype) while arrows denote direct causal influences between these variables. Adapted from S. Wright (1920).

In the context of cognitive neuroscience and other fields, DAGs are used to delineate the plausible causal pathways that might explain observed correlations, thus enabling researchers to make informed inferences about the underlying causal mechanisms. These diagrams can vary in complexity from simple to vast networks involving numerous variables and connections. To provide some examples, this section will explore the most basic forms of causal diagrams and explain the basic concepts with examples, then highlight how this will apply to the study of causal cognition with LLMs.

1.2.2.1 Simple Causal Diagram

The simplest causal diagram is a direct causal relationship between two variables, often represented as:

$$A \rightarrow B$$

This can be read as “A” causes “B.” This diagram indicates that changes in “A” will lead to changes in “B” and assumes that no other variables interfere. For example, in a health study, “smoking” (A) leads to “lung cancer” (B).

1.2.2.2 Mediator

A mediator variable acts as an intermediary in a causal pathway, transmitting the effect of the initial variable to the outcome variable. The structure looks like:

$$A \rightarrow M \rightarrow B$$

Here, “A” affects “B” “through “M”, where “M” is the mediator. For example, consider a study on stress (A) impacting cognitive function (B) through the mediator of cortisol levels (M) (Lupien et al., 2007). Here, cortisol levels mediate the relationship by translating the physiological impact of stress into cognitive changes. These effects would not be conferred without the mediator variable.

1.2.2.3 Fork Junction (Common Cause)

A fork junction, otherwise known as a common cause structure or a confounder, involves a single variable causing two other variables. This structure is depicted as:

$$A \leftarrow C \rightarrow B$$

Here, “C” is a common cause of both “A” and “B”. For instance, socioeconomic status (C) might influence both education level (A) and health outcomes (B). Socioeconomic status is the common factor affecting both aspects. An example in cognitive neuroscience would showcase how “neuroinflammation” (C) might lead to “cognitive decline” (A) as well as “mood disorders” (B) in aging populations (Heneka et al., 2014).

1.2.2.4 Collider

A collider is a variable that is causally influenced by at least two other variables, forming a structure like:

$$A \rightarrow C \leftarrow B$$

In this diagram, both “A” and “B” independently cause “C”. A practical example is when genetic predisposition (A) and high-fat diet (B) both lead to obesity. Here, obesity is a collider, influenced by both genetic factors and diet. An example from cognitive neuroscience might involve genetic predisposition to anxiety (A) and environmental stressors (B) both contributing to development of Post-Traumatic Stress Disorder, or PTSD (C) (Yehuda & LeDoux, 2007). In this structure, PTSD is a collider that results from both genetic and environmental causal pathways.

Understanding these different junctions in causal diagrams is critical for correctly interpreting the causal relationships in empirical data. Mediators help us understand the mechanism or process through which an effect occurs, while colliders can introduce bias if not correctly handled, particularly in statistical models. Common causes must be considered to avoid confounding effects, where an external factor influences both the independent and dependent variables, potentially misleading conclusions about their direct causal relationship. Causal inference methods have evolved significantly with the development of Judea Pearl’s do-calculus, which provides a theoretical framework for using DAGs to derive causal inferences from observational data by simulating interventions. This approach allows researchers to hypothesize and validate causal relationships, moving beyond mere correlation to rigorously test and confirm the underlying causal mechanisms within the data by integrating causal discovery and causal estimation (Pearl & Mackenzie, 2018).

1.2.3 Causal Estimation

In the evolving study of causality, once the structures of causal relationships are discerned through causal discovery, the next crucial step involves causal estimation. This process aims to quantify the strength and direction of the effects that variables have on one another within a defined causal framework. Causal estimation addresses the question "What happens to one variable when another is manipulated?" It is at the heart of moving from observational correlations (rung 1) to conclusions about causation that can inform effective interventions (rung 2).

An accepted and standard framework for causal estimation in contemporary research is Judea Pearl's do-calculus, which permits the derivation of causal effects from data by simulating interventions and observing potential outcomes (Pearl, 2000). To outline how this works, this section will provide some simple example scenarios that benefit from do-calculus.

1.2.3.1 Single Variable Intervention

Suppose we have a model where smoking "S" causes an increase in the risk of lung cancer "C". We are interested in finding out the causal effect of smoking on lung cancer. Using do-calculus notation, we want to calculate:

$$P(C\,|do(S = 1))$$

This expression tells us the probability of cancer (C) given that we intervene to make someone a smoker, regardless of their natural inclination to smoke or not. The do-operator, "do(S=1)", indicates that we're setting the smoking variable "S" to 1 (smoker) artificially, simulating an intervention.

1.2.3.2 Adjusting for Confounders

Consider a scenario where we're studying the effect of a new drug on blood pressure, and we know that age (A) is a confounder because it affects both the likelihood of receiving the drug (D) and the outcome, which is blood pressure (B). The causal effect of the drug on blood pressure while controlling for age would involve calculating the following:

$$P(B\ |do(D = 1))$$

To calculate this using observational data while adjusting for age, do-calculus might involve using a rule that allows us to "open" or "close" paths in a causal diagram to isolate the effect of the drug from the influence of age.

1.2.3.3 Mediator Analysis

Imagine we are interested in not just whether a treatment (T) affects recovery (R) but also how it does so via mediator like improved medication adherence (M). To understand the direct effect of treatment on recovery, excluding the mediation effect, we could use do-calculus to compute:

$$P(R\ |do(T = 1), do(M = natural))$$

Here, "do(M=natural)" implies that we allow the mediator "M" to take its natural value as influenced by "T" rather than intervening on "M".

These examples are conceptual and illustrate the power of do-calculus in providing mathematical notational to enable causal inference from non-experimental data in a way that not available with traditional statistical methods. The use of causal diagrams in conjunction with do-calculus helps in making assumptions about the data generation process explicit. To summarize, do-calculus provides a way to mathematically simulate an intervention with the "do" operation (e.g. do(X=x)), which isolates the effect of setting a variable X to a specific value x, independent of its usual causal influences. Beyond the examples provided, do-calculus also allows researchers to address counterfactual queries—questions about what would happen under hypothetical scenarios (rung 3). Lastly, do-calculus aids in deriving estimators that are theoretically grounded in causal assumptions, allowing statisticians and researchers to move from associative to causal inferences, which is particularly important when experimental data may be limited or unethical to obtain.

1.2.4 Types of Causal Queries

In the study of causality, researchers often aim to estimate the effects of interventions from observational data. Different types of causal queries can be used depending on the nature of the research question and the available data. Each type of query focuses on estimating different causal estimands. This section will present three of the key types of causal queries commonly addressed in the literature, one example for each rung on Pearl's Ladder of Causation.

1.2.4.1 Average Treatment Effect (ATE)

The Average Treatment Effect (ATE) is a causal estimand that measures the expected difference in outcomes between units treated with an intervention and those that are not, across the entire population. This type of query would fall on rung 2 on Pearl's Ladder of Causation. Mathematically, it is defined as:

$$ATE = E[Y(1) - Y(0)]$$

Where Y(1) and Y(0) represent the potential outcomes under treatment and control, respectively. ATE provides a comprehensive view of the treatment's impact, assuming that the treatment assignment is independent of potential outcomes—an assumption often referred to as ignorability (Rosenbaum & Rubin, 1983).

1.2.4.2 Explaining Away Effect

The explaining away effect occurs when the presence of one cause diminishes the impact or likelihood of another cause for a given effect. This effect is particularly relevant in the context of conditional probabilities where two competing causes influence a common effect (Cruz et al., 2020). A query involving this effect would fall on rung 1 on Pearl's Ladder of Causation.

To provide an example, imagine a scenario involving an alarm system in a house. Let's denote A as the event "Alarm goes off," B as the event "Burglary occurs," and C as the event "Earthquake occurs." Both a burglary and an earthquake can trigger an alarm. If the alarm goes off and you then learn there was an earthquake, the likelihood of a burglary having occurred decreases because the earthquake provides a sufficient explanation for the alarm. We can express this scenario if an alarm going off (A=1), knowing an earthquake occurred (C=1), reducing the likelihood of a burglary (B=1) using conditional probabilities:

$$P(B = 1|A = 1, C = 1)$$

The probability of B given the event A and event C encapsulate the explaining away effect, considering how C influences the likelihood of B when A also occurs.

1.2.4.3 Natural Direct Effect (NDE)

The natural direct effect (NDE) is a causal estimand used to quantify the direct influence of a treatment or intervention on an outcome, independent of a specific mediating variable. This estimand is particularly useful when aiming to isolate the effect of a treatment that acts through various pathways, distinguishing the portion of the effect that does not pass through a given mediator. This effect is typically defined in counterfactual mediation analysis literature (VanderWeele, 2011) and falls on rung 3 of Pearl's Ladder of Causation.

NDE quantifies how much the outcome would change if we could change the treatment while blocking any mediation effects, thus holding the mediator at the level it would naturally assume without the treatment. For a treatment variable X, a mediator M, and outcome Y, the mathematical notation for NDE would be as follows:

$$NDE = E[Y_{1,M(0)} - Y_{0,M(0)}]$$

Here, $Y_{1,M(0)}$ represents the potential outcome when the treatment is applied but the mediator is held at the level it would take under the control condition, and $Y_{0,M(0)}$ is the outcome under control with the mediator also at the control level. For instance, in a health intervention study, suppose a new drug is introduced to reduce stress levels (X) and one of the mediators through which the drug works is by improving sleep quality (M). The outcome (Y) is the reduction in symptoms of depression. NDE would quantify the impact of the drug on depression symptoms assuming sleep quality remains as it would be without the drug, to help isolate the direct pharmacological effects of the drug from its effects on sleep quality.

It is important to note that these examples of causal queries are not exhaustive, but merely introduce some foundational concepts essential for understanding causal estimation at each rung. This introduction of the science of causality, from Pearl's Ladder of Causation to causal discovery to causal estimation, collectively form the bedrock of causal inference today, offering a structured framework to formalize how humans conceptualize and reason about causality. Highlighting these components is important as we aim to bridge the gap between theoretical causality and practical, algorithm reasoning demonstrated by AI systems, particularly LLMs. The application of this formal framework to understand the causal reasoning capabilities of LLMs remains largely unexplored. The causal inference methodologies explored in this introduction provide a promising framework to explore causal reasoning in LLMs, particularly as causal inference bridges observing patterns in data to discerning cause-and-effect relationships, and present criticisms against the ability of LLMs to reason about causality largely center around algorithms being complex pattern recognizers.

1.3 The Current State of Understanding Causal Reasoning Capabilities of LLMs

The ability of LLMs to engage in causal reasoning has been at the forefront of discussion regarding AI's causal cognition abilities. These models have demonstrated significant capabilities in generating coherent and contextually appropriate text, suggesting a form of complex pattern recognition with the potential for causal reasoning. One could also argue that despite their sophistication, LLMs do not inherently understand causality in the way humans do, and instead replicate causal inference through pattern recognition learned from vast data sets rather than through intrinsic understanding of cause and effect (E. M. Bender & Koller, 2020; Marcus & Davis, 2019). Just within the past couple of years, there has been significant debate regarding this topic.

Earlier studies steered the conversation towards the argument that while LLMs can generate text that appears causally coherent, their “understanding” is fundamentally statistical, rooted in data co-occurrence rather than a conceptual grasp of cause and effect (Naveed et al., 2024). Empirical assessments using benchmarking tasks such as the Heuristic Analysis for Natural Language Inference Systems (HANS), suggest that LLM performance is heavily influenced by their training data’s structure, which may not necessarily reflect a true causal reasoning process (McCoy et al., 2019). This perspective is continued to be supported by the viewpoint that LLMs do not inherently understand causality but rather mimic causally-informed speech by regurgitating patterns, dubbing them “causal parrots” (Zečević et al., 2023). However, this viewpoint that LLMs are simply patterns recognizers rather than genuine reasoners is simply one side of the argument.

Conversely, proponents of LLMs’ causal reasoning abilities highlight their potential in simulating some aspects of causal understanding. Studies have shown that with appropriate fine-tuning and additional structured causal data, LLMs like GPT-3 can exhibit improved performance on tasks requiring causal inference (Evans & Grefenstette, 2018). LLMs have been used to generate hypotheses about causal relationships which are then validated through empirical data, suggesting that these models might be capable of supporting causal discovery processes in scientific research (Schwab et al., 2019). Additionally, some argue that the ability of LLMs to generalize from provided examples to novel contexts can be seen as a rudimentary form of causal reasoning, as it involves extrapolating the underlying causal mechanisms that govern observed phenomena (Zhang et al., 2023). Recently, researchers at Microsoft Research proposed that LLMs can indeed grasp some basic aspects of causal mechanisms. This revelation suggests a potential for these models to contribute novel insights into causal reasoning research (Kıcıman et al., 2023). Further contributing to this discourse, a study supported the notion that while LLMs may not independently derive causal relationships, they can effectively use pre-existing causal knowledge embedded within their training data to make seemingly causal inferences (Cai et al., 2023). This ability could be interpreted as a form of causal reasoning, albeit heavily reliant on prior human input, putting a spin on what critics previously labelled a lack of true causal reasoning. Overall, this ongoing debate sheds some light on the complexity of AI's engagement with causal reasoning—a foundational element of human cognitive processes—and highlights the importance of continued empirical research to disentangle the capabilities and limitations of LLMs in this domain.

Amid the current academic discourse surrounding the causal reasoning capabilities of LLMs lies a significant gap in the application of *formalized causal inference frameworks* to evaluate AI systems. In a recent paper, Jin et al. (2024) begin to address this by generating structured causal queries grounded in symbolic logic and Pearl's Ladder of Causation, however there remains much to be explored in this line of investigation. The foundational methodologies of causal inference, which move beyond mere pattern recognition to discern genuine cause-and-effect relationships, offer a robust template for investigating the depths of causal reasoning within LLMs. Critics argue that LLMs function primarily as complex pattern recognizers, lacking true causal understanding. Taking a causal inference approach, which requires recognizing patterns in the data, could critically transform our approach to evaluating AI cognition. Thus, leveraging this formal framework to probe the causal reasoning faculties of LLMs is not only novel and interdisciplinary, but also contributes to the ongoing debate through a perspective that was previously unexplored.

1.4 Motivation

In "Causal Cognition: A Multidisciplinary Debate," the complex interplay between causal cognition and core cognitive processes is explored. Memory crucially underpins causal cognition by archiving past causal relationships to guide future behaviour, as inferred by Tolman's cognitive maps (Tolman, 1948). Causal reasoning enables predictions based on established relationships (Gopnik & Schulz, 2007). This reasoning extends into problem-solving, where understanding causality helps identify effective solutions by pinpointing and manipulating contributing factors, especially by distinguishing them from confounders (Keil, 2006). In decision-making, insights from causal cognition for strategic planning and anticipation of outcomes, as causal beliefs shape decisions under uncertainty (Kahneman & Tversky, 1982). Together, these cognitive capabilities not only underscore human and animal adaptation but significantly inform the ongoing advancement of AI. There is an increasing recognition of need for AI that can understand and process causality, reflecting a broader call within the field of the development of causal AI (Pearl & Mackenzie, 2018), and update our notions of causal cognition as we continue to understand them better in humans and machines (E. M. Bender & Koller, 2020).

Applying the study of causal cognition to machine cognition, specifically to LLMs, the motivation is threefold:

First, as AI systems become increasingly autonomous, the ability to understand and predict the consequences of action as based on causal reasoning becomes essential for ensuring these systems operate safely and ethically. Causal understanding helps AI predict the consequences of actions and adjust behavior based on structured inference, a crucial aspect towards developing dependable autonomous systems (Russell & Norvig, 2016).

Second, as LLMs are tasked with processing complex and nuanced human languages, their ability to model and understand causal relationships directly impacts their effectiveness and reliability in real-world applications. This is particularly important in the field of health AI, where discerning between a confounder and a true causal risk factor has significant implications for patients, public health, and vulnerable populations. Understanding causality improves AI's performance across various applications such as epidemiology and economics, but also to improve transfer learning in models (Bengio et al., 2019).

Third, there remains much to be gained from comparative studies assessing causal cognition in intelligent systems like LLMs versus human and non-human animals' causal cognitive abilities. Such studies enrich our understanding of intelligence and cognitive processes across different entities, providing insights that can lead to improved AI systems, build upon and inspire improvement of what we have learned about human and animal intelligence.

By studying causal cognition comprehensively, from its biological roots to its implementation in artificial systems, we can better understand a fundamental aspect of intelligence that is crucial for survival and sophisticated interaction with the world. This understanding can lead to advancements in cognitive neuroscience, artificial intelligence, and various applied domains where predicting and understanding causality is key.

1.5 Thesis Overview

This introduction (Chapter 1) explored a history and background of cognition as it relates to artificial intelligence, causal cognition, a preliminary background of causal inference, and the background and motivation for this study. The remaining components of this thesis aim to address two fundamental questions about the causal reasoning capabilities of LLMs.

In Chapter 2, we will explore the first research question: *how well do LLMs perform on tasks that require causal reasoning?* This question seeks to evaluate the capabilities of current LLMs in understanding and applying causal knowledge in structured scenarios.

In Chapter 3, we address the second research question: *what type of information is important in the prompt, and how will manipulating this information affect LLMs' causal reasoning abilities?* This explores the efficacy of different prompting strategies, which is crucial for optimizing the performance of LLMs on tasks requiring a nuanced understanding of causality.

Chapter 2

2 Can Large Language Models Do Causal Inference?

In this section, we turn our attention to exploring the first research question*: how well do large language models (LLMs) perform on causal reasoning tasks?* This inquiry challenges LLMs—specifically frontier models such as OpenAI's GPT-4 and Anthropic's Claude 3 Opus—not only to generate linguistically coherent outputs but also demonstrate a deeper cognitive capability by assessing their accuracy in responding to questions that require causal inference to correctly answer. In doing so, the aim is to assess an LLMs' ability to process and potentially reason about causal relationships in a formalized framework.

Previous research in understanding causal reasoning in LLMs has often focused on commonsense causality (Ho et al., 2022; Zečević et al., 2023; Zhang et al., 2023), treating these models as knowledge bases (Jiang et al., 2020; Petroni et al., 2019; Shin et al., 2020). This approach, however, does not adequately test a model's ability to perform causal inference based on formal rules based on the framework outlined by Pearl's Ladder of Causation. Our study aims to fill this gap by applying the structured framework of causal inference to assess LLMs' causal reasoning capabilities. There is a concern that LLMs might merely replicate causal information from their training data without genuine inference, essentially acting as "causal parrots." To investigate this, we use the Causal Ladder (CLADDER) dataset, which consists of questions designed to evaluate and benchmark formal causal reasoning abilities of LLMs (Jin et al., 2024).

2.1 The CLADDER Dataset

The CLADDER dataset is described in detail in a paper presented at the Neural Information Processing Systems (NeurIPS) conference (Jin et al., 2024). In this section, the key aspects of the CLADDER dataset as it relates to this study will be described, alongside the motivation for choosing this dataset for this study.

The CLADDER dataset was composed to address the gap in reasoning benchmarking tasks in evaluating causal inference in natural language. The dataset consists of causal questions posed in natural language that are grounded in symbolic query and ground truth answers derived through the causal inference engine (Figure 7), which abides by the rules of causal inference. Overall, the CLADDER dataset comprises of 10,112 questions that are distributed to probe and assess the ability to answer various causal queries relating to each rung of Pearl's Ladder of Causation (Figure 5).

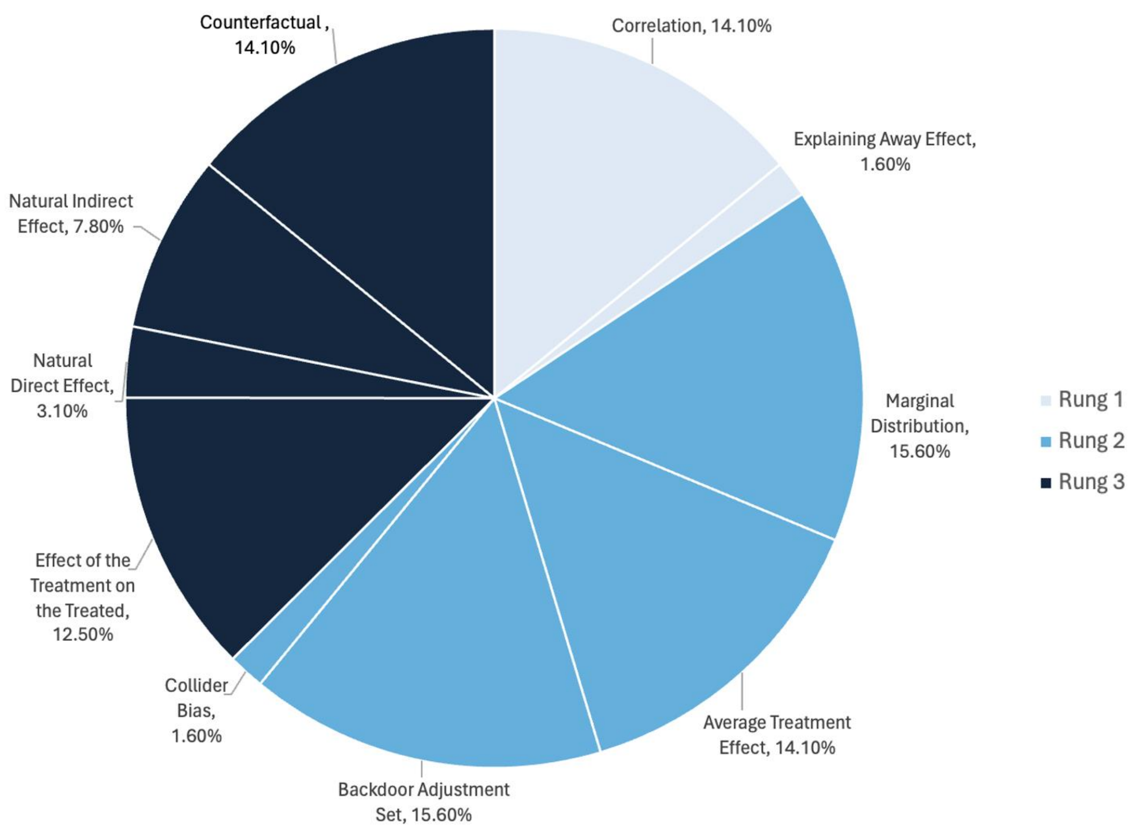


Figure 5. Distribution of CLADDER Dataset on the rungs of Pearl's Ladder of Causation and types of causal queries, adapted from Jin et al. (2024).

The questions are also structured around several types of causal graphs (Appendix 1), which give rise to scenarios that require different causal inference abilities. For each question, a query, information about the model, and other context data about the scenario is included. Additionally, ground-truth explanation with step-by-step reasoning is also included in the dataset (Figure 6).

Question: Imagine a self-contained, hypothetical world with only the following conditions, and without any unmentioned factors or causal relationships:

Encouragement level (X) has a direct effect on **studying habit (O)** and **exam score (Y)** . **Studying habit (O)** has a direct effect on **exam score (Y)**.

For students who are not encouraged, the probability of high exam score is 12%. For students who are encouraged, the probability of high exam score is 64%.

Will encouragement decrease the chance of a high exam score?

Ground truth answer: No

Correct steps leading to the ground-truth answer:

1) Parse the causal graph: **Mediation**
   Subskill: Causal Relation Extraction

2) Classify the query type: **Average Treatment Effect**
   Subskill: Causal Question Classification

3) Formulate the query to its symbolic form:
   E[Y | do(X = 1)] - E[Y | do(X = 0)]
   Subskill: Formalization

4) Collect the available data:
   P(Y=1 | X=0) = 0.12
   P(Y=1 | X=1) = 0.64
   Subskill: Semantic Parsing

5) Derive the estimand using causal inference:
   P(Y=1|X=1) - P(Y=1|X=0)
   Subskill: Formal Causal Inference

6) Solve for the estimand by plugging in the relevant data in Step 4
   0.64 - 0.12 = 0.52
   0.52 > 0 therefore encouragement increases the chance of a high exam score, so the final answer is “No” (it does not decrease)
   Subskill: Arithmetic

Figure 6. Example of a question in the CLADDER dataset and Formal Correct Answering Steps. The initial question posed in natural language, followed by a symbolic representation of the causal graph and query, and steps detailing the step-by-step process including formulating the causal estimand and applying appropriate do-calculus and probabilistic calculations. Adapted from Jin et al. (2024).

2.1.1 CLADDER Dataset Generation

The question generation process for the CLADDER dataset was guided by the causal inference engine (Figure 7). Initially, a series of inputs—comprising triples of causal queries, graphs, and corresponding data—was created. These triples are designed to ensure that each query has a clear, definitive ground truth answer that can be derived based on available data. LLMs are known to struggle with calculation-heavy tasks (Hendrycks et al., 2021; Stolfo et al., 2023), and since the aim of this dataset is the test the causal reasoning capabilities of LLMs rather than their ability to perform complex calculations, graphs with three to four variables arranged in common configurations were used. These configurations were selected from a variety of sources within the literature such as statistics textbooks, where graph structures are typically used to demonstrate foundational problems in causal inference.

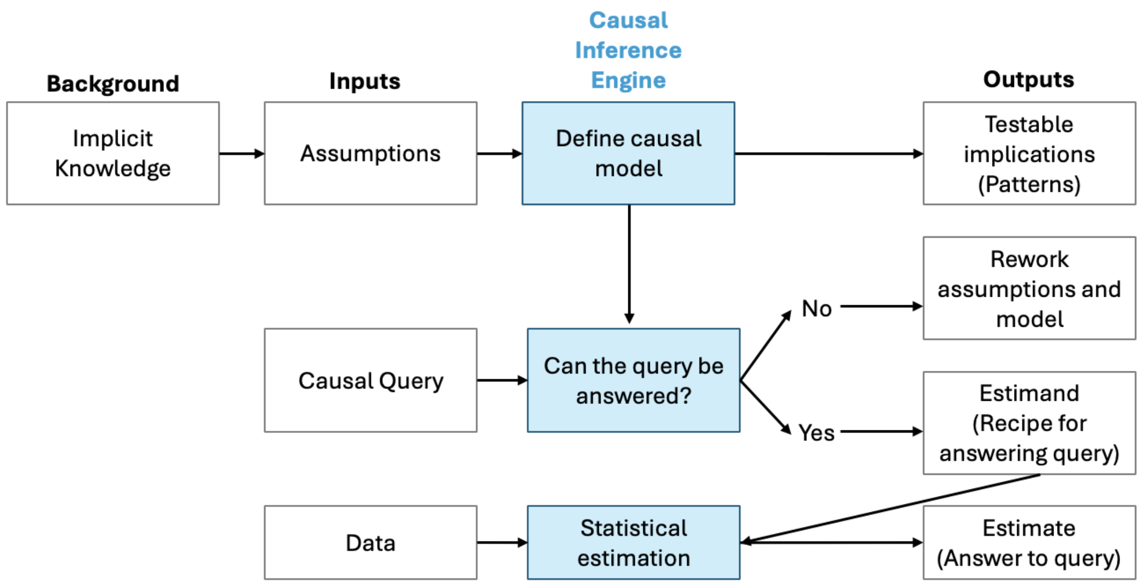


Figure 7. Causal Inference Engine Schematic. Inputs are fed into a causal inference engine, which processes them to compute an estimable expression of the query, provided a viable solution exists. This process is designed to compute if a causal query can be answered given the data and model, and if so, solve the query and provide an answer for the causal estimand. Adapted from Pearl & Mackenzie (2018).

To collect common query types in each rung, Jin et al. (2024) continued to draw from causal inference literature. Several different types of queries can be categorized under each rung of Pearl's Ladder of Causation. On rung 1, queries primarily involve probability distributions such as marginal and conditional probabilities, testing an LLM's capacity to accurately respond to these foundational questions that have to do with association. Rung 2 queries can include the concept of Average Treatment Effect (ATE), probing questions like "how will the outcome change if treatment (A) changes from x' to x?" Finally, for rung 3, the focus shifts to counterfactual reasoning, which explore hypothetical alterations in causal scenarios (Appendix 2).

Next, Causal inference rules on specified graphs and queries were applied to derive a causal estimand. For queries on Rung 2 of Pearl's Ladder, simplification using the rules of do-calculus to reduce terms to Rung 1 (conditional and marginal probabilities) was used to solve the estimand. Rung 3 queries require the application of counterfactual causal inference methods to evaluate the estimand and establish a ground truth answer. The estimand specifies the necessary terms to include as "available data" in the prompt, ensuring that sufficient information is provided to correctly determine the answer to the question.

Lastly, symbolic questions and answers are then transformed into narratives understandable in natural language. For each causal graph, two to five stories were constructed, each story incorporating variable names as node identifiers within the graph. These stories were selected from examples in commonly cited causal inference literature. The verbalization process involves converting symbolic variables into semantic concepts to create a plausible narrative of the underlying causal mechanisms. This narrative is then translated into natural language through templates developed by Jin et al., (2024). The code base for the data generation process can be found at the following GitHub repository: https://github.com/causalNLP/cladder

In summary, the dataset was generated through an algorithmic procedure, which benefits from zero human annotation cost, controllability, and a decreased likelihood that the data was previously seen before by the model during the training phase of LLMs.

2.2 Why CLADDER?

2.2.1 Unexplored Areas that Benefit from Using this Dataset

Jin et al. (2024) evaluated multiple LLMs using the CLADDER dataset, reporting an overall accuracy of 62.03% by GPT-4, indicating the challenging nature of the task for LLMs. **However, their analysis lacked granularity beyond the categorization of questions into different rungs of Pearl's Ladder of Causation**. Specifically, they did not explore variations in LLM performance based on the type of causal query or structure of causal graph involved in the questions. Instead, their focus was on comparing LLM responses across narratives that aligned with commonsense reasoning versus those that involved anti-commonsense and formal causal inference reasoning. Furthermore, the rapid evolution in the field of generative AI, marked by continual advancements and new releases from OpenAI, Anthropic, and various open-source models, has quickly outdated the findings of Jin et al., (2024). This creates a need for ongoing research to keep pace with technological developments in this area.

2.2.2 Uniqueness from other reasoning benchmarking tasks for LLMs

Currently, benchmarking tasks that relate to causal reasoning typically fall into one of three categories (Zhang et al., 2023):

> Type 1: Identifying causal relationships using domain knowledge. Tasks in this category primarily test LLMs' ability to use stored world knowledge to infer causal relationships. These are generally the simplest type of casual reasoning tasks, requiring LLMs to retrieve and apply information from their training data rather than generate new causal insights.
>
> Type 2: Discovering new knowledge from data. These tasks are designed to assess whether LLMs can generate new knowledge by analyzing given data sets. Challenges include interpreting business strategies or predicting medical treatment outcomes based on historical data. This category requires a higher level of causal reasoning, engaging with intervention-level reasoning.
>
> Type 3: Quantitatively Estimating the Consequences of Actions. The most complex category, these tasks involve quantifying the effects of specific actions, such as adjusting medication dosages based on prior outcomes. These tasks demand that LLMs apply principles of counterfactual reasoning to propose actionable solutions or predictions.

The majority of existing benchmarks can be categorized as Type 1 tasks, which primarily assess an LLM's ability to recognize causal relationships through domain knowledge, often resulting in evaluations of simple associative reasoning (Ho et al., 2022; Zečević et al., 2023; Zhang et al., 2023). This approach does not robustly challenge the LLM's deeper causal reasoning capacities. In contrast, the CLADDER dataset is designed to cover the full spectrum of benchmarking tasks—Type 1 through Type 3—thereby allowing for a nuanced assessment of LLMs' abilities to differentiate between mere correlation and true causation, as well as compare LLMs' performance between the different types of questions. As such, the CLADDER dataset is an essential tool for this thesis. Additionally, it includes metadata for each question and outlines formal causal inference steps. The enables granular analysis into the types of questions on which LLMs excel or underperform, highlighting specific limitations in their causal reasoning capabilities. Formal causal inference steps enable the investigation of specific type of information necessary for LLMs to arrive at accurate conclusions, which contributes towards answering the second research question. Overall, these features are vital for addressing the research questions posed in this thesis, allowing for a deeper exploration of LLMs' capabilities in causal inference.

2.3 Methods

2.3.1 Filtering the data

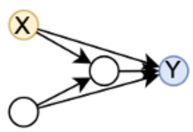
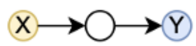
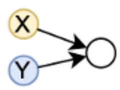
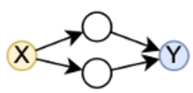
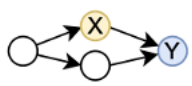
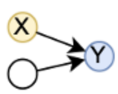
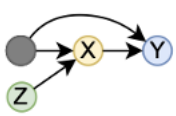
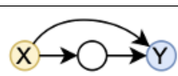
For the experiments conducted in this thesis, only questions from the CLADDER dataset that included full reasoning steps were included. This selection criterion was necessary to maintain consistency across the experiments in both this and the subsequent chapter, facilitating a direct comparison of different prompting strategies. After filtering the CLADDER dataset to only include questions with the full reasoning steps, 1392 questions were used for the experiments, with the following distribution across rungs, queries (Table 1), and graphs (Table 2).

Table 1: Number of Questions for Each Rung and Query Type in the Filtered CLADDER dataset

Query Type	Rung	Number of Questions Using this Query Type	Percentage of Total Dataset (1392 Questions)
Correlation	1	296	21.3%
Explaining Away Effect	1	32	2.3%
Marginal Distribution	1	296	21.3%
Average Treatment Effect	2	264	19.0%

Collider Bias	2	32	2.3%
Effect of the Treatment on the Treated	3	224	16.1%
Natural Direct Effect	3	96	6.9%
Natural Indirect Effect	3	152	10.9%

Table 2: Number of Questions for Each Graph Type in the Filtered CLADDER dataset

Graph Type	Graph Structure*	Number of Questions Using this Graph	Percentage of Total Dataset (1392 Questions)
Arrowhead		288	20.7%
Chain		120	8.6%
Collision		128	9.2%
Confounding		128	9.2%
Diamond		120	8.6%
Diamond Cut		32	2.3%
Fork		128	9.2%
Front Door		40	2.9%
Instrumental Variable		120	8.6%
Mediation		288	20.7%

*X represents the treatment variable, Y represents the outcome variable, and Z represents the mediator variable

2.3.2 Simple Prompting

To explore how current LLMs respond to questions that require varying levels of causal inference to answer correctly, a simple prompting approach was used to establish a baseline. This prompting approach simply fed each question from the filtered dataset in the simplest, most direct way. Each prompt contained three main components—background, given information about the question, and the question itself—to provide the model with the same amount of information that would be considered sufficient for a human to go through the causal inference steps and get to the ground truth answer. An example of a prompt is presented below, broken down by the three components:

1. **Background**: "Imagine a self-contained, hypothetical world with only the following conditions, and without any unmentioned factors or causal relationships. Encouragement level has a direct effect on studying habit and exam score. Studying habit has a direct effect on exam score."
2. **Given information about the question**: "For students who are not encouraged, the probability of high exam score is 12%. For students who are encouraged, the probability of high exam score is 64%."
3. **Question**: "Will encouragement decrease the chance of a high exam score?"

For the simple prompting approach, these three components were read from the filtered dataset (stored in a JSON file) and concatenated to form a prompt (Appendix 4). Each question was passed through the OpenAI API Server with temperature 0 to prompt GPT-3.5 Turbo (20 billion parameters), GPT-4 (an estimated 8 language models, each with 220 billion parameters), and GPT-4 Turbo (number of parameters not released to the public) using the filtered dataset of 1392 questions. Similarly, the Anthropic API Server, also with temperature 0, was used to prompt Claude 3 Opus (an estimated 2 trillion parameters) with the same dataset and prompting strategy. These frontier models were selected due to their prevalence in current reasoning research (López Espejel et al., 2023) as well as their demonstrated ability to reason about mathematics (Zhou et al., 2023) and conceptual structure (Singh et al., 2023). On the server side, there is no memory cache that each API call is pulling from (Zheng et al., 2023), as OpenAI and Anthropic do not maintain chat history for the API. In the context of these experiments, this means that for each causal inference question passed to the model in the form of API calls, the model does not 'learn' from past questions and answers each questions separately from the others.

Before touching upon the assessment of responses from LLMs, it is important to note that for each question, the instruction from the 'system' role (a parameter that can be manipulated in the API call) guided the LLM to respond with the first word being "Yes" or "No" using the following system message: "You are an expert in causal inference. The following question is not a typical commonsense query, but rather a meticulously designed question created by a professor specializing in causal inference, intended to assess mastery of the course content. Start your answer with "yes" or "no," followed by additional reasoning or evidence to support your explanation." As such, for the scoring, the first word of the LLM response was taken and compared to the ground truth (not case sensitive). A simple true or false scorer was used that accepted any of "No," "False," "Incorrect," or "Not necessarily" as a member of the false family, and any of "Yes," "True," or "Correct" as a member of the true family. If the first word of the LLM response belongs in the false family, and the ground truth answer for that question was "No," then the scorer assessed the response as correct (i.e. the prediction matches the ground truth). Similarly, if the first word of the LLM response to the true family, and the ground truth answer for that question was "Yes," then the scorer assessed the response as correct. If family of words that the LLM responds with does not match the ground truth, then the response was assessed as incorrect. The accuracy was determined for LLM responses using the following formula:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Metadata about the response types were saved alongside the responses in a JSON file, and accuracies were calculated for overall, query type, and graph type. Excel version 16.78 was used to generate bar graphs, and the Seaborn package in python version 3.11 was used to generate the heatmaps.

2.4 Results

For the response accuracy on the 1392 causal inference questions, GPT-3.5 Turbo had an overall accuracy of 59.41%, GPT-4 had an overall accuracy of 73.71%, GPT-4 Turbo had an overall accuracy of 72.20%, and Claude 3 Opus had an overall accuracy of 69.97% (Figure 8 and Appendix 3). For GPT-3.5 Turbo and GPT-4, these overall accuracies can be compared with the performance on the entire CLADDER dataset (10,112 questions) reported by (Jin et al., 2024). For GPT-3.5, they reported an accuracy of 52.18% and for GPT-4, an accuracy of 62.03%.

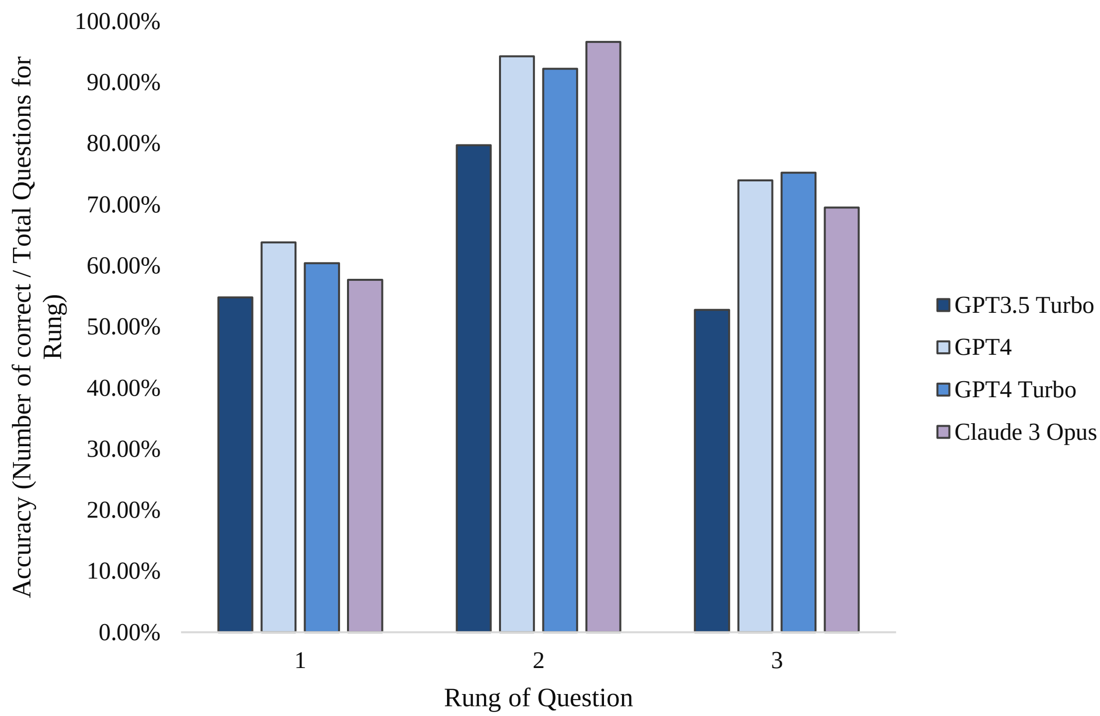


Figure 8. Accuracy of responses to causal inference questions by LLMs broken down by the rung on Pearl's Ladder of Causation that the question assesses.

Overall, the group of LLMs that we studied answer questions that require rung 2 level thinking on Pearl's Ladder of Causation with higher accuracy than rung 3 and rung 1 level questions. To assess whether this is due to higher or lower accuracy in responses for a particular query or graph type skewing the overall accuracy, accuracies were broken down by query type and graph type (Figure 9).

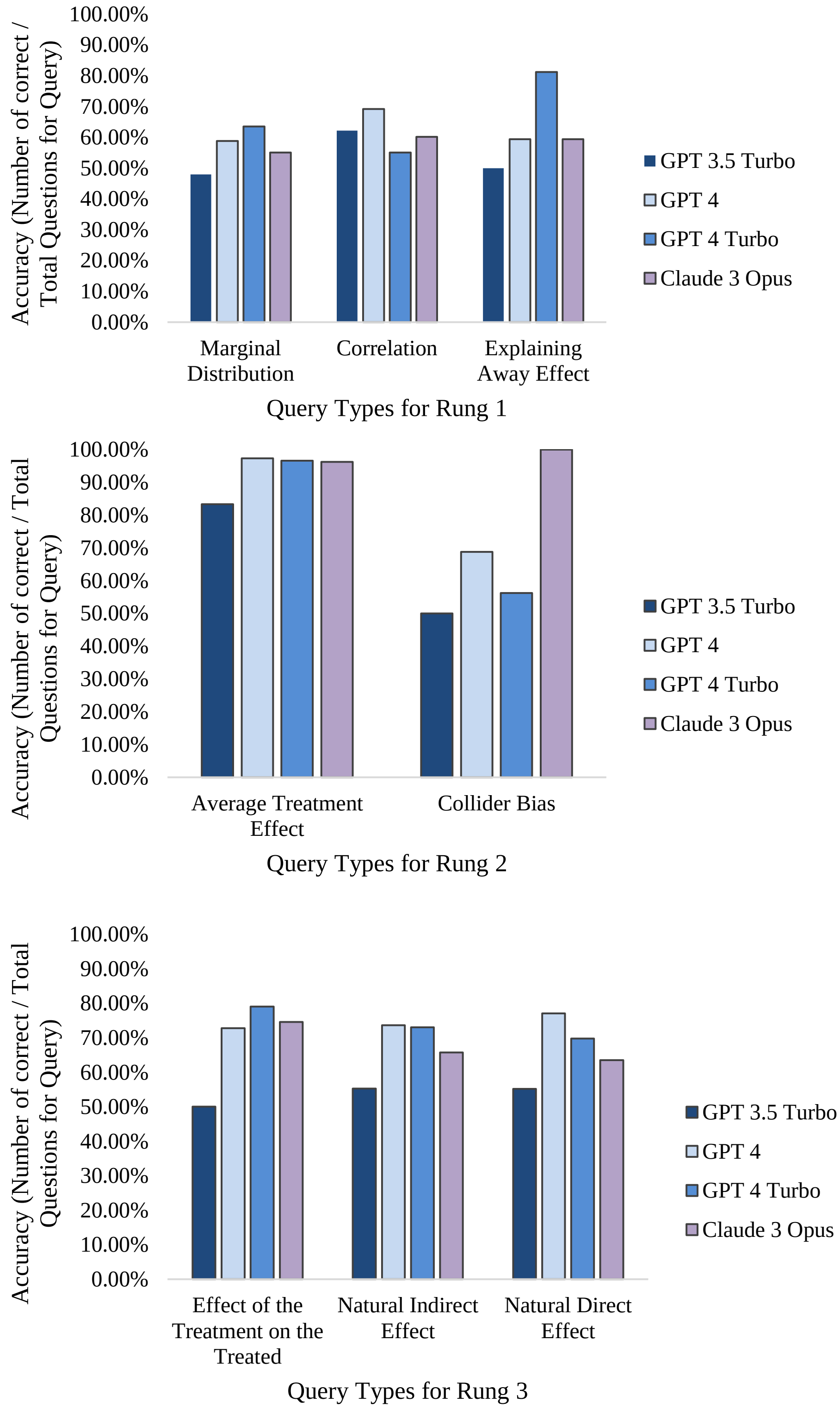


Figure 9. Accuracy of responses to causal inference questions by LLMs organized by the query type.

Lastly, to examine whether LLMs performed with a higher accuracy on a specific type of graph within a specific rung on the Ladder of Causation (since each type of graph can be used for questions across many different rungs), heatmaps were generated to organize this information (Figures 10, 11, 12, and 13).

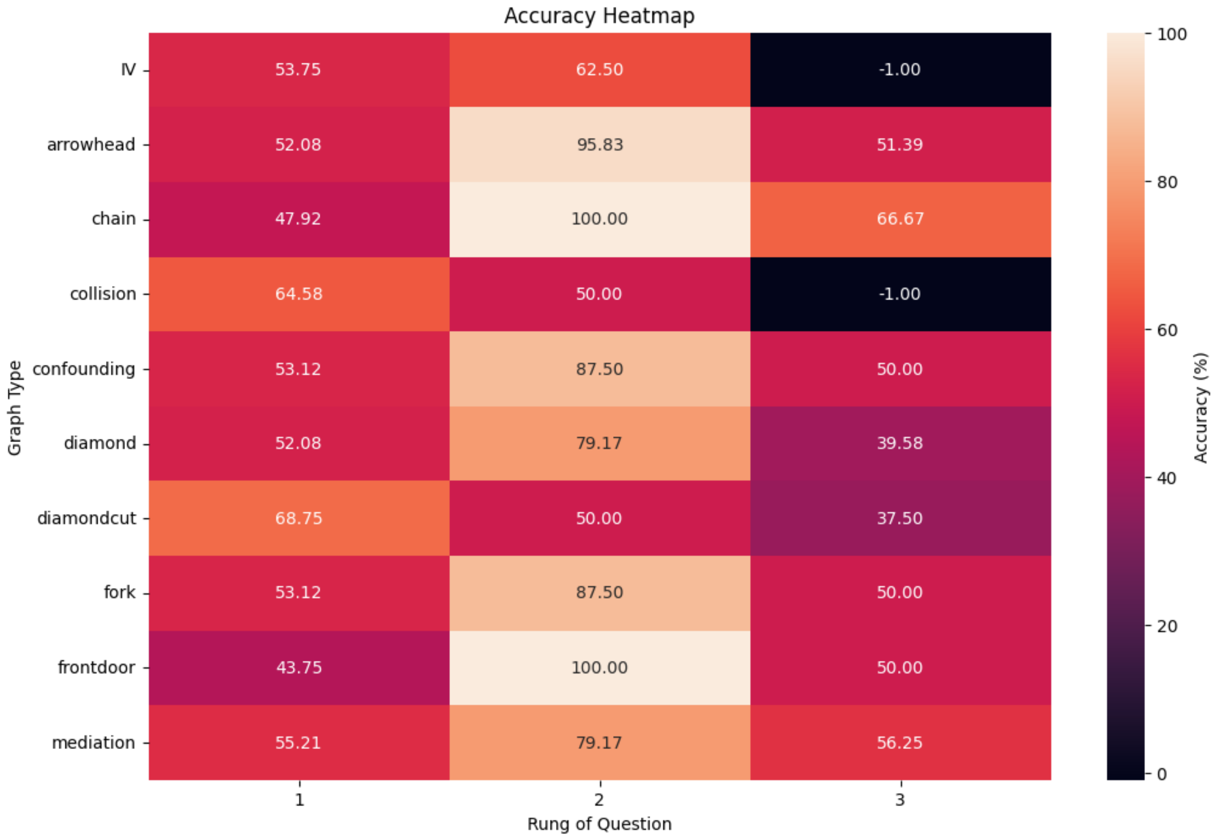


Figure 10. Accuracy heatmap of GPT-3.5 Turbo responses to causal inference questions organized by rung and graph type. A value of -1.00 denotes that there were no questions to evaluate the intersection of a particular rung type and graph.

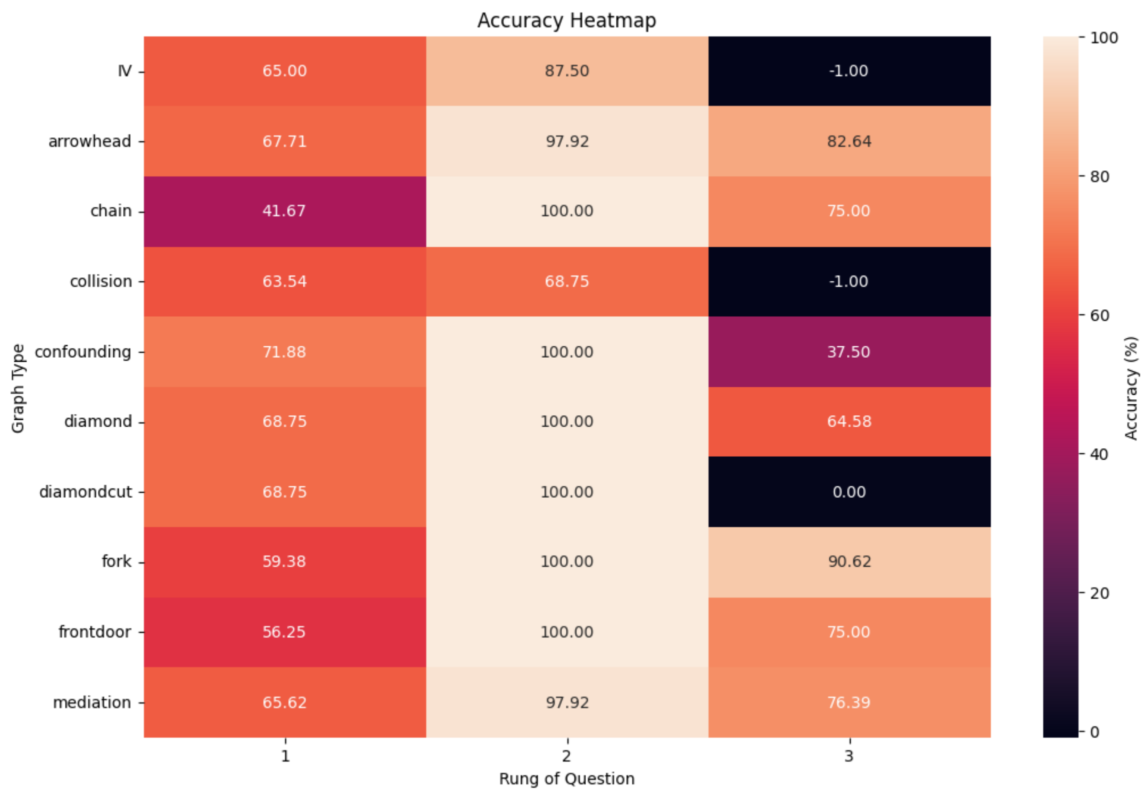


Figure 11. Accuracy heatmap of GPT-4 responses to causal inference questions organized by rung and graph type. A value of -1.00 denotes that there were no questions to evaluate the intersection of a particular rung type and graph.

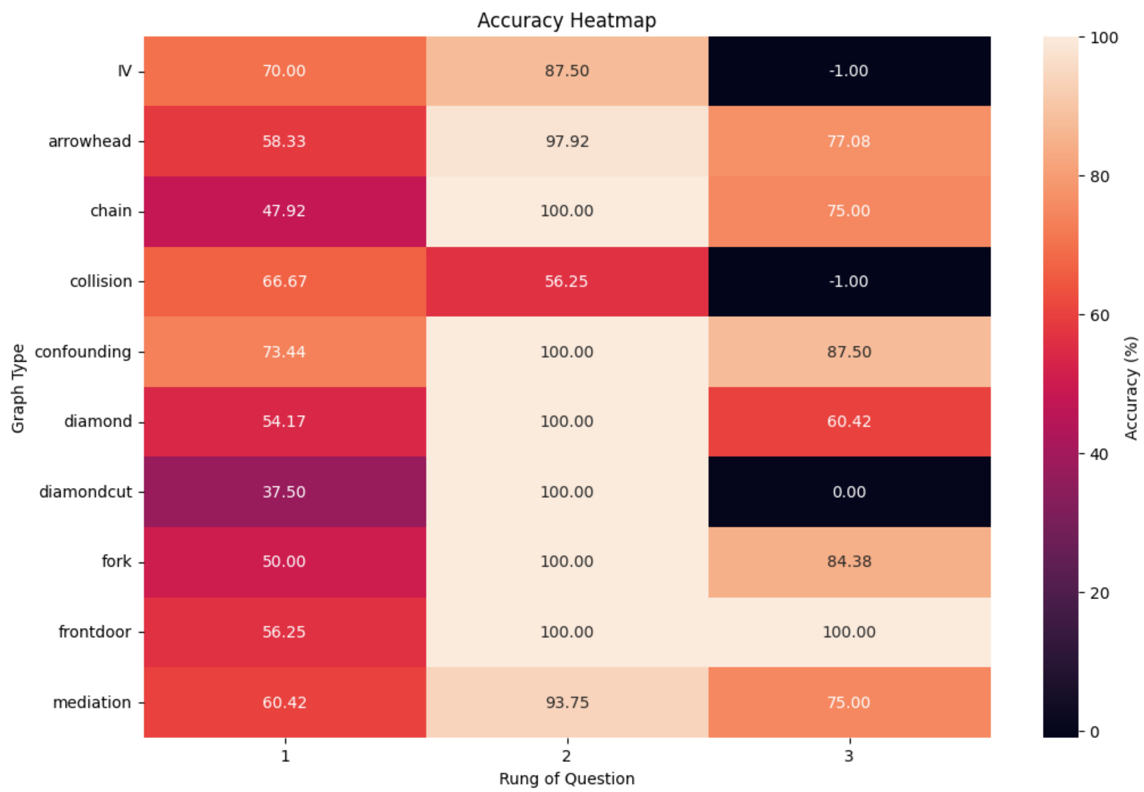


Figure 12. Accuracy heatmap of GPT-4 Turbo responses to causal inference questions organized by rung and graph type. A value of -1.00 denotes that there were no questions to evaluate the intersection of a particular rung type and graph.

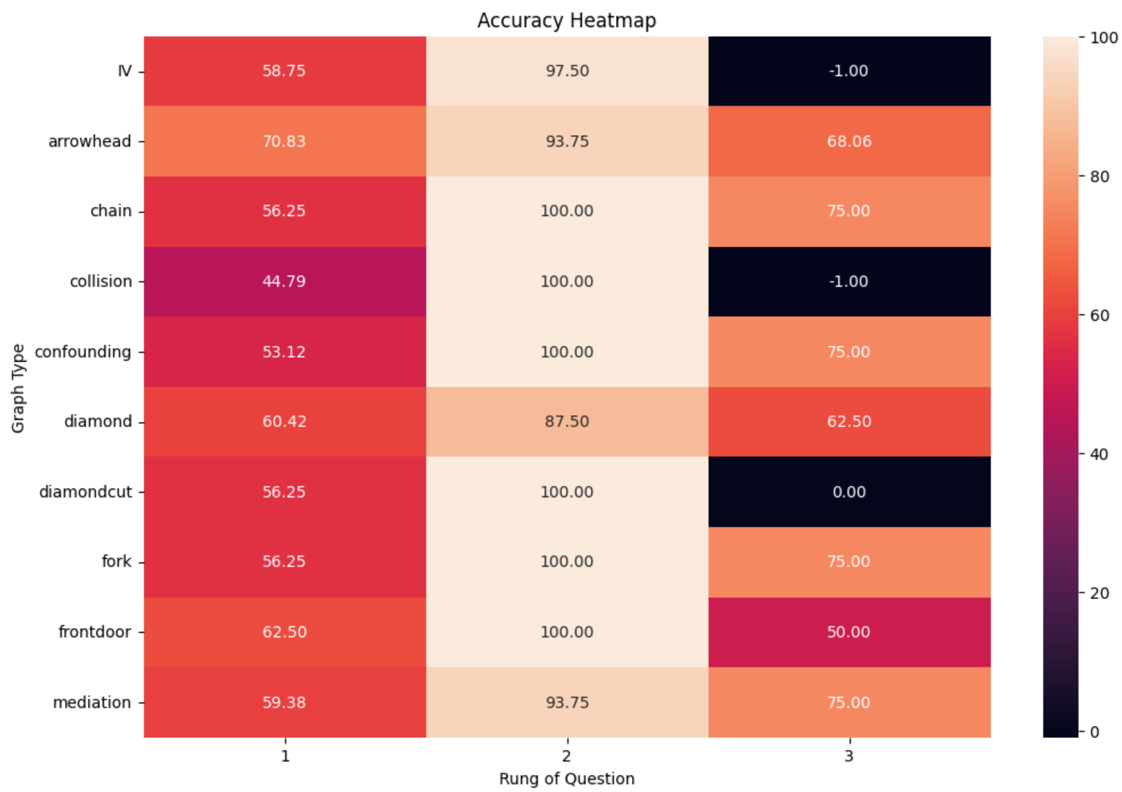


Figure 13. Accuracy heatmap of Claude-3 Opus responses to causal inference questions organized by rung and graph type. A value of -1.00 denotes that there were no questions to evaluate the intersection of a particular rung type and graph.

2.5 Discussion

2.5.1 Differences in Accuracy Between Different Rungs on Pearl's Ladder of Causation

Our study using the filtered CLADDER dataset to evaluate causal reasoning in LLMs reveals important insights into their capabilities and limitations. Notably, our results indicate that LLMs perform more accurately on questions that involve interventions (rung 2 on Pearl's Ladder of Causation) compared to those dealing with association (rung 1) or counterfactual reasoning (rung 3), as depicted in Figure 8. This difference in performance suggests that LLMs are better at handling scenarios that require direct manipulation of variables rather than those necessitating the understanding of hypothetical situations when it comes to formal causal reasoning. Interestingly, their weaker performance on associative and counterfactual questions may challenge the prevalent assertion that LLMs can answer Type 1 causal questions thanks to their large collection of knowledge (Zhang, 2024). The CLADDER dataset, despite being built to assess formal causal reasoning, draws from highly cited sources such as textbooks to generate the stories used in the questions. Further investigation into the prompting strategies, training data, and model architecture that influence the accuracy of Type 1 question responses is required to determine the limitations of LLMs' performance on these types of tasks. The results from this experiment show that in a formal causal reasoning task, LLMs have greater difficulty answering questions of association and counterfactuals compared to questions of intervention.

2.5.2 Differences in Accuracy of Response Between Different Models

The disparity in performance on interventional versus counterfactual questions highlights a fundamental challenge in LLMs' ability to engage in complex causal reasoning. Counterfactual reasoning, which involves imagining alternative realities and outcomes, is a cognitively challenging task that differentiated the abilities of the models tested in this experiment. Models with a greater number of parameters such as GPT-4, GPT-4 turbo, and Claude 3 Opus performed with high accuracy than GPT-3.5 Turbo on rung 3 questions as seen in Figure 8, and this was the largest difference in accuracy between different models for questions across the same rung type. GPT-3.5 Turbo performed just above random chance (52.75%) contrasted with the higher performance accuracies of GPT-4 (73.94%), GPT-4 Turbo (75.21%), and Claude 3 Opus (69.49%). As shown in Figure 9, this difference in accurately is seen consistently for rung 3 query types, and for rung 3 graphs (Figures 10, 11, 12, and 13). Such findings challenge the assertions regarding causal reasoning in LLMs in studies using models like GPT-3.5 turbo, which may have prematurely concluded LLMs' incapacity of abstract causal reasoning. This raises the potential for further improvements in LLMs, particularly in their ability to handle complex causal questions. Supporting this notion, recent advancements in causal AI emphasize the importance of designing AI systems that can integrate and reason with causal models (Bengio et al., 2019). This could lead to LLMs that not only better understand causality but also more effectively apply it in practical, real-world scenarios. The ongoing development and scaling of LLMs paralleled with the ongoing research in causal AI is moving in a promising direction, towards overcoming some of the limitations observed in earlier models and potentially leading to more robust and capable systems in terms of causal reasoning.

Conversely, there are specific intersections of graph and rung types where GPT-3.5 Turbo outperformed the newer models. Specifically, for the diamond cut graph type questions, while GPT-4, GPT-4 Turbo, and Claude 3 Opus performed well on rung 1 and rung 2, these newer models answered rung 3 diamond cut graph type questions with 0.00% accuracy, while GPT-3.5 Turbo responses had a 37.50% accuracy on these questions. This is likely due to small sample size error, as there were only 32 diamond cut graph type questions in the filtered dataset (16 were rung 1 questions, 8 were rung 2 questions, and 8 were rung 3 questions). Small sample size error in question types like diamond cut, which are also less common in the literature, is a limitation to this study. However, this highlights a gap not only in LLM and AI literature regarding the reasoning capabilities with respect to data organized in these structures, but also in causal inference. Due to the number of possibilities of different graph structures that could be formed simply with 3 or 4 variables, not all of them are equally explored and documented. The diamond cut graph structure is one such possibility that has not been widely investigated but can be applied in future studies.

Another notable difference in accuracy was observed between Anthropic's Claude 3 Opus and OpenAI's GPT models for collider bias questions (Figure 9). Collider bias occurs when an exposure and outcome each influence a third common variable and that variable is controlled for by design or analysis (Holmberg & Andersen, 2022). In other words, this specific type of bias arises when conditioning on a common effect of two or more causes. An example of collider bias could be if researchers are studying the effects of a new drug on heart disease risk. If they control for a variable like cholesterol level—which the drug itself affects—then they might observe misleading relationships between the drug and other risk factors due to collider bias introduced by conditioning on cholesterol. Claude 3 Opus performed perfectly on these types of questions, with an accuracy of 100.00% on collider bias queries, whereas the OpenAI models performed with much lower accuracy. GPT-3.5 Turbo was correct in answering 50.00% of the questions, while GPT-4 Turbo was correct with 56.25% of questions, both answering around as accurately as random chance, whereas GPT-4 was correct in 68.75% of collider bias queries (Figure 9). This highlights a difference that can arise due to different data used by different organizations and individuals when training LLMs.

However, this discrepancy could also be prone to small sample bias, as there was only a total of 32 questions about collider bias in the dataset. Importantly, the DAG that was required to correctly reason through all collider bias questions were collision type graphs. There was a total of 128 questions corresponding to collision type graphs in the dataset (96 for rung 1, 32 for rung 2). In Figures 10, 11, 12 and 13, we can see questions involving collision type graphs consistently brought down the accuracy of rung 2 type questions for OpenAI GPT models—these are the 32 collider bias questions. What is of interest is the models' accuracy on collision type questions for rung 1. We can see that for Claude 3 Opus, even though it answers with perfect accuracy on collider bias questions, the accuracy of rung 1 questions that use a collision type graph was 44.79% (i.e. about random chance). Rung 1 type questions using a collision graph was also challenging for OpenAI GPT models (64.58% accuracy with GPT-3.5 Turbo, 63.54% accuracy with GPT-4, and 66.67% accuracy with GPT-4 Turbo). Overall, this finding supports the previous notion that LLMs are better at reasoning with rung 2 type questions, given that Claude 3 Opus seemingly performs well with collider bias questions but poorly with collision type rung 1 questions, but further investigation on collision type problems with a larger dataset should be conducted to eliminate the possibility of small sample bias.

Chapter 3

3 What Information Do Large Language Models Need for Causal Reasoning?

In this chapter, we address our second research question: *What type of information is important in a prompt to help LLMs answer causal inference questions?* We explore how variations in the information presented have an impact on the accuracy of LLM response to the same filtered CLADDER dataset used in the previous chapter. The motivation for line of inquiry includes 1) to understand the components of effective prompts when it comes to causal inference questions, 2) to explore whether LLMs benefit from structuring information in the form of a causal inference engine, and 3) to further delineate the boundaries of LLMs' causal reasoning proficiency under different information conditions.

Prompt engineering is an emerging field in the domain of artificial intelligence that focuses on optimizing the prompt inputs to improve the performance of language models, particularly LLMs (Sahoo et al., 2024). Understanding how to write a prompt to optimize responses from LLMs is becoming increasing important with the explosion in the popularity of models like OpenAI's GPT series. The aim of this practice is to enhance how LLMs understand and respond to user queries effectively (Liu et al., 2021). By designing prompts that incorporate elements and information typically helpful for humans to structure causal inference queries, we can assess how the inclusion of such information influences LLM performance.

At its core, prompt engineering involves crafting queries that guide the LLM in generating responses that are not only relevant but also contextually rich and nuanced. This is particularly critical in complex problem-solving scenarios where the precision of the input significantly influences the output quality (Shin et al., 2020). An important approach within this field is chain-of-thought (CoT) prompting, which encourages the model to "think aloud" by sequentially reasoning through a problem before arriving at an answer. This method has shown promise in mimicking human-like reasoning patterns, thereby enhancing the model's ability to tackle intricate causal and logical questions (Wei et al., 2023). By systematically structuring prompts to simulate a step-by-step unpacking of complex issues, researchers can potentially unlock different cognitive capabilities in AI, and explore causal cognition in a novel way.

3.1 Chain-of-Thought Prompting

3.1.1 Why Chain-of-Thought (CoT)?

Research has shown that CoT prompting can dramatically improve the performance of LLMs across various domains, including mathematics, common-sense reasoning, and causal inference. Wei et al., (2023) demonstrated that this method enabled LLMs to perform better on arithmetic and common-sense reasoning tasks by breaking down problems into intermediate steps, thus making the models "thought process" visible and logically coherent. Dodge et al., (2019) highlighted how structured prompts can lead to more precise and contextually appropriate outputs in natural language understanding tasks, showing that structured prompting helps mitigate some of the typical failures of LLMs. These failures can include generating plausible but incorrect answers or glossing over complex reasoning steps. By forcing the model to articulate its reasoning, CoT prompting not only improves interpretability but also helps researchers and users to fine-tune the models more effectively. As such, CoT prompting is a tool that can enhance LLM performance, and also sheds light on how AI thinks through complex cognitive tasks.

3.1.2 Standard Chain-of-Thought (CoT) Prompting

Standard CoT prompting opened a frontier in prompt engineering to enhance LLM performance on complex tasks. This method involves crafting prompts that guide the model to articulate its reasoning process step-by-step before arriving at a conclusion. It mirrors human-like problem-solving by having the model "think aloud" as it processes information, providing not just answers to the prompt but also the logical path leading to those answers (Wei et al., 2023). An example is provided in Figure 14. This method is worth investigating because it has shown potential to significantly improve the accuracy of AI responses on reasoning tasks. For instance, when applied to GPT-3, standard CoT prompting uncovered the emergent ability in sufficiently large models to handle nuanced and complex questions more effectively (Wei et al., 2023). CoT prompting can be especially useful in educational settings, where explaining the process of arriving at an answer is as important as the answer itself, and in situations requiring detailed decision-making support, such as complex data analysis tasks. However, there are limitations to consider. For one, this strategy relies heavily on the initial prompt design—poorly designed prompts can lead to misleading or incorrect chains of thought, exacerbating errors rather than clarifying reasoning. While this can be a powerful approach to improving performance on complex tasks in sufficiently large models, the effectiveness of CoT prompting largely depends on the underlying model's training and capacity; it may not yield similar benefits across different models or smaller models (Wei et al., 2023).

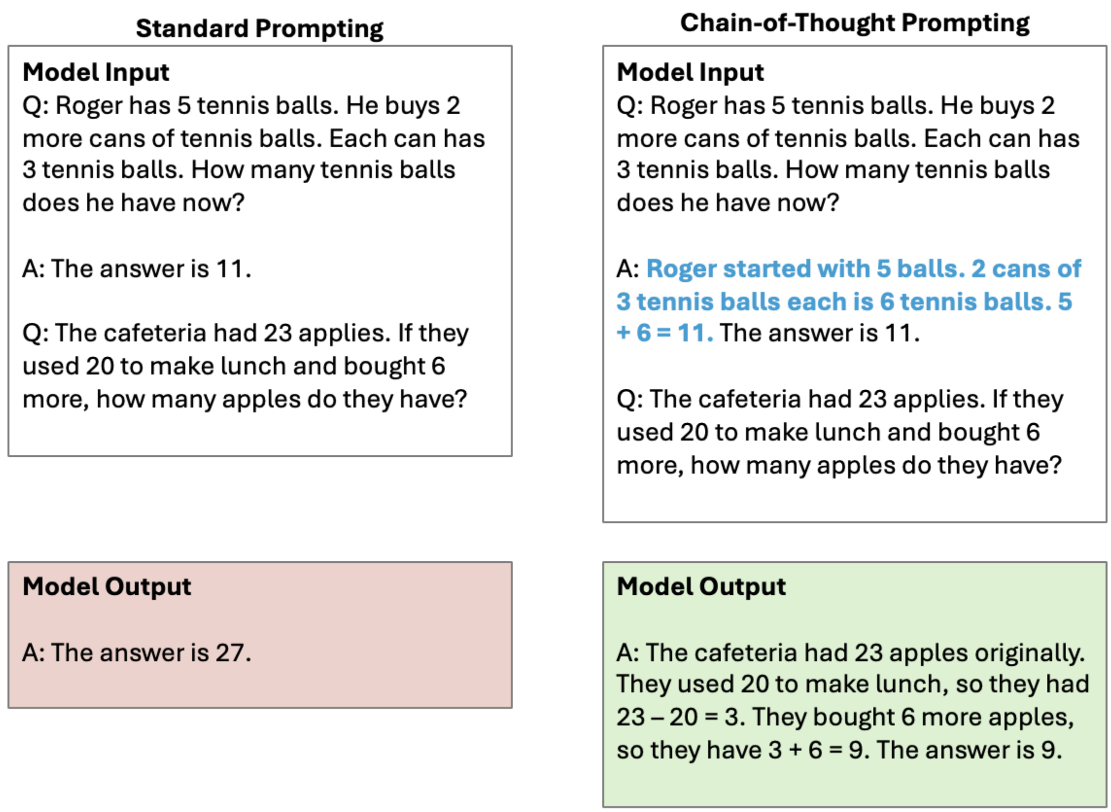


Figure 14. Chain-of-thought prompting compared to standard prompting, adapted from (Wei et al., 2023).

3.1.3 Zero-shot CoT Prompting

Zero-shot CoT prompting is a method where the directive “Let’s think step by step” is added to the prompt, encouraging the model to break down its reasoning into sequential, comprehensible steps (Kojima et al., 2023). This approach is advantageous when there are no specific examples available to provide in the prompt. However, zero-shot prompting may not always lead to correct or a meaningful sequence of thought, particularly if the model has not been trained on related tasks or lacks the necessary knowledge base.

3.1.4 Causal Chain-of-Thought Prompting

The Causal Chain-of-Thought (Causal CoT) prompting strategy is an approach to enhancing causal reasoning capabilities in LLMs. This methodology draws directly from the operational logic of a causal inference engine, which systematically decomposes complex causal reasoning tasks into manageable, symbolically grounded steps (Jin et al., 2024). Causal CoT prompting is designed to simulate the sequence of analytical tasks performed by human experts in causal inference by following the multi-step reasoning process shown in Figure 15. These tasks include identifying the relevant causal graph, specifying the causal query, gathering necessary information, then deriving and evaluating the estimand to come to a formal solution.

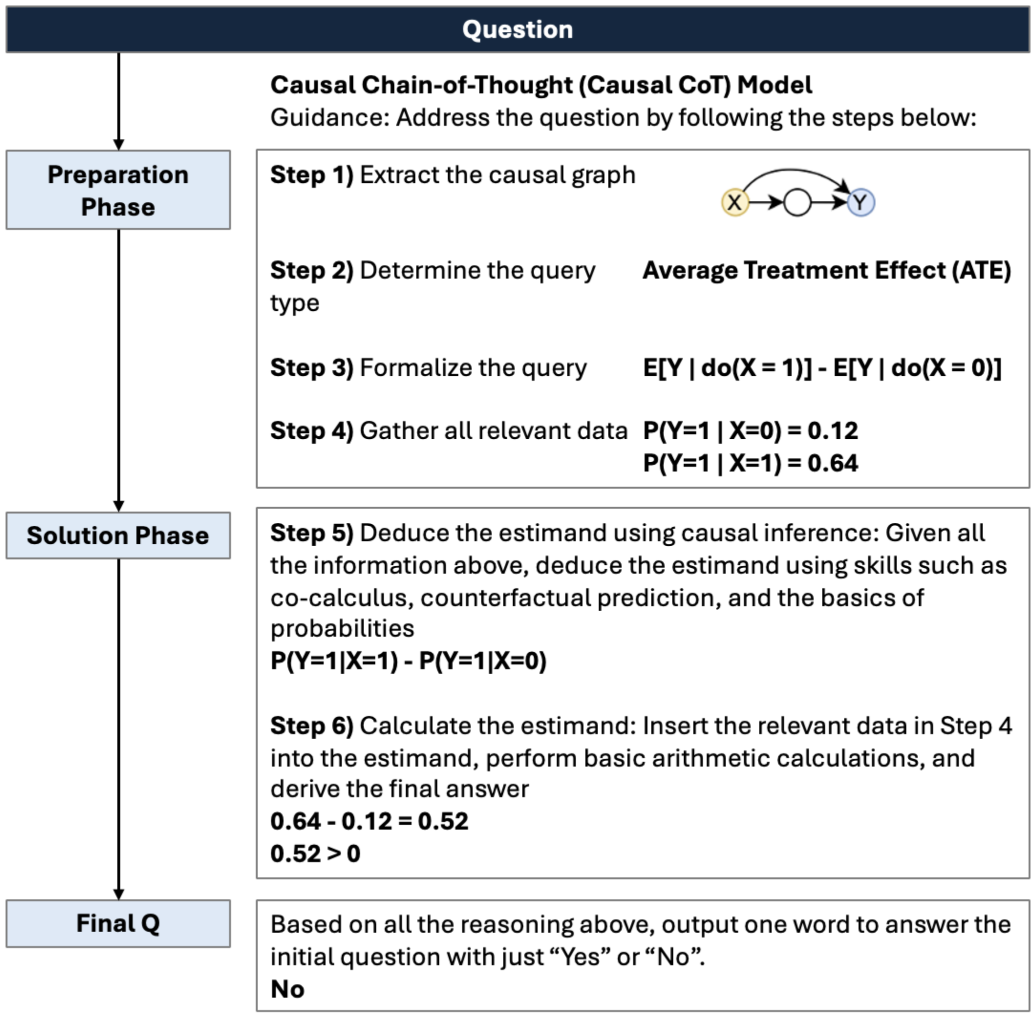


Figure 15. Causal CoT Prompting Strategy, adapted from (Jin et al., 2024).

Using this approach, Jin et al. (2024) observed an improvement to the performance of GPT-4 on the CLADDER dataset, increasing overall accuracy from 62.03% to 70.40% in answering causal reasoning questions. This demonstrates the potential of structured, stepwise prompting strategies to improve reasoning responses of LLMs, particularly in contexts requiring causal analysis. However, the original study did not explore the adaptability of Causal CoT across different LLMs, nor did it evaluate the impact of modifying the information introduced in the prompt. Therefore, this study aims to extend Jin et al.'s methodology by applying Causal CoT across various LLMs and systematically assess the effects of altering the reasoning step information on performance outcomes. This exploration aims to uncover the benefits that LLMs derive from specific steps in the

Causal CoT prompting strategy using an ablation study. As state-of-the-art language models still struggle with multi-step mathematical reasoning (Cobbe et al., 2021), we hypothesize that largest difference in accuracy will be observed with providing arithmetic information (i.e. providing the estimand calculations in steps 4 and 6 in Figure 15).

3.2 Methods

3.2.1 Modified Causal Chain-of-Thought Prompting Strategy

Given its alignment with our research goals, we used the Causal CoT prompting strategy as a foundation, while also incorporating elements from other prompting strategies to give LLMs the best chance to answer accurately based on previous studies on reasoning tasks. In addition to providing step-by-step instructions inspired by the causal inference engine, we also provided the correct answer for each step up to the final solution for each question. The aim of this modification was to mirror the effects of standard CoT by providing a correct example of the reasoning pathway (Wei et al., 2023). Additionally, "Let's think step by step" was appended to the question to bring in the effects of zero-shot CoT prompting (Kojima et al., 2023). To show the components of the prompt, an example prompt (Appendix 5) used in these experiments is as follows:

1. **Background**: "Imagine a self-contained, hypothetical world with only the following conditions, and without any unmentioned factors or causal relationships. Encouragement level has a direct effect on studying habit and exam score. Studying habit has a direct effect on exam score."
2. **Given information about the question**: "For students who are not encouraged, the probability of high exam score is 12%. For students who are encouraged, the probability of high exam score is 64%."
3. **Question**: "Will encouragement decrease the chance of a high exam score? Let's think step by step. First, let's define the variables: Let X = encouragement level; V2 = studying habit; Y = exam score."
4. **Causal Reasoning Steps:**

"**Step 1)** Extract the causal graph: Identify the causal graph that depicts the relationships in the scenario. The diagram should simply consist of edges denoted in 'var1 → var2' format, separated by commas.

Answer: X → Y2, X → Y, V2 → Y

Step 2) Determine the query type: Identify the type of query implied by the main question. Please answer only with one of the choice options: correlation, marginal distribution, explaining away effect (exp_away), average treatment effect (ate), backdoor adjustment set (backadj), collider_bias, effect of the treatment on the treated (ett), natural direct effect (nde), natural indirect effect (nie), and counterfactual deterministic.

Answer: ate

Step 3) Formalize the query: Translate the query into its formal mathematical expression based on its type, utilizing the 'do(.)' notation or counterfactual notations as needed.

Answer: E[Y | do(X = 1)] - E[Y | do(X = 0)]

Step 4) Gather all relevant data: Extract all the available data. Your answer should contain nothing but marginal probabilities and conditional probabilities in the form 'P(...)=...' or 'P(...|...)=...', each probability being separated by a semicolon. Answer: P(Y=1|X=1) - P(Y=1|X=0)

Step 5) Deduce the estimand: Given all the information above, deduce the estimand using skills such as do-calculus, counterfactual prediction, and the basics of probabilities. Answer step by step.

Answer: P(Y=1 | X=0) = 0.12 P(Y=1 | X=1) = 0.64

Step 6) Calculate the estimand: Insert the relevant data in Step 4 into the estimand, perform basic arithmetic calculations, and derive the final answer. Answer step by step.

Answer: 0.64 - 0.12 = 0.52

Step 7) Finally, based on all the reasoning above, output your answer the initial question starting with 'Yes' or 'No'."

3.2.2 Step Ablations

For each of the models tested (GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo), the filtered CLADDER dataset of 1392 questions was used to prompt the LLMs with 4 different configurations: full steps (as mentioned above), without steps 1 and 2 (information about question type), without steps 3 and 5 (organizing information), and without steps 4 and 6 (arithmetic information). Claude 3 Opus and other Anthropic models were excluded from this study due to constraints on computing resources. The limitations on Anthropic API usage and higher cost of inference prevented the repetition of these experiments for Claude 3 Opus.

3.2.3 Overall Study Design

Similar to the simple prompting approach, these components (including the appropriate steps) were read from the filtered dataset (stored in a JSON file) and concatenated, then passed through the OpenAI API Server with temperature 0 to prompt GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo using the filtered dataset of 1392 questions. Again, to keep things consistent with the simple prompting approach, the instruction from the ‘system’ role guided the LLM to respond with the first word being “Yes” or “No” using the following system message: “You are an expert in causal inference. The following question is not a typical commonsense query, but rather a meticulously designed question created by a professor specializing in causal inference, intended to assess mastery of the course content. Start your answer with “yes” or “no,” followed by additional reasoning or evidence to support your explanation.” For the scoring, again the first word of the LLM response was taken and compared to the ground truth (not case sensitive), and the same true or false scorer was used to compare the prediction with the ground truth.

3.3 Results

The response accuracy on the 1392 causal inference questions for each of the configurations is shown in Figure 16. Overall, providing the full set of causal inference steps in the prompt resulted in a decrease in accuracy compared to the baseline simple prompting approach that was used in the previous chapter. All configurations of the step ablations also resulted in varying levels of decreased accuracy compared to baseline across all models, with the exception of ablating steps 4 and 6 with GPT-4 Turbo.

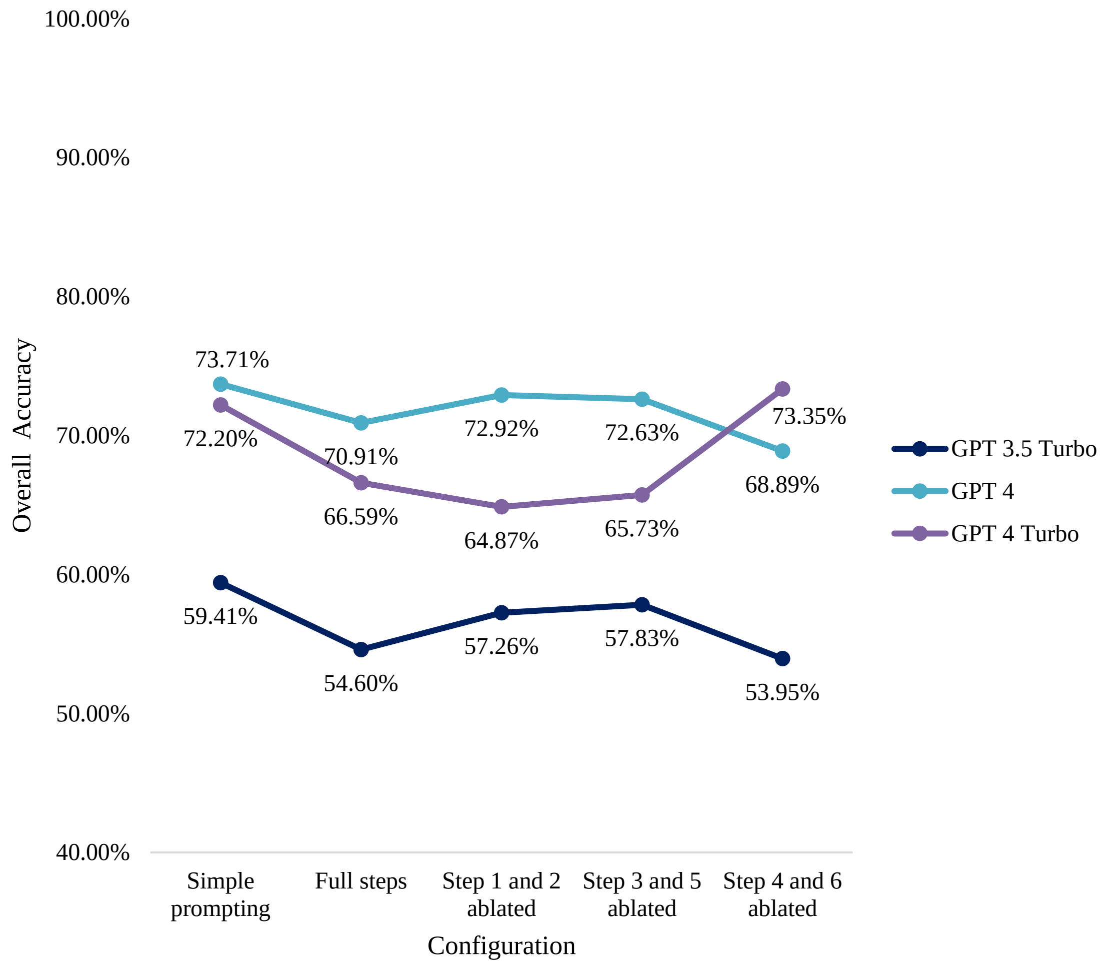


Figure 16. Overall Accuracies for GPT 3.5 turbo, GPT 4, and GPT 4 turbo on the Chain-of-Thought prompting task.

To look with higher granularity at whether the causal inference steps resulted in this decrease in accuracy across different rungs and query types, the change in accuracy was calculated between the baseline (simple prompting) and each of the configurations for every model. Table 3 shows this data broken down by rung and query type, and Table 4 shows this broken down by graph type.

Table 3: Change in accuracy relative to baseline with the configurations of the Chain-of-Thought ablation study, broken down by rung and query type.

Query Type (Rung Type)	Model	Simple prompting	Full steps	Step 1 and 2 ablated	Step 3 and 5 ablated	Step 4 and 6 ablated
Correlation (Rung 1)	GPT 3.5 Turbo	0.00%	-4.39%	-6.42%	10.48%	-5.74%
	GPT 4	0.00%	5.41%	13.51%	14.19%	-11.82%
	GPT 4 Turbo	0.00%	7.09%	4.05%	12.16%	2.70%
Explaining Away Effect (Rung 1)	GPT 3.5 Turbo	0.00%	0.00%	0.00%	0.00%	-3.12%
	GPT 4	0.00%	25.00%	28.13%	12.50%	6.25%
	GPT 4 Turbo	0.00%	-31.25%	-31.25%	-34.38%	-34.38%
Marginal Distribution (Rung 1)	GPT 3.5 Turbo	0.00%	1.02%	5.41%	-5.74%	1.69%
	GPT 4	0.00%	-3.72%	-3.38%	-6.08%	3.72%
	GPT 4 Turbo	0.00%	-4.05%	-6.42%	-17.23%	19.93%
Average Treatment Effect (Rung 2)	GPT 3.5 Turbo	0.00%	-21.21%	-7.57%	-12.12%	-23.86%
	GPT 4	0.00%	0.00%	0.38%	-0.76%	-1.52%
	GPT 4 Turbo	0.00%	-5.68%	-7.20%	-4.92%	-0.76%
Collider Bias (Rung 2)	GPT 3.5 Turbo	0.00%	12.50%	3.13%	3.13%	-6.25%
	GPT 4	0.00%	18.75%	21.88%	0.00%	-6.25%
	GPT 4 Turbo	0.00%	6.25%	-18.75%	0.00%	-21.88%
Effect of the Treatment on the Treated (Rung 3)	GPT 3.5 Turbo	0.00%	-2.23%	-2.23%	-4.02%	2.68%
	GPT 4	0.00%	-20.54%	-25.00%	-14.73%	-11.16%
	GPT 4 Turbo	0.00%	-22.77%	-20.98%	-19.64%	-19.64%
Natural Direct Effect (Rung 3)	GPT 3.5 Turbo	0.00%	-3.13%	-2.08%	4.17%	-3.13%
	GPT 4	0.00%	-6.25%	-3.13%	-4.17%	-9.38%
	GPT 4 Turbo	0.00%	-5.21%	-9.38%	-3.13%	7.29%
Natural Indirect Effect (Rung 3)	GPT 3.5 Turbo	0.00%	1.98%	-0.65%	0.00%	-0.65%
	GPT 4	0.00%	-3.95%	0.66%	-2.63%	-3.29%
	GPT 4 Turbo	0.00%	-5.26%	-2.63%	-2.63%	3.95%

Table 4: Change in accuracy relative to baseline with the configurations of the Chain-of-Thought ablation study, broken down by graph type.

Graph Type	Model	Simple prompting	Full steps	Step 1 and 2 ablated	Step 3 and 5 ablated	Step 4 and 6 ablated
Arrowhead	GPT 3.5 Turbo	0.00%	-3.13%	-3.47%	-1.04%	-7.64%
	GPT 4	0.00%	-10.42%	-7.64%	-4.51%	-9.03%
	GPT 4 Turbo	0.00%	-9.03%	-8.33%	-9.38%	1.04%
Chain	GPT 3.5 Turbo	0.00%	-11.66%	-8.33%	-7.50%	-13.33%
	GPT 4	0.00%	1.67%	5.00%	5.83%	-1.67%
	GPT 4 Turbo	0.00%	-1.67%	0.00%	0.00%	5.00%
Collision	GPT 3.5 Turbo	0.00%	3.90%	2.34%	-5.47%	-0.78%
	GPT 4	0.00%	11.72%	16.41%	7.03%	-5.47%
	GPT 4 Turbo	0.00%	3.91%	-6.25%	-3.91%	-16.41%
Confounding	GPT 3.5 Turbo	0.00%	-7.81%	-3.13%	-1.56%	-1.56%
	GPT 4	0.00%	0.78%	1.56%	-1.56%	2.34%
	GPT 4 Turbo	0.00%	-15.63%	-14.84%	-10.16%	-1.56%
Diamond	GPT 3.5 Turbo	0.00%	3.33%	5.83%	0.83%	3.33%
	GPT 4	0.00%	-7.50%	-5.83%	-5.00%	-8.33%
	GPT 4 Turbo	0.00%	-1.67%	0.00%	-1.67%	5.00%
Diamond Cut	GPT 3.5 Turbo	0.00%	-12.50%	0.00%	9.38%	-6.25%
	GPT 4	0.00%	-9.38%	-6.25%	-9.38%	-6.25%
	GPT 4 Turbo	0.00%	6.25%	15.63%	12.50%	9.38%
Fork	GPT 3.5 Turbo	0.00%	-8.60%	-3.91%	-1.56%	-5.47%
	GPT 4	0.00%	-8.59%	-6.25%	-3.91%	-3.13%
	GPT 4 Turbo	0.00%	-0.78%	-4.69%	-3.13%	3.13%
Front Door	GPT 3.5 Turbo	0.00%	0.00%	-2.50%	-10.00%	-5.00%
	GPT 4	0.00%	-2.50%	-2.50%	-5.00%	-7.50%
	GPT 4 Turbo	0.00%	-12.50%	-27.50%	-10.00%	-7.50%
Instrumental Variable	GPT 3.5 Turbo	0.00%	-5.84%	1.66%	4.16%	-5.00%
	GPT 4	0.00%	17.50%	15.83%	13.33%	4.17%
	GPT 4 Turbo	0.00%	-8.33%	-13.33%	-16.67%	2.50%
Mediation	GPT 3.5 Turbo	0.00%	-7.29%	-4.16%	-1.39%	-7.64%
	GPT 4	0.00%	-8.33%	-6.60%	-5.56%	-7.29%
	GPT 4 Turbo	0.00%	-6.60%	-7.99%	-6.60%	5.90%

3.4 Discussion

3.4.1 Findings Contradict with Previous Work on Causal CoT Prompting

Our study found that simple prompting, which provides straightforward background information and questions, resulted in higher accuracy compared to the modified Causal CoT prompting approach (Figure 16). This finding was contrary to the findings by Jin et al. (2024) in their study on Causal CoT and the CLADDER dataset. Jin et al. reported an increase in accuracy from 62.03% with GPT-4 (without Causal CoT) to 70.40% accuracy with the GPT-4 with Causal CoT, and they observed this increase in accuracy across all three rungs of causal questions. In our study, this was not observed—providing the full steps configuration only increased accuracy with specific rung 1 (correlation, explaining away effect) and rung 2 (collider bias) questions, and even then, inconsistently in some but not all models (Table 3). Rather, performance accuracy on most questions—particularly rung 3 questions—decreased compared to the simple prompting approach. Similarly, when we look at questions broken down by graph type, there was an inconsistent change in accuracy performance across graph types compared to simple prompting (Table 4). In terms of overall accuracy, by including the full causal reasoning steps in the prompt, the models performed with a decrease rather than the expected increase in accuracy.

These findings from this study suggest that future work on validating chain-of-thought prompting strategies is warranted. **Our results indicate that LLMs do not benefit from CoT techniques when assessed with formal causal reasoning tasks, even though the models tested were sufficiently large (Wei et al., 2023), suggesting that formal causal reasoning may require other approaches to optimize performance.** There remains much to be uncovered for future studies investigating context-specific optimization of causal reasoning in AI.

3.4.2 Large Language Models Approach Formal Causal Reasoning Differently Compared to Humans

Our findings reveal that structural frameworks suitable for human causal reasoning might not be a suitable framework for LLMs to reason about causality. The development of causal inference as a discipline has evolved over many decades through contributions from multiple fields, culminating in frameworks like the causal inference engine, which logically structures reasoning in a manner intuitive to human thinkers (Pearl & Mackenzie, 2018). However, our study indicates that providing LLMs with structured logical prompts that mirror human approaches to formal causal reasoning using the causal inference engine does not enhance LLM performance. We observed that simple prompting consistently yielded higher accuracy compared to the configurations that provided additional information to solve the formal causal reasoning task (Figure 15), which was unexpected. This discrepancy raises questions about the alignment of LLM capabilities with human-like reasoning structures and suggests that the architecture of these models or their training data may lack crucial elements necessary for understanding, learning, or applying causal reasoning step-by-step (Brown et al., 2020). Potentially, the additional cognitive load imposed by CoT prompting, intended to emulate human reasoning processes in a causal inference task, may not align with the intrinsic information processing capabilities of LLMs. In a broader sense, these results contribute to the present discourse on LLM consciousness by providing evidence that LLMs do not 'think' in a human-like manner.

While it was hypothesized that providing arithmetic information would be particularly beneficial, this was not supported by our results, thus opening new avenues for exploring how AI can effectively engage in causal reasoning. This investigation into the 'artificial mind' reveals that unlike human minds, the steps required for solving complex cognitive tasks such as causal reasoning in LLMs may require a re-evaluation of what effectively aids these models, inviting further research into optimizing AI for complex, causal problem-solving.

3.4.3 Future Work and Limitations

One potentially promising approach is adapting a new technique used to improve mathematical reasoning performance for LLMs (Chen et al., 2024). This novel approach named 'AlphaMath' leverages Monte Carlo Tree Search (MCTS) to enhance mathematical reasoning without the need for human-annotated training data. MCTS is a search algorithm used for making optimal decisions, notably in game play and other areas where predicting the outcome is important (Chaslot et al., 2021). Chen et al. (2024) show that by using a well-pre-trained LLM and leveraging MCTS, LLMs can identify the correct mathematical reasoning process without human annotation. The integration of LLMs with MCTS is a technique previously used in strategic games like AlphaGo (Silver et al., 2017) to optimize the model's problem-solving abilities in mathematics. The arithmetic reasoning steps involved in formal causal reasoning could be enhanced by this approach, and it would also eliminate the costly need for human-annotated step-by-step solutions for causal inference.

Additionally, the method presented by Chen et al. (2024) allows the model to learn from both correct and incorrect solution paths. AlphaMath operates by iteratively generating and refining solution paths, using a policy model to predict potential steps and a value model to evaluate these steps' correctness. This aligns closely with the approach that was taken in Chapter 3 of this thesis with the modified Causal Chain-of-Thought prompting strategy. Future studies comparing this approach with the integration of MCTS for causal inference steps can provide valuable insight into optimizing these models for causal reasoning, as well as further contribute to cognitive research by exploring this as a different 'thinking strategy'. Overall, this future direction highlights the potential of MCTS integrated with LLMs to autonomously refine and enhance AI's reasoning capabilities, particularly in domains requiring complex, multi-step reasoning like in causal reasoning.

Despite their impressive capabilities, the extent to which LLMs truly 'understand' language remains debatable. As with other neural networks, their operation primarily involves pattern recognition—LLMs are fundamentally designed to predict the next word in a sequence based on patterns observed in their training data. This approach can yield responses that are superficially plausible yet lack deep semantic coherence, questioning the depth of their linguistic comprehension (E. M. Bender & Koller, 2020). This limitation is relevant to our study, which used prompts drawn from sources like statistics textbooks and websites—content likely included in the training data of LLMs' responses might simply reflect correct answers already present in their training datasets. Therefore, it is possible that LLM responses might merely reflect correct answers already present in their training data. The CLADDER paper by Jin et al. (2024) addresses this limitation by comparing LLM responses to standard questions with responses to modifies questions containing nonsensical variables. The study observed no consistent improvement or decline in LLM performance with these altered prompts, suggesting that modifying prompts to exclude training data does not impact LLM performance on causal reasoning tasks. Nonetheless, it is continually important in AI and LLM research to consider the influence of training data of responses. Moreover, our study adds to the ongoing debate on whether mere data sufficiency drives LLM reasoning. While adaptations to the causal Chain-of-Thought prompting did not uniformly improve LLM causal reasoning performance, exploring alternative strategies to improve reasoning such as incorporating MCTS (Chen et al., 2024) could potentially expand LLM cognitive capabilities.

A notable limitation of using LLMs for causal reasoning tasks is the lack of direct comparability to human cognitive processes. Particularly in this study, LLMs operated within a binary or limited output framework, producing "Yes" or "No" responses. This stands in contrast to human reasoning, which is nuanced and can integrate a broad spectrum of responses based on context, emotion, and cultural norms (Hagendorff et al., 2023). To address these discrepancies, future studies might expand upon this exploration through the integration of more flexible response mechanisms in LLMs that mirror the complexity of human thought processes. Furthermore, the causal reasoning methodologies used in these studies were reliant on directed acyclic graphs, which present another misalignment with human cognition. Human reasoning is thought to be recurrent, where previous knowledge and experiences continually inform and reshape our understanding of cognitive processes. This cyclic nature of human thought allows for the adaptation and evolution of ideas, a feature that is structurally unsupported in the acyclic computational frameworks used in this study. This structural limitation not only affects LLMs' ability to simulate human-like reasoning but also its effectiveness in tasks requiring iterative and evolving thought processes, such as complex problem-solving and learning. Lastly, the exploration of these themes inevitably raises deeper questions about the nature of reasoning and cognition beyond the confines of language. The current focus on verbal or textual outputs from models like LLMs may overlook the vast spectrum of human cognitive activities that are non-verbal, such as spatial awareness, emotional intelligence, and unconscious processing (Taylor, 2001). Rethinking the frameworks and outputs of these models to accommodate these forms of cognition could provide valuable insight to enhance their utility and performance. Incorporating multimodal learning systems that integrate visual, textual, and sensory data might provide a more holistic approach to AI development (Bewersdorff et al., 2024). These limitations underscore the importance of interdisciplinary AI research, particularly in towards how we frame and interpret the capabilities of LLMs within the context of cognitive neuroscience.

The opaque nature of complex models like LLMs presents significant challenges, often referred to as the “black box” problem. This problem describes how the internal workings of these models are not easily discernible, making it difficult to understand how they process information and arrive at conclusions (Linardatos et al., 2020). This opacity complicates their use in cognitive science, where understanding the mechanisms of processing is an important area of investigation. Furthermore, the potential biases embedded within the training data of these models raise ethical and methodological concerns, especially when these systems are applied in contexts that simulate human decision-making processes. These challenges and concerns should continue to be addressed while we progress in cognitive neuroscience research towards understanding AI cognition and human cognitive processes. Continued research at the intersection of AI and cognitive science is essential to not only improve model transparency, but also delineate the similarities and differences between AI and human cognition.

References

Bender, A. (2020). What Is Causal Cognition? *Frontiers in Psychology*, *11*, 3. https://doi.org/10.3389/fpsyg.2020.00003

Bender, A., & Beller, S. (2019). The Cultural Fabric of Human Causal Cognition. *Perspectives on Psychological Science*, *14*(6), 922–940. https://doi.org/10.1177/1745691619863055

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., & Pal, C. (2019). *A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms* (arXiv:1901.10912). arXiv. http://arxiv.org/abs/1901.10912

Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., Kasneci, G., Zhai, X., & Nerdel, C. (2024). *Taking the Next Step with Generative Artificial Intelligence: The Transformative Role of Multimodal Large Language Models in Science Education* (arXiv:2401.00832). arXiv. http://arxiv.org/abs/2401.00832

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal Reasoning in Rats. *Science*, *311*(5763), 1020–1022. https://doi.org/10.1126/science.1121872

Bringsjord, S., & Bringsjord, A. (2012). Kluge: The Haphazard Construction of the Human Mind. *Philosophical Psychology*, *25*(2), 301–305. https://doi.org/10.1080/09515089.2011.583015

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. http://arxiv.org/abs/2005.14165

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (arXiv:2308.08708). arXiv. http://arxiv.org/abs/2308.08708

Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. (pp. xiv, 254). MIT Press.

Cai, H., Liu, S., & Song, R. (2023). *Is Knowledge All Large Language Models Needed for Causal Reasoning?* (arXiv:2401.00139). arXiv. http://arxiv.org/abs/2401.00139

Chaslot, G., Bakkes, S., Szita, I., & Spronck, P. (2021). Monte-Carlo Tree Search: A New Framework for Game AI. *Proceedings of the AAAI Conference on Artificial*

Intelligence and Interactive Digital Entertainment, *4*(1), 216–217. https://doi.org/10.1609/aiide.v4i1.18700

Chen, G., Liao, M., Li, C., & Fan, K. (2024). *AlphaMath Almost Zero: Process Supervision without process* (arXiv:2405.03553). arXiv. http://arxiv.org/abs/2405.03553

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). *Training Verifiers to Solve Math Word Problems* (arXiv:2110.14168). arXiv. http://arxiv.org/abs/2110.14168

Cruz, N., Hahn, U., Fenton, N., & Lagnado, D. (2020). Explaining Away, Augmentation, and the Assumption of Independence. *Frontiers in Psychology*, *11*, 502751. https://doi.org/10.3389/fpsyg.2020.502751

Darema, F. (2004). Dynamic Data Driven Applications Systems: A New Paradigm for Application Simulations and Measurements. In M. Bubak, G. D. Van Albada, P. M. A. Sloot, & J. Dongarra (Eds.), *Computational Science—ICCS 2004* (Vol. 3038, pp. 662–669). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24688-6_86

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, *56*(2), 384–398. https://doi.org/10.1016/j.neuron.2007.10.004

Dehaene, S., Dupoux, E., Mehler, J., Cohen, L., Paulesu, E., Perani, D., van de Moortele, P. F., Lehéricy, S., & Le Bihan, D. (1997). Anatomical variability in the cortical representation of first and second language. *Neuroreport*, *8*(17), 3809–3815. https://doi.org/10.1097/00001756-199712010-00030

Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show Your Work: Improved Reporting of Experimental Results. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2185–2194. https://doi.org/10.18653/v1/D19-1224

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., … Olah, C. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.

Evans, R., & Grefenstette, E. (2018). *Learning Explanatory Rules from Noisy Data* (arXiv:1711.04574). arXiv. http://arxiv.org/abs/1711.04574

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202. https://doi.org/10.1007/BF00344251

Giere, R. N., & Salmon, W. C. (1988). Scientific Explanation and the Causal Structure of the World. *The Philosophical Review*, *97*(3), 444. https://doi.org/10.2307/2185459

Goldberg, Y. (2019). *Assessing BERT's Syntactic Abilities* (arXiv:1901.05287). arXiv. http://arxiv.org/abs/1901.05287

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT press.

Gopnik, A., & Schulz, L. (Eds.). (2007). *Causal Learning: Psychology, Philosophy, and Computation* (1st ed.). Oxford University PressNew York. https://doi.org/10.1093/acprof:oso/9780195176803.001.0001

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, *3*(10), 833–838. https://doi.org/10.1038/s43588-023-00527-x

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). Measuring Mathematical Problem Solving With the MATH Dataset. In J. Vanschoren & S. Yeung (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (Vol. 1). https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf

Heneka, M. T., Kummer, M. P., & Latz, E. (2014). Innate immune activation in neurodegenerative disease. *Nature Reviews Immunology*, *14*(7), 463–477. https://doi.org/10.1038/nri3705

Hinton (Director). (2023, May 9). *"Godfather of AI" Geoffrey Hinton Warns of the "Existential Threat" of AI | Amanpour and Company* [Video recording]. https://www.youtube.com/watch?v=Y6Sgp7y178k

Ho, M., Sharma, A., Chang, J., Saxon, M., Levy, S., Lu, Y., & Wang, W. Y. (2022). *WikiWhy: Answering and Explaining Cause-and-Effect Questions* (arXiv:2210.12152). arXiv. http://arxiv.org/abs/2210.12152

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holmberg, M. J., & Andersen, L. W. (2022). Collider Bias. *JAMA*, *327*(13), 1282. https://doi.org/10.1001/jama.2022.1820

Hume, D. (1779). An enquiry concerning human understanding. In D. Hume, *Essays and treatises on several subjects, Vol 2: Containing An enquiry concerning human understanding, A dissertation on the passions, An enquiry concerning the principles of morals, and The natural history of religion.* (pp. 3–212). Unknown Publisher. https://doi.org/10.1037/11713-001

Jefferies, M. E., & Yeap, W.-K. (2008). *Robotics and cognitive approaches to spatial mapping*. Springer.

Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2020). *How Can We Know What Language Models Know?* (arXiv:1911.12543). arXiv. http://arxiv.org/abs/1911.12543

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Adauto, F. G., Kleiman-Weiner, M., Sachan, M., & Schölkopf, B. (2024). *CLadder: Assessing Causal Reasoning in Language Models* (arXiv:2312.04350). arXiv. http://arxiv.org/abs/2312.04350

Johnson-Laird, P. N. (1995). *Mental models: Towards a cognitive science of language, inference, and consciousness* (6. print). Harvard Univ. Press.

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed). Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1982). The Psychology of Preferences. *Scientific American*, *246*(1), 160–173. https://doi.org/10.1038/scientificamerican0182-160

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, *109*(2), 224–234. https://doi.org/10.1016/j.cognition.2008.08.010

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). *Causal Reasoning and Large Language Models: Opening a New Frontier for Causality* (arXiv:2305.00050). arXiv. http://arxiv.org/abs/2305.00050

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). *Large Language Models are Zero-Shot Reasoners* (arXiv:2205.11916). arXiv. http://arxiv.org/abs/2205.11916

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Kummer, H. (1996). Causal knowledge in animals. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal Cognition* (pp. 26–36). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198524021.003.0002

Kummer, H., & Goodall, J. (2003). Conditions of Innovative Behaviour in Primates. In S. M. Reader & K. N. Laland (Eds.), *Animal Innovation* (pp. 223–236). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198526223.003.0010

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Li, J., Zaikin, A., Zhang, X., & Chen, S. (2022). Editorial: Closed-loop iterations between neuroscience and artificial intelligence. *Frontiers in Systems Neuroscience*, *16*, 1002095. https://doi.org/10.3389/fnsys.2022.1002095

Lin, Z., Feng, M., Santos, C. N. dos, Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). *A Structured Self-attentive Sentence Embedding* (arXiv:1703.03130). arXiv. http://arxiv.org/abs/1703.03130

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy (Basel, Switzerland)*, *23*(1), 18. https://doi.org/10.3390/e23010018

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing* (arXiv:2107.13586). arXiv. http://arxiv.org/abs/2107.13586

Lombard, M., & Gärdenfors, P. (2017). Tracking the evolution of causal cognition in humans. *Journal of Anthropological Sciences*, *95*, 219–234. https://doi.org/10.4436/JASS.95006

López Espejel, J., Ettifouri, E. H., Yahaya Alassan, M. S., Chouham, E. M., & Dahhane, W. (2023). GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, *5*, 100032. https://doi.org/10.1016/j.nlp.2023.100032

Lupien, S. J., Maheu, F., Tu, M., Fiocco, A., & Schramek, T. E. (2007). The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and Cognition*, *65*(3), 209–237. https://doi.org/10.1016/j.bandc.2007.02.007

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust* (First edition). Pantheon Books.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). *Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference* (arXiv:1902.01007). arXiv. http://arxiv.org/abs/1902.01007

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435. https://doi.org/10.2307/2183914

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). *A Comprehensive Overview of Large Language Models* (arXiv:2307.06435). arXiv. http://arxiv.org/abs/2307.06435

Newell, A., & Simon, H. A. (1972). *Human problem solving.* (pp. xiv, 920). Prentice-Hall.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex.* (pp. xv, 430). Oxford Univ. Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* (pp. xvi, 384). Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Pennartz, C. M. A., Farisco, M., & Evers, K. (2019). Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach. *Frontiers in Systems Neuroscience*, *13*, 25. https://doi.org/10.3389/fnsys.2019.00025

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language Models as Knowledge Bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. https://doi.org/10.18653/v1/D19-1250

Pinker, S. (1999). *How the mind works* (1. publ). Norton.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Russell, S. J., & Norvig, P. (with Davis, E., & Edwards, D.). (2016). *Artificial intelligence: A modern approach* (Third edition, Global edition). Pearson.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications* (arXiv:2402.07927). arXiv. http://arxiv.org/abs/2402.07927

Scassellati, B., Admoni, H., & Matarić, M. (2012). Robots for use in autism research. *Annual Review of Biomedical Engineering*, *14*, 275–294. https://doi.org/10.1146/annurev-bioeng-071811-150036

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). *Towards Causal Representation Learning* (arXiv:2102.11107). arXiv. http://arxiv.org/abs/2102.11107

Schwab, P., Linhardt, L., & Karlen, W. (2019). *Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks* (arXiv:1810.00656). arXiv. http://arxiv.org/abs/1810.00656

Shettleworth, S. J. (2010). *Cognition, evolution, and behavior, 2nd ed.* (pp. xiii, 700). Oxford University Press.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). *AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts* (arXiv:2010.15980). arXiv. http://arxiv.org/abs/2010.15980

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Singh, M., Cambronero, J., Gulwani, S., Le, V., & Verbruggen, G. (2023). *Assessing GPT4-V on Structured Reasoning Tasks* (arXiv:2312.11524). arXiv. http://arxiv.org/abs/2312.11524

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis.* (p. 457). Appleton-Century.

Sloman, S. (2005). *Causal Models: How People Think about the World and Its Alternatives* (1st ed.). Oxford University PressNew York. https://doi.org/10.1093/acprof:oso/9780195183115.001.0001

Spelke, E. S. (2000). Core knowledge. *American Psychologist*, *55*(11), 1233–1243. https://doi.org/10.1037/0003-066X.55.11.1233

Sternberg, R. J. (Ed.). (1999). *The Nature of Cognition*. The MIT Press. https://doi.org/10.7551/mitpress/4877.001.0001

Stolfo, A., Jin, Z., Shridhar, K., Schölkopf, B., & Sachan, M. (2023). *A Causal Framework to Quantify the Robustness of Mathematical Reasoning with Language Models* (arXiv:2210.12023). arXiv. http://arxiv.org/abs/2210.12023

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction, 2nd ed.* (pp. xxii, 526). The MIT Press.

Taylor, E. W. (2001). Transformative learning theory: A neurobiological perspective of the role of emotions and unconscious ways of knowing. *International Journal of Lifelong Education*, *20*(3), 218–236. https://doi.org/10.1080/02601370110036064

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In *The adapted mind: Evolutionary psychology and the generation of culture.* (pp. 19–136). Oxford University Press.

VanderWeele, T. J. (2011). Controlled direct and mediated effects: Definition, identification and bounds. *Scandinavian Journal of Statistics, Theory and Applications*, *38*(3), 551–563. https://doi.org/10.1111/j.1467-9469.2010.00722.x

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Warstadt, A., Singh, A., & Bowman, S. R. (2019). *Neural Network Acceptability Judgments* (arXiv:1805.12471). arXiv. http://arxiv.org/abs/1805.12471

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (arXiv:2201.11903). arXiv. http://arxiv.org/abs/2201.11903

Weisberg, D. S., & Gopnik, A. (2013). Pretense, Counterfactuals, and Bayesian Causal Models: Why What Is Not Real Really Matters. *Cognitive Science*, *37*(7), 1368–1381. https://doi.org/10.1111/cogs.12069

Wright, S. (1920). The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs. *Proceedings of the National Academy of Sciences*, *6*(6), 320–332. https://doi.org/10.1073/pnas.6.6.320

Wright, G. H. von. (2004). *Explanation and understanding*. Cornell University Press.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yehuda, R., & LeDoux, J. (2007). Response Variation following Trauma: A Translational Neuroscience Approach to Understanding PTSD. *Neuron*, *56*(1), 19–32. https://doi.org/10.1016/j.neuron.2007.09.006

Zečević, M., Willig, M., Dhami, D. S., & Kersting, K. (2023). *Causal Parrots: Large Language Models May Talk Causality But Are Not Causal* (arXiv:2308.13067). arXiv. http://arxiv.org/abs/2308.13067

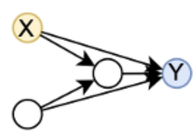
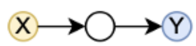
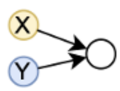
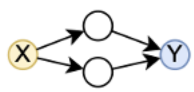
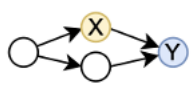
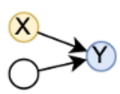
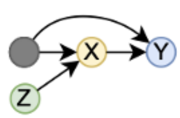
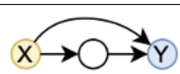
Zhang, C., Bauer, S., Bennett, P., Gao, J., Gong, W., Hilmkil, A., Jennings, J., Ma, C., Minka, T., Pawlowski, N., & Vaughan, J. (2023). *Understanding Causality with Large Language Models: Feasibility and Opportunities* (arXiv:2304.05524). arXiv. http://arxiv.org/abs/2304.05524

Zheng, H., Shen, L., Tang, A., Luo, Y., Hu, H., Du, B., & Tao, D. (2023). *Learn From Model Beyond Fine-Tuning: A Survey* (arXiv:2310.08184). arXiv. http://arxiv.org/abs/2310.08184

Zhou, A., Wang, K., Lu, Z., Shi, W., Luo, S., Qin, Z., Lu, S., Jia, A., Song, L., Zhan, M., & Li, H. (2023). *Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-based Self-Verification* (arXiv:2308.07921). arXiv. http://arxiv.org/abs/2308.07921

Appendices

Appendix 1: Number of Questions for Each Graph Type in the CLADDER dataset

Graph Type	Graph Structure*	Number of Questions Using this Graph	Percentage of Total Dataset (10,112 Questions)
Arrowhead		1264	12.5%
Chain		1106	10.9%
Collision		632	6.2%
Confounding		948	9.4%
Diamond		1106	10.9%
Diamond Cut		948	9.4%
Fork		948	9.4%
Front Door		1106	10.9%
Instrumental Variable		790	7.8%
Mediation		1264	12.5%

*X represents the treatment variable, Y represents the outcome variable, and Z represents the mediator variable

Appendix 1: Number of Questions for Each Query Type in the CLADDER dataset

Query Type	Rung	Number of Questions Using this Query Type	Percentage of Total Dataset (10,112 Questions)
Correlation	1	1422	14.1%
Explaining Away Effect	1	158	1.6%
Marginal Distribution	1	1580	15.6%
Average Treatment Effect	2	1422	14.1%
Backdoor Adjustment Set	2	1580	15.6%
Collider Bias	2	158	1.6%
Effect of the Treatment on the Treated	3	1264	12.5%
Natural Direct Effect	3	316	3.1%
Natural Indirect Effect	3	790	7.8%
Counterfactual	3	1422	14.1%

Appendix 2: Number of Causal Inference Questions LLMs Answered Correctly in Simple Prompting Task

		Number of Questions that were answered Correctly by Model			
Rung of Question	Total number of Questions	GPT 3.5 Turbo	GPT 4	GPT 4 Turbo	Claude 3 Opus

1	624	342	398	377	360
2	296	236	279	273	286
3	472	249	349	355	328

Appendix 3: Prompt and Response Example of Causal Inference with Simple Prompting

```
{
        "Prompt": "You are an expert in causal inference.
The following question is not a typical commonsense query,
but rather a meticulously designed question created by a
professor specializing in causal inference, intended to
assess mastery of the course content. Be sure to start your
response with either the answer \"Yes\" or \"No\".Imagine a
self-contained, hypothetical world with only the following
conditions, and without any unmentioned factors or causal
relationships: Vaccination status has a direct effect on
vaccination reaction and getting smallpox. Getting smallpox
has a direct effect on smallpox survival. Vaccination
reaction has a direct effect on smallpox survival.For
unvaccinated individuals, the probability of smallpox
survival is 51%. For vaccinated individuals, the
probability of smallpox survival is 52%.Will vaccination
increase the chance of smallpox survival?",
        "response": "No. In the given hypothetical world,
vaccination does not increase the chance of smallpox
survival. Although the probability of smallpox survival is
slightly higher for vaccinated individuals (52%) compared
to unvaccinated individuals (51%), this difference cannot
be attributed to a causal effect of vaccination on smallpox
survival.\n\nThe key to understanding this is the presence
of a collider bias. In the described causal structure,
vaccination status affects both vaccination reaction and
getting smallpox, while both vaccination reaction and
getting smallpox affect smallpox survival. This means that
smallpox survival is a collider variable, and conditioning
on it (by comparing survival probabilities) opens up a non-
causal path between vaccination status and smallpox
survival through the collider.\n\nAs a result, the observed
```

```
difference in smallpox survival probabilities between
vaccinated and unvaccinated individuals is not due to a
causal effect of vaccination on survival, but rather due to
the collider bias introduced by conditioning on the
collider variable. In other words, the difference in
survival probabilities is a statistical artifact arising
from the specific causal structure and the act of
conditioning on the collider, rather than a genuine causal
effect of vaccination on survival.\n\nTherefore, based on
the given information and causal relationships in this
hypothetical world, vaccination does not increase the
chance of smallpox survival.",
        "first_word": "No.",
        "ground_truth": "yes",
        "prediction_correct": 0,
        "type of graph for question": "diamond",
        "type of query": "ate",
        "rung of question": 2
    }
```

Appendix 4: Prompt and Response Example of Causal Inference with Modified Causal Chain-of-Thought Prompting

```
{
        "Prompt": "Imagine a self-contained, hypothetical
world with only the following conditions, and without any
unmentioned factors or causal relationships: Encouragement
level has a direct effect on studying habit and exam score.
Studying habit has a direct effect on exam score. For
students who are not encouraged and do not study hard, the
probability of high exam score is 14%. For students who are
not encouraged and study hard, the probability of high exam
score is 50%. For students who are encouraged and do not
study hard, the probability of high exam score is 44%. For
students who are encouraged and study hard, the probability
of high exam score is 80%. For students who are not
encouraged, the probability of studying hard is 30%. For
students who are encouraged, the probability of studying
hard is 68%. If we disregard the mediation effect through
studying habit, would encouragement level negatively affect
exam score?   Let's think step by step  First, let's define
the variables: Let X = encouragement level; V2 = studying
habit; Y = exam score. Step 1) Extract the causal graph:
Identify the causal graph that depicts the relationships in
the scenario. The diagram should simply consist of edges
denoted in 'var1 -> var2' format, separated by commas. \n
```

Answer: X->V2,X->Y,V2->Y Step 2) Determine the query type: Identify the type of query implied by the main question. Please answer only with one of the choice options: correlation, marginal distribution, explaining away effect (exp_away),average treatment effect (ate), backdoor adjustment set (backadj), collider_bias, effect of the treatment on the treated (ett), natural direct effect (nde), natural indirect effect (nie), and counterfactual deterministic. \n Answer: nde Step 3) Formalize the query: Translate the query into its formal mathematical expression based on its type, utilizing the 'do(.)' notation or counterfactual notations as needed. \n Answer: E[Y_{X=1, V2=0} - Y_{X=0, V2=0}] Step 4) Gather all relevant data: Extract all the available data. Your answer should contain nothing but marginal probabilities and conditional probabilities in the form 'P(...)=...' or 'P(...|...)=...', each probability being separated by a semicolon. \n Answer: \\sum_{V2=v} P(V2=v|X=0)*[P(Y=1|X=1,V2=v) - P(Y=1|X=0, V2=v)] Step 5) Deduce the estimand: Given all the information above, deduce the estimand using skills such as do-calculus, counterfactual prediction, and the basics of probabilities. Answer step by step. \n Answer: P(Y=1 | X=0, V2=0) = 0.14\nP(Y=1 | X=0, V2=1) = 0.50\nP(Y=1 | X=1, V2=0) = 0.44\nP(Y=1 | X=1, V2=1) = 0.80\nP(V2=1 | X=0) = 0.30\nP(V2=1 | X=1) = 0.68 Step 6) Calculate the estimand: Insert the relevant data in Step 4 into the estimand, perform basic arithmetic calculations, and derive the final answer. Answer step by step. \n Answer: 0.30 * (0.80 - 0.44) + 0.68 * (0.50 - 0.14) = 0.30 Step 7) Finally, based on all the reasoning above, output one word to answer the initial question with just 'Yes' or 'No'. ",
        "response": "No",
        "expected_response": "no",
        "last_word": "No",
        "prediction_correct": 1,
        "type of graph for question": "mediation",
        "type of query": "nde",
        "rung of question": 3

Curriculum Vitae

Name:	Vicky Chang
Post-secondary Education and Degrees:	University of Western Ontario London, Ontario, Canada 2018-2022 Honours Bachelor of Medical Sciences University of Western Ontario London, Ontario, Canada 2022-2024 MSc Candidate, Neuroscience
Honours and Awards:	NSERC Undergraduate Student Research Award 2021
Related Work Experience	Teaching Assistant The University of Western Ontario COMPSCI 1033 Multimedia and Communications COMPSCI 2212 Introduction to Software Engineering COMPSCI 2120 Coding Essentials COMPSCI 2034 Data Analytics – Principles and Tools Dementia Prevention Initiative Principal Investigator: Dr. Vladimir Hachinski Advantages of Causal Inference in Dementia Prevention