Testing GPT-4's Ability to Reason Causally

Aimon Benfield-Chand and Rajvir Singh

Introduction

In the realm of cognitive abilities, humans possess a remarkable ability for causal reasoning, which sets us apart from other intelligent systems. This capacity to discern and comprehend causal relationships within the surrounding environment is an intrinsic aspect of human cognition. This ability is essential to our own survival, as it allows humans to predict which of our actions will maximize our survival. Our understanding of causation therefore drives human decision making, and is "at the heart of moral learning" according to David Hume (Pearl 2018, 10). Without the ability to model the consequences of ours' and others' actions, our ability to live in social communities would be greatly restricted. Causal inference allows individuals to make sense of the intricate web of cause and effect that shapes how events occur within all parts of the world. Cognitively, this involves the ability to identify patterns, create connections, and deduce the consequences of, and the precursors to, events.

The importance of causal reasoning goes beyond human's great ability to do so. Causal inference has profound implications for the development and evaluation of artificial intelligence, such as GPT-4. As we continue to create more sophisticated AI models, it becomes increasingly crucial to explore whether GPT-4 can causally infer. By investigating the capacity of GPT-4 to engage in causal inference, we unlock the potential for the system to not only comprehend but also reason about cause and effect relationships in the world. Such an advancement would mark

a significant step forward in AI capabilities, enabling it to navigate complex scenarios, make informed decisions, and exhibit a higher level of contextual understanding.

Through this project, we plan to test GPT-4's ability to reason about causation by providing reasoned responses to counterfactual scenarios. But first, let's define this. Simply put, a counterfactual is a hypothetical scenario that explores what would have happened if some aspect of the past had been different. This involves imagining an alternative course of events that diverges from the historical or observed reality. Counterfactuals are tightly connected with causal influences and computing counterfactuals rely heavily on the knowledge of causation (Balke 1994). That is, to know how a system would have been in the event of some alternative hypothetical event, one must understand the causal interactions that define the system in the first place.

Intricacies of causal modeling and its impact on various variables and agents lies at the core of Judea Pearl's groundbreaking research, exemplified by the introduction of the "do"-calculus within a causal model (Pearl 2018). In unraveling the realm of counterfactuals, we can discern three distinct layers: the observational layer, the intervention layer, and the counterfactual layer. Picture the observational layer as the repository of an agent's accumulated knowledge and experiences. In contrast, the intervention layer represents the dynamic actions an agent undertakes. Lastly, the counterfactual layer embodies the agent's ability to speculate and predict alternative outcomes based on different courses of action within the intervention layer, all while drawing from their knowledge in the observational and intervention layers (Correa 2021, 2).

Within the scope of our project, we aim to assess GPT-4's skills in applying a causal model both implicitly and explicitly. To achieve this, we will provide GPT-4 with prompts that

describe hypothetical scenarios that specify a subset of the three layers described above. In the first experiment, we provide only the intervention, and ask about the counterfactual, while leveraging GPT-4's domain knowledge of history to provide the observational layer of the causal model. In the second experiment, we provide a complete causal model defining causal relationships between variables, an observation, and an intervention, and ask GPT-4 to reason about a counterfactual. In the first part of experiment three, we specify the observation and the intervention, and instruct GPT-4 to reason about the counterfactual. In the second part of the experiment, however, we specify the observation and the counterfactual, and instruct GPT-4 to reason about the necessary intervention to produce the counterfactual from the observation. While several of these experimental scenarios therefore test the three layers in the same way, they use highly different knowledge domains and types of responses. It is also worth noting that the three experiments most drastically vary in how GPT-4's domain knowledge is used to complete the causal model; that is, in some experiments, the scenarios rely heavily on GPT-4's knowledge of the domain, while in others, domain knowledge is completely unnecessary.

It is worth noting that some of these scenarios will be articulated in informal prose, while others will use Pearl's formal "do"-calculus, prompting GPT-4 to craft its responses given a symbolic specification of the causal model (Pearl 2012). By contrasting GPT-4's performance in generating responses to these scenarios in both the informal and formal styles, we aim to see whether the use of an explicit symbolic causal model can aid an AI system's ability to discern and infer causation. Our results may provide valuable insights into the potential benefits conferred by embracing a causal model within the realm of AI inference.

Another question we seek to answer is the ability to evaluate causal scenarios when prompted with different processes for reasoning. To seek the answer to this question, we evaluate

how well GPT-4 is able to reason using three different methods: Input/Output (I/O), Chain of Thought (CoT), and Tree of Thought (ToT). Input/Output is essentially where GPT-4 will give a solution without giving a reason to back up its reasoning. This approach to reasoning has been shown to lead to more incorrect answers when compared to CoT and ToT. CoT is a common method for reasoning where GPT-4 comes to an answer after evaluating a series of events that are related. CoT has been seen to perform better than I/O but it is still to be seen whether it will perform better than ToT when tested with a causal scenario. Tree of Thought is a tree-based approach to reasoning where the best answers are recursively chosen while the less ideal answers are pruned (Yao 2023). This method of prompting LLMs with ToT is a recent innovation so we strive to see how well it will compare with other methods such as I/O and Chain of Thought when testing GPT-4's ability to causally infer. The results could provide information on the ability of GPT-4 to provide correct reasoning when coming to a conclusion for a causal scenario that it has been prompted with.

Experiment 1: Essay

Design

In the first experiment, our goal is to evaluate GPT-4's ability to understand and reason about causation in the real-world. To do this, we task GPT-4 with answering an open-ended counterfactual question regarding a complex historical topic in essay form, while deliberately selecting a topic for which information is likely to be well-represented in GPT-4s training data. We choose to use the domain of real-world history because it is naturally rich in causal forces. This allows us to evaluate GPT-4's ability to reason about complex causal scenarios without having to specify the relevant causal factors ourselves. In other words, we aim to leverage the

latent historical knowledge represented in GPT-4's training data to provide both the causal variables and the causal interactions that constitute the causal model. We acknowledge that GPT-4s training data related to the historical topic thus plays a significant role in its responses. However, our goal is to see how well it can synthesize this knowledge into a reasoned counterfactual argument, which we believe requires a non-trivial level of causal reasoning abilities, similar to those displayed by a university student when responding to an essay question on a history exam. Given a counterfactual prompt, we compare GPT-4's ability to reason about the causal counterfactual scenario using three different methods of prompt specification: Input/Output, CoT, and ToT. By doing so, our goal is to see how well of a response it can supply to a causal scenario based on the method of reasoning.

The causal scenario we prompted GPT-4 to write an essay about was "What might have happened if the U.S. had pursued a policy of containment instead of direct military intervention during the Vietnam War?". Additionally, based on the method of reasoning, we prompt GPT-4 with a methodology for doing reasoning. First, we have GPT-4 generate three different thesis statements that each answer the counterfactual question. Then, for Input/Output, we simply ask GPT-4 to write an entire essay for each thesis start to finish. For CoT, we prompt GPT-4 to write the essay paragraph by paragraph, where it generates and then selects the best of three versions of the next paragraph based on how well it connects to the previous paragraphs. For ToT, we have GPT-4 recursively generate the essay. This is done by prompting GPT-4 to first select the best of the three thesis statements to use. Then, we have GPT-4 generate three sets of three topic sentences that support the thesis and select one of these sets to use. For each of these topic sentences, we have GPT-4 generate nine pieces of supporting evidence and select three of these

to use. Finally, we compile this content into an essay outline and have GPT-4 synthesize the outline into a final draft.

We specifically choose to represent I/O, CoT, and ToT in this way so that all three methods would generate the content of three essays, while only selecting the best one in the end. The variation between the methods is where this decision of "what is best" comes into play. In the I/O method, the decision is only made at the end after generating the three entire essays in parallel. In the CoT method, the decision is at every step in a linear writing process, such that GPT-4 generates and then chooses between three thesis statements, three introductions, three first body paragraphs, three second body paragraphs, and so on to produce exactly one essay where each paragraph is the best of three options. In the ToT method, the decision is made at each level of the generated tree. This can be thought of as selecting the ideas at each level of an essay outline that best support the idea at their parent level.

To evaluate which approach for reasoning causal scenarios performed best, we have GPT-4 generate a set of 10 evaluation criteria. We then have GPT-4 select the best of the three I/O essays using these criteria, and afterwards select the best of the three essay methods (best I/O, CoT, ToT). The experiment was performed using the OpenAI AI with GPT-4, which allowed us to control the temperature hyperparameter of GPT-4's responses. We used temperature=0 to guarantee deterministic and replicable results for this experiment.

Results

Scores given by GPT-4 to each final essay based on GPT-4's own criteria

	I/O	CoT	ToT
1. Clarity and coherence of the thesis statement	9	9	8
2. Logical organization and structure	9	9	8

3. Depth and quality of counterfactual analysis	9	8	7
4. Use of evidence and examples	8	7	7
5. Critical thinking and originality	9	8	7
6. Persuasiveness of the argument	9	8	7
7. Clarity and precision of language	9	9	8
8. Proper formatting and citation	9	8	8
9. Engagement with counterarguments	8	7	6
10. Quality of the conclusion	9	9	8
Total	88	82	74

GPT-4's Reasoning:
"Based on the scores, Essay 3 is the best of the three essays. It has a clear and coherent thesis statement, a logical organization and structure, and provides an in-depth counterfactual analysis. The essay effectively supports its arguments with relevant historical evidence and examples, demonstrates critical thinking and originality, and presents a persuasive argument. The language is clear and precise, and the essay follows proper formatting and citation guidelines. Additionally, Essay 3 engages with counterarguments and has a strong conclusion that effectively summarizes the main points and restates the thesis statement. Overall, Essay 3 provides the most comprehensive and compelling analysis of the counterfactual question."

Analysis

The results of the experiment seem to indicate that simple input / output prompting methods are preferable for writing tasks like essay generation. The I/O essay tied or outperformed the CoT and ToT essays on every single criterion. On the one hand, this result is surprising. We hypothesized that the added reasoning structure and focus on a single essay would yield an advantage to the CoT and ToT methods over the I/O method. For humans, the tools of outlining and weighing multiple potential ideas or arguments is often very helpful when writing essays. However, it appears that the end-to-end generation process of the I/O method was

sufficient for generating clear and well-reasoned essays and that the additional structures of the CoT and ToT methods only complicated the process without providing any clear benefit to GPT-4. Indeed, it is somewhat reasonable to think that a large language model would be good at generating well-reasoned argumentative essays without any additional help. After all, natural language in text is exactly what GPT-4 was designed to model.

Onto the main question: what do these results tell us about GPT-4's ability to reason about causation? At least in humans, the ability to construct a convincing counterfactual argument certainly indicates some level of causal understanding. Yet this is likely not true for a language model like GPT-4. While GPT-4's ability to argue about counterfactuals at a level similar to humans does demonstrate its skills in reasoning, it does not imply any level of understanding. Rather, GPT-4 displays mere competence in its ability to grapple with causation through counterfactual reasoning. This competence is undoubtedly learned from GPT-4's training data, and it is likely that most of the ideas of each essay's argument were drawn directly from this training data. Still, this is true for everyone. The goal of an essay question on a history exam is to see how well a student can synthesize the course materials into their own argument. On the whole, it appears that GPT-4 exhibits this skill in the experimental results. While the experiment does not therefore demonstrate GPT-4's ability to reason about novel causation, it does demonstrate GPT-4's ability to reason in a domain filled by causation.

Extensions

There are many possible extensions of the experiment. We chose to represent the decision at each step in CoT and ToT be which thesis or main idea to expand on. Because of this, whatever decision was made earlier on in the CoT and ToT process (an earlier paragraph or a higher outline level, respectively) has the effect of pruning many possible ideas from being

considered later on in the process. One variation of this would be to allow for backtracking, where the decision of which content to select is made after all the content is generated. This change would be most pronounced in the ToT method, allowing for GPT-4 to generate the entire tree of the essay outline initially and then later backtrack up the tree to select the best ideas by pruning the already-generated branches. Alternatively, each level of ToT could be represented as each sequential paragraph in the essay, just like CoT. With backtracking, this version of ToT could similarly expand out each paragraph option branch into a full tree that includes options for the later paragraphs before eventually pruning down the tree to select the best version. Our reasoning for not attempting this approach, however, is that it causes ToT to generate exponentially more content the deeper the tree, which would make it difficult to fairly compare to the I/O and CoT methods. Nonetheless, it would be an interesting extension for exploring the specific application of ToT to writing tasks.

Experiment 2: Causal Model and Do Calculus

Design

In the second experiment, we aim to test GPT-4's ability to identify mistakes in causal reasoning concerning a completely specified causal model. To do this, we give GPT-4 a series of counterfactual statements regarding the causal model and ask it to evaluate whether each statement is correct or incorrect according to the provided causal model. These counterfactual statements are specified in both natural language (i.e. english) and in Judea Pearl's formal do-calculus notation. We then compare GPT-4's ability to identify logical flaws in counterfactual statements when prompted in the do-calculus notation vs natural language. By doing so, we aim

to see if Pearl's formal causal notation enhances GPT-4's ability to reason about causation over standard language.

We choose to use the NBA playoffs as inspiration for our causal model because it offers easy specification of causal interactions and is a real-world scenario that GPT-4 has been trained on. This latter quality allows GPT-4 to harness its domain knowledge of basketball and the NBA in order to effectively interpret and contextualize the causal model we specify. However, our causal model is a greatly simplified version of the NBA playoffs, which eliminates the potential for GPT-4 to accurately reason about the counterfactual statements we provide simply by regurgitating its training data.

Our causal model consists of the following variables: "Injured" (I), "Home Court Advantage in the Semifinals" (HCS), "Home Court Advantage in the Finals" (HCF), "Victory" (V), "Semifinal Champion" (SC), "Finals Champion" (FC), "Skill Advantage" (SA), "Finals Most Valuable Player" (MVP), and "Hall of Fame" (HF). The causal interactions between these variables are defined as follows.

- Injured (I) → Victory (V): A team getting injured (Team(Injury=1)) decreases that team's chance of victory.
- Injured (I) → Skill (SA): A team getting injured (Team(Injured=1)) reduces that team's skill advantage.
- Home Court Advantage (HC) → Skill Advantage (SA): A team having home court advantage (Team(HC=1)) increases that team's skill advantage, since they play better at home.
- Semifinal Champion (SC) → Finals Champion (FC): A team becoming semifinal champion Team(HC=1) increases that team's chance of becoming finals champion.
- Skill Advantage (SA) → Victory (V): A team possessing a skill advantage over their opponent (Team(SA=1)) positively influences their chance of victory.
- Finals Champion (FC) → Most Valuable Player (MVP): A team becoming finals champion (Team(FC=1)) increases the chance that one of their player's will win most value player Player(MVP=1).
- Victory (V) → Semifinal Champion (SC): A team winning a game in the playoffs (Team(V=1)) increases their chance of becoming semi-finals champion.
- Victory (V) → Finals Champion (V): A team winning a game in the playoffs (Team(V=1) increases their chance of becoming finals champion.

- Finals Most Valuable Player (MVP) → Hall of Fame (HF): A player winning the Finals MVP (Player(MVP=1)) increases the player's chance of becoming a hall of fame player.

Results

Based on the results of the experiment, GPT-4 performed better at finding the mistake in the counterfactuals when presented the counterfactual in natural language as compared to Pearl's do-calculus. When prompted in natural language, GPT-4 was able to determine the mistake 100% of the time. On the other hand, when prompted in the do calculus, GPT-4 determined the mistake 50% of the time. These results mean that GPT-4 is able to causally reason better with natural language prompting compared to when prompted with the do-calculus. These results were surprising to see. We hypothesize that since the do calculus provided a formal way of interpreting the causal model that this would lead to better performance compared to natural language. However, in the results, natural language prompting led to GPT-4 providing the correct results along with better reasoning in the scenarios presented.

We believe that this was the case because GPT-4 had a more clear understanding of the counterfactual and could do a better job of translating it to fit the causal model when using natural language than with the do-calculus. Additionally, since LLMs like GPT-4 are trained on datasets of natural language, it makes sense that they would be able to interpret a question presented in a natural language compared to a formal language like the do-calculus.

Experiment 3: Mathematical Counterfactuals

In the third experiment, we continue to test GPT-4's ability to causally reason by evaluating its capacity to find the correct solution to counterfactual scenarios. Counterfactual scenarios are strongly linked with causal scenarios since computing counterfactuals rely on

knowledge of causal reasoning. In this experiment, we pose two different types of counterfactual problems to GPT-4, which we label change scenarios and solve scenarios. Both scenarios use linear math equations with three variables x, y, and z as their domain due to the simplicity of their causal interactions. By representing causation as the effects of three variables on the result of a simple linear equation, we strive to see how well GPT-4 can reason about counterfactuals in simple models where correctness is clearly defined. Through the framework of a causal model, this can be seen as specifying the causal variables (i.e. x, y, and z) and interactions in the scenario (the variable coefficients), while using GPT-4's domain knowledge of math and arithmetic as the rules for how to manipulate these causal interactions.

Part 1: Change Scenarios

Design

In change scenarios, we state to GPT-4 an observation that '3x - 9y + 5z = -11' given the current values of x, y, and z. We then provide GPT-4 with an intervention to the variables in the form of an augmented assignment (e.g. "x is decremented by 2"). Given the observation and intervention, we ask GPT-4 to determine the new result on the RHS of the equation. In this type of scenario, the mathematical equation and GPT-4's prior knowledge of math would be the observational layer, the augmented assignment to each of the variables would be the intervention layer, and the counterfactual layer would be the response GPT-4 gives based on its observational and intervention layers.

In this experiment, we plan to utilize two methods of reasoning, I/O, and CoT, where we will run a set of 15 scenarios using each of these methods. In I/O, GPT-4 will give the solution to the question without providing any reasoning. In CoT, GPT-4 will give a step-by-step

explanation of how it came to its result. By utilizing two reasoning methods, we see whether GPT-4 can causally reason better based on the reasoning method. Additionally, these scenarios will be run using the GPT-4 API with the temperature hyperparameter equal to zero to guarantee deterministic results.

Results

After experimenting, we found that GPT-4 exhibits a great improvement in performance when prompted with Chain of Thought (CoT) examples rather than Input/Output examples. Specifically, GPT-4 achieved 100% accuracy with CoT prompting as opposed to 0% accuracy with I/O prompting after fifteen trials. These results suggest that CoT prompting is far superior to I/O prompting for this type of task. However, this is not surprising. By design, the CoT few-shot examples broke down the problem of calculating the counterfactual response from some intervention into simple steps requiring only basic arithmetic calculation. On the other hand, the I/O few-shot examples stated only the correct answer without any intermediate reasoning. Whereas the CoT few-shot prompting forced GPT-4 to copy its structure by breaking down the problem step-by-step into manageable calculations, the I/O few-shot prompting forced GPT-4 to output the result without any intermediate calculation. For GPT-4 to output the correct answer under I/O prompting, it would have needed to perform something closer to the recall of memorized facts than logical problem-solving.

In the framework of causation, these results suggest that step-by-step reasoning is fundamental to deduction. Without the ability to work out sequentially how changes to each variable produced a change to the resulting value, the I/O method was completely unable to give an accurate counterfactual based on the known prior and intervention. Conversely, the CoT method enabled GPT-4 to achieve perfect accuracy by reasoning out the impact of each variable

intervention on the counterfactual response. Thus, we find that GPT-4 demonstrates competence and accuracy regarding simple counterfactuals when allowed to reason. However, GPT-4 demonstrates poor causal competency when forced to intuit a counterfactual output without any intermediate reasoning.

Part 2: Solve Scenarios

Design

In the solve scenarios, we test GPT-4's ability to causally reason by determining how well it can come up with an intervention after providing it with the observational and counterfactual layers. In this type of counterfactual, we provide GPT-4 with an equation like "2x + 3y - 4z = -11" and values for each of the variables that satisfy the equation. In the previous equation, the values would be "x=2, y=-1, z=3". Then, we ask GPT-4 to find the most parsimonious integer-valued assignments for x, y, and z such that the equation now evaluates to some different integer. In this context, we define a parsimonious solution as one that assigns new values to the fewest number of the variables x, y, and z. If multiple solutions change the same number of variables, we tell GPT-4 to select the solution that changes these variables by the smallest absolute amount. By providing GPT-4 with a constraint such as finding the parsimonious solution, we force GPT-4 to select the most optimal solution from the infinite sub-optimal solutions that exist. This allows for easier evaluation of performance and requires additional reasoning by GPT-4.

In the solve experiment, we compare I/O and ToT as methods for reasoning where the structure of the response is not specified or constrained. We evaluate each of the reasoning methods by comparing their ability to accurately produce the correct solution to three different

‘solve’ problems. In terms of prompts, the only difference between the two methods is that the ToT prompts explicitly instruct GPT-4 to solve the problem using the Tree of Thought framework, whereas the input/output prompts specify no such structure requirement. To perform these tests, we use the ChatGPT browser application instead of the API in order to simulate a normal use-case of GPT-4 and to enable the use of plugins–specifically, Wolfram and Linkreader. Though the Wolfram plugin could be used to solve larger portions of the problems, we allow GPT-4 to only use Wolfram for small arithmetic computations. Our goal with this is to eliminate arithmetic errors in order to isolate the effect of reasoning on problem accuracy. With the Linkreader plugin enabled and a link to the seminal ToT research paper included in our ToT prompts, GPT-4 can use the linkreader to read the contents of the ToT paper and thereby apply ToT without us having to specify the methodology ourselves. While this may result in less precise applications of ToT to the problems, it better reflects the normal use case of GPT-4, allowing us to test whether ToT is still useful without the high overhead involved in manually specifying the inputs and outputs as is done in the Princeton DeepMind paper. Indeed, if ToT can be implemented by GPT-4 automatically without human involvement, its potential as a general reasoning tool is greatly bolstered, including in the domain of causal inference.

Results

The following tables summarize the performance of ChatGPT-4 with Input/Output prompting and Tree of Thought (ToT) prompting for the three problems. Note, each value represents the number of trials (out of three) for each prompting style that satisfied the criterion.

Problem 1: Changing any of x, y, or z individually yields solutions

	(A) Parsimony	(B) Valid Solution	(C) Systematic Approach	(D) Correct Reasoning

I/O	2	2	1	3
ToT	1	2	3	1

Problem 2: Changing either y or z individually yields solutions

	(A) Parsimony	(B) Valid Solution	(C) Systematic Approach	(D) Correct Reasoning
I/O	2	3	3	2
ToT	2	2	3	0

Problem 3: Changing x and y yields solutions

	(A) Parsimony	(B) Valid Solution	(C) Systematic Approach	(D) Correct Reasoning
I/O	0	0	1	2
ToT	1	1	3	0

From the results, it appears that ToT performed slightly worse than I/O for the simpler first two problems, but performed slightly better for the more complex third problem. To explain this result, it helps to examine GPT-4s performance on criterion C and criterion D. In terms of criterion C, ToT prompting led GPT-4 to follow a systematic approach when solving the problem far more frequently than did I/O prompting. Generally, systematic approaches are more advantageous when solving complex problems than when solving simple ones. Alternatively, a naive approach might suffice for simple problems and be easier to correctly implement than a more sophisticated approach. The performance of GPT-4 between problems 1, 2, and 3 seems to be subject to this tradeoff. Whereas the first two problems only require a change in value to a single variable, problem 3 requires two variables to be changed. This means that a relatively unsystematic approach is more likely to find the optimal solution to the first two problems than to the third one, where a deeper search for possible solutions is needed. In this case, the enhanced problem-solving structure required by ToT appears to have been worth the logical overhead required for its correct implementation. For simple problems, however, the unsystematic

approach produced by I/O prompting appears sufficient, and the additional structure required by ToT often led to mistakes in reasoning and longer responses which were more prone to arithmetic errors, as evidenced by the poor performance on criterion D for the ToT prompting. However, confidence in these trends is weak due to the limited sample size of the results. From just the nine trials collected of ToT and I/O prompting, it seems that GPT-4 performed almost identically on criterion A and criterion B. Whether this inconclusive result is due to insufficient data remains to be known. One extension of this experiment would therefore be to perform many more trials.

Within the lens of causal reasoning, these results suggest that GPT-4's abilities are limited to problems with simple causal interactions. The goal of 'solving' problems is to determine the best intervention to produce some counterfactual response given some ground-truth observation. For scenarios where only a small intervention needed to be made (i.e. changing the value of only one variable) to produce the desired counterfactual, GPT-4 appears to be fairly proficient, notwithstanding the small sample size of our data. But for scenarios where less proximate interventions must be made (i.e. changing the values of multiple variables), GPT-4 appears decisively less competent regardless of the reasoning method (I/O or ToT).

This result is intuitive from a causal viewpoint since fewer causal interactions are involved the more direct the effect of an intervention on the counterfactual. In the real world, this type of 'simplicity' can be viewed as a function of either the absolute magnitude of the intervention or of the temporal distance between the counterfactual and intervention. Given a counterfactual, it is harder to deduce the precise intervention that produced a divergence in the timeline the farther back that intervention is from the counterfactual. Likewise, it is harder to deduce the intervention the smaller it is in magnitude. Applying these principles, GPT-4's

reduced accuracy on the more complex ‘solve’ problems makes sense, since they require less obvious interventions that are more distantly connected to the difference between the desired counterfactual and the ground-truth observation.

One rebuttal to this point is that the causal relationships in the ‘solve’ problems are not difficult for humans to maneuver and should then be as easy for GPT-4. While it is true that humans can easily solve these types of problems, this is due to the use of mathematical reasoning tools rather than causal reasoning tools. It is understood that GPT-4 is itself bad at math, which is a clear disadvantage that must be accounted for when comparing GPT-4’s performance to a human’s. To explore exactly how much of an effect mathematical skill and intuition has, if any, on GPT-4’s performance, one future extension could be to apply the concept of the ‘solve’ problems to tasks in a domain more suited to GPT-4’s skillset, such as writing and reading.

Conclusion

In this paper, we explored GPT-4's capacity for causal reasoning—a crucial cognitive ability that distinguishes human intelligence and has significant implications for the development of artificial intelligence. We recognized the importance of causal inference in human cognition, enabling us to understand and predict the consequences of actions, make informed decisions, and engage in social interactions. Our objective was to assess whether GPT-4 could replicate this ability, contributing to AI's contextual understanding and decision-making capabilities.

To evaluate GPT-4's causal inference, we designed a series of experiments. The first experiment involved an open-ended causal scenario, where GPT-4 was prompted to reason using three different methods: Input/Output, Chain of Thought, and Tree of Thought. We aimed to

identify the most effective approach for reasoning causal scenarios by assessing GPT-4's responses.

The second experiment focused on GPT-4's comprehension of causal reasoning by testing its ability to identify mistakes in causal questions. We presented NBA-themed scenarios framed in both formal and natural language, comparing GPT-4's responses to determine its reasoning based on question structure.

The third experiment examined GPT-4's capacity to handle mathematical counterfactuals, which rely on causal reasoning. We challenged GPT-4 with ‘change’ scenarios and ‘solve’ scenarios to evaluate its ability to reason about interventions and determine outcomes based on mathematical equations.

Overall, our exploration of GPT-4's causal inference abilities revealed its potential to reason, evaluate scenarios, and identify errors in causal reasoning. By leveraging symbolic causal models and different reasoning approaches, GPT-4 demonstrated its ability to engage with causal models and contribute to contextual comprehension and decision-making.

Our research highlights the significance of incorporating causal models and reasoning techniques in AI development. By integrating formal and informal questioning styles, AI systems like GPT-4 can enhance their ability to discern, infer, and reason about causation. These advancements have implications for decision-making, problem-solving, and the potential of AI systems to navigate complex scenarios.

In conclusion, our research offers insights into the causal inference abilities of GPT-4. By further exploring and refining the integration of causal models, AI systems can bridge the gap between human-like reasoning and machine intelligence. This opens opportunities for AI to

navigate the intricate web of cause-and-effect relationships, contributing meaningfully to various domains and societal challenges.

References

Balke, A., & Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. Uncertainty Proceedings 1994, 46–54. https://doi.org/10.1016/b978-1-55860-332-5.50011-0.

Correa, J. A., Lee, S., & Bareinboim, E. (2021). Nested Counterfactual Identification from Arbitrary Surrogate Experiments. In *arXiv (Cornell University)* (Vol. 34). Cornell University. https://arxiv.org/pdf/2107.03190.

Goldman, A. I. (1967). A Causal Theory of Knowing. The Journal of Philosophy, 64(12), 357–372. https://doi.org/10.2307/2024268.

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

Pearl, J. (2012). The do -calculus revisited. In *arXiv (Cornell University)* (pp. 3–11). Cornell University. https://arxiv.org/pdf/1210.4852.

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

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *ArXiv [Cs.CL]*. Retrieved from http://arxiv.org/abs/2305.10601.

Appendix

NOTE: Full results viewable at
https://drive.google.com/drive/folders/1hU1W5Z0_kyky1tAZmiOhfnAQHPgtz4UL?usp=sharing

Appendix A: Essays

A.1: I/O Essays

Essay 1 (Selected as best)

Introduction: The Vietnam War remains one of the most controversial and divisive conflicts in American history, with its legacy continuing to shape U.S. foreign policy and public opinion. Central to the debate surrounding the war is the question of whether the United States should have pursued a policy of containment, rather than direct military intervention, in its efforts to prevent the spread of communism in Southeast Asia. This essay argues that if the U.S. had pursued a policy of containment instead of direct military intervention during the Vietnam War, the spread of communism in Southeast Asia would have been slowed, ultimately leading to a more stable and less costly outcome for the United States and its allies.

Body Paragraph 1: One of the most significant consequences of pursuing a policy of containment rather than direct military intervention would have been the reduced loss of life and economic costs for the United States and its allies. The Vietnam War resulted in the deaths of over 58,000 American soldiers and an estimated 2 million Vietnamese civilians and military personnel. Additionally, the war cost the United States approximately $168 billion, which equates to over $1 trillion in today's currency. By focusing on containment, the U.S. could have avoided these devastating human and financial costs, while still working to prevent the spread of communism in the region. This approach would have allowed the U.S. to allocate resources more effectively, both domestically and in other areas of foreign policy.

Body Paragraph 2: A policy of containment would have also allowed the United States to focus on strengthening regional allies and preventing the so-called "domino effect," in which the fall of one country to communism would lead to the fall of its neighbors. By providing military and economic aid to countries in Southeast Asia, the U.S. could have bolstered their ability to resist communist influence and insurgency. This approach would have been more in line with the successful containment strategies employed during the early years of the Cold War, such as the Marshall Plan in Europe. By focusing on building strong regional alliances, the U.S.

could have created a more effective barrier against the spread of communism, without the need for direct military intervention.

Body Paragraph 3: Finally, by pursuing a policy of containment rather than direct military intervention, the United States could have avoided many of the negative consequences that arose from its involvement in the Vietnam War. The war led to widespread anti-war protests and social unrest in the United States, as well as a loss of faith in the government and its institutions. Furthermore, the war had a lasting impact on U.S. foreign policy, with the so-called "Vietnam Syndrome" leading to a reluctance to engage in military interventions in subsequent years. By avoiding direct military intervention, the U.S. could have maintained greater domestic stability and preserved its international reputation, while still working to prevent the spread of communism in Southeast Asia.

Conclusion: In conclusion, if the United States had pursued a policy of containment instead of direct military intervention during the Vietnam War, it is likely that the spread of communism in Southeast Asia would have been slowed, ultimately leading to a more stable and less costly outcome for the United States and its allies. By avoiding the devastating human and financial costs of the war, focusing on strengthening regional alliances, and preventing the negative consequences of direct intervention, the U.S. could have achieved its strategic objectives in a more effective and sustainable manner. While it is impossible to know for certain how history might have unfolded under different circumstances, this counterfactual analysis offers valuable insights into the potential benefits of a containment strategy during the Vietnam War and its implications for U.S. foreign policy more broadly.

Essay 2

Introduction: The Vietnam War remains one of the most controversial and divisive conflicts in modern history, with the United States' direct military intervention in the region sparking widespread debate and criticism. This essay will explore the counterfactual scenario in which the U.S. pursued a policy of containment instead of direct military intervention during the Vietnam War. The thesis statement for this essay is: If the U.S. had pursued a policy of containment instead of direct military intervention during the Vietnam War, the conflict would have escalated into a larger regional war, drawing in other major powers and leading to a more dangerous and unstable global situation.

Body Paragraph 1: If the United States had chosen to pursue a policy of containment during the Vietnam War, it is likely that the conflict would have escalated into a larger regional war. Containment, as a strategy, would have involved the U.S. providing military and economic support to South Vietnam and other regional allies, while avoiding direct military intervention. This approach would have emboldened North Vietnam and its

communist allies, such as China and the Soviet Union, to increase their support for the Viet Cong and other communist forces in the region. As a result, the conflict would have spread beyond Vietnam's borders, engulfing neighboring countries such as Laos, Cambodia, and potentially even Thailand. This escalation would have led to a more protracted and destructive war, with devastating consequences for the people and nations involved.

Body Paragraph 2: The escalation of the Vietnam War into a larger regional conflict would have drawn in other major powers, further complicating the situation and increasing the risk of a global confrontation. With the U.S. pursuing a policy of containment, the Soviet Union and China would have been more inclined to increase their military and economic support for North Vietnam and other communist forces in the region. This could have led to a dangerous arms race and proxy war between the U.S. and its communist rivals, with each side seeking to outmaneuver the other and gain the upper hand in the region. The involvement of these major powers would have raised the stakes of the conflict, making it more difficult to achieve a negotiated settlement and increasing the likelihood of a larger, more destructive war.

Body Paragraph 3: The escalation of the Vietnam War into a larger regional conflict, involving other major powers, would have created a more dangerous and unstable global situation. The increased tensions between the U.S., the Soviet Union, and China would have heightened the risk of a broader confrontation, potentially even leading to a direct conflict between these nuclear-armed superpowers. Furthermore, the spread of the conflict to neighboring countries would have destabilized the entire region, creating a power vacuum and opening the door for other extremist groups and ideologies to take root. This instability would have had far-reaching consequences, not only for Southeast Asia but also for the global balance of power and the prospects for peace and security in the decades to come.

Conclusion: In conclusion, if the United States had pursued a policy of containment instead of direct military intervention during the Vietnam War, the conflict would have escalated into a larger regional war, drawing in other major powers and leading to a more dangerous and unstable global situation. While the U.S.'s direct military intervention in Vietnam was undoubtedly costly and controversial, this counterfactual analysis suggests that a policy of containment would have had even more dire consequences for the region and the world. As we reflect on the lessons of the Vietnam War, it is important to consider not only the mistakes that were made but also the potential consequences of alternative strategies and policies.

Essay 3

Introduction: The Vietnam War remains one of the most controversial and divisive conflicts in American history, with its origins rooted in the Cold War struggle between the United States and the Soviet Union. The U.S. chose to engage in direct military intervention in Vietnam, ultimately leading to a costly and protracted conflict with significant consequences for both the United States and the region. This essay will argue that if the U.S. had pursued a policy of containment instead of direct military intervention during the Vietnam War, the United States would have lost credibility as a global superpower, leading to a decline in its influence and a shift in the balance of power towards the Soviet Union and China.

Body Paragraph 1: If the United States had chosen a policy of containment rather than direct military intervention in Vietnam, it would have risked losing credibility as a global superpower. The U.S. had established itself as a defender of democracy and capitalism in the post-World War II era, and its commitment to these principles was a key component of its global influence. By not intervening militarily in Vietnam, the U.S. would have appeared weak and indecisive in the face of communist aggression, undermining its reputation as a strong and reliable ally. This loss of credibility would have had significant consequences for the United States' ability to project power and influence around the world, as other nations might have questioned its commitment to defending democratic values and containing the spread of communism.

Body Paragraph 2: A policy of containment in Vietnam would have also led to a decline in U.S. influence in the region and around the world. The United States had established a network of alliances and partnerships in Asia, including with countries like South Korea, Japan, and the Philippines, which were critical to its efforts to contain the spread of communism. By not intervening militarily in Vietnam, the U.S. would have sent a message to its allies that it was not fully committed to their defense, potentially leading to a weakening of these alliances and a reduction in American influence in the region. Furthermore, the perception of American weakness and indecision would have emboldened the Soviet Union and China, who might have sought to exploit the situation to expand their own influence and challenge the United States in other parts of the world.

Body Paragraph 3: Finally, a policy of containment in Vietnam would have led to a shift in the balance of power towards the Soviet Union and China. With the United States appearing weak and indecisive, the Soviet Union and China would have been more likely to pursue aggressive policies in other parts of the world, potentially leading to a further spread of communism and an increase in global tensions. This shift in the balance of power could have had significant consequences for the overall stability of the international system, as the United States would have been less able to

counteract the influence of its rivals and maintain the global order it had helped to establish in the aftermath of World War II.

Conclusion: In conclusion, if the United States had pursued a policy of containment instead of direct military intervention during the Vietnam War, it would have risked losing credibility as a global superpower, leading to a decline in its influence and a shift in the balance of power towards the Soviet Union and China. While the human and financial costs of the Vietnam War were undoubtedly high, the alternative scenario of containment would have had significant consequences for the United States and the international system as a whole. By intervening militarily in Vietnam, the U.S. demonstrated its commitment to defending democratic values and containing the spread of communism, even at great cost, and maintained its position as a global superpower in the face of rising challenges from the Soviet Union and China.

I/O Essay Comparison Evaluation

Essay 1:
1. Clarity and coherence of the thesis statement: 9
2. Logical organization and structure: 9
3. Depth and quality of counterfactual analysis: 8
4. Use of evidence and examples: 8
5. Critical thinking and originality: 8
6. Persuasiveness of the argument: 9
7. Clarity and precision of language: 9
8. Proper formatting and citation: 8 (citations not included)
9. Engagement with counterarguments: 7
10. Quality of the conclusion: 9

Essay 2:
1. Clarity and coherence of the thesis statement: 8
2. Logical organization and structure: 9
3. Depth and quality of counterfactual analysis: 7
4. Use of evidence and examples: 7
5. Critical thinking and originality: 7
6. Persuasiveness of the argument: 7
7. Clarity and precision of language: 8
8. Proper formatting and citation: 8 (citations not included)
9. Engagement with counterarguments: 6
10. Quality of the conclusion: 8

Essay 3:
1. Clarity and coherence of the thesis statement: 8
2. Logical organization and structure: 9
3. Depth and quality of counterfactual analysis: 8
4. Use of evidence and examples: 8
5. Critical thinking and originality: 8

6. Persuasiveness of the argument: 8
7. Clarity and precision of language: 9
8. Proper formatting and citation: 8 (citations not included)
9. Engagement with counterarguments: 7
10. Quality of the conclusion: 8

Based on the scores, Essay 1 is the best of the three essays. It has a clear and coherent thesis statement, a logical organization and structure, and a persuasive argument. The depth and quality of counterfactual analysis, use of evidence and examples, and critical thinking and originality are also strong in this essay. While it could improve in engaging with counterarguments and providing proper citations, the overall quality of the essay is high, making it the best among the three.

A.2: CoT Essay

Introduction: The Vietnam War remains a contentious and divisive topic in American history, with many questioning the wisdom of direct military intervention in the conflict. If the U.S. had pursued a policy of containment instead of direct military intervention during the Vietnam War, the spread of communism in Southeast Asia would have been slowed, ultimately leading to a more stable and less costly outcome for the United States and its allies. This essay will explore the potential consequences of such a policy shift, examining the implications for regional stability, the human and financial costs of the war, and the broader geopolitical landscape.

Body Paragraph 1: One of the primary consequences of a containment policy during the Vietnam War would have been the slowing of communist expansion in Southeast Asia. By focusing on economic and political support for non-communist governments in the region, the United States could have strengthened these countries' ability to resist communist influence. For example, the U.S. could have provided more substantial aid to countries like Thailand and Malaysia, which were facing their own communist insurgencies. This approach would have been similar to the Marshall Plan in Europe, which helped rebuild Western European economies and prevent the spread of communism after World War II. By bolstering the economies and political stability of non-communist countries in Southeast Asia, the United States could have created a more effective barrier against the spread of communism.

Body Paragraph 2: A containment policy would have allowed the United States to focus on providing economic and military aid to South Vietnam and other countries in the region, rather than committing large numbers of American troops to the conflict. This approach would have been similar to the successful Marshall Plan in Europe, which helped to rebuild Western European economies and prevent the spread of communism after World War II. By investing in the development and stability of Southeast Asian nations, the U.S. could have created a stronger bulwark against communist expansion. Additionally, this strategy would have allowed the U.S. to maintain a more flexible and adaptable foreign policy, as it would not have been as deeply entangled in the conflict as it was with direct military intervention.

Body Paragraph 3: Another possible outcome of the U.S. pursuing a policy of containment during the Vietnam War would have been the increased likelihood of a negotiated settlement between North and South Vietnam. Without direct American military intervention, the conflict may have remained more localized, allowing for diplomatic efforts to take center stage. The U.S. could have worked with regional and international partners to broker a peace agreement that would have maintained the sovereignty of

South Vietnam while addressing the concerns of the communist North. This approach could have prevented the devastating consequences of a prolonged and bloody conflict, ultimately saving countless lives and resources.

Conclusion: In conclusion, had the U.S. chosen a policy of containment over direct military intervention in the Vietnam War, the spread of communism in Southeast Asia would have been slowed, leading to a more stable and less costly outcome for the United States and its allies. This alternative approach would have allowed the U.S. to maintain its global influence while avoiding the devastating human and financial costs of a prolonged war. Ultimately, this counterfactual scenario demonstrates the importance of carefully considering the long-term consequences of foreign policy decisions and the potential benefits of pursuing diplomatic solutions over military force.

A.3: ToT Essay

Introduction: The Vietnam War remains one of the most controversial and divisive conflicts in American history. The United States' decision to engage in direct military intervention in Vietnam has been the subject of much debate and analysis. This essay will explore the counterfactual scenario of the U.S. pursuing a policy of containment instead of direct military intervention during the Vietnam War. The thesis of this essay is that if the U.S. had pursued a policy of containment, the spread of communism in Southeast Asia would have been slowed, ultimately leading to a more stable and less costly outcome for the United States and its allies.

Body Paragraph 1: The first supporting argument for this thesis is that a containment policy would have allowed the U.S. to focus on diplomatic and economic efforts to limit the spread of communism in Southeast Asia. The U.S. could have provided economic aid to non-communist countries in the region, strengthening their economies and making them less susceptible to communist influence. Diplomatic efforts could have been made to establish alliances with key regional powers, creating a network of countries committed to resisting communism. Additionally, the U.S. could have supported local anti-communist movements and leaders, providing them with resources and training to effectively counter communist forces. This approach would have allowed the U.S. to address the spread of communism without resorting to direct military intervention.

Body Paragraph 2: The second supporting argument is that the absence of direct military intervention would have reduced the human and financial costs of the Vietnam War for the United States. Over 58,000 American soldiers lost their lives in the Vietnam War, a number that would have been much lower if the U.S. had not engaged in direct military intervention. The Vietnam War also led to widespread anti-war protests and social unrest in the United States, which could have been mitigated if the U.S. had not engaged in direct military intervention. Furthermore, the U.S. military's involvement in Vietnam led to a loss of credibility and trust in the U.S. government, both domestically and internationally, which could have been avoided with a containment policy.

Body Paragraph 3: The third supporting argument is that a more stable Southeast Asia would have emerged as a result of a containment policy, fostering stronger relationships between the U.S. and its regional allies. Containment would have encouraged regional cooperation among non-communist countries, leading to the formation of alliances and partnerships that could counterbalance the spread of communism. By avoiding direct military intervention, the U.S. would have been able to maintain a more positive image in the eyes of Southeast Asian nations, fostering goodwill and cooperation rather than resentment and anti-American sentiment. Moreover,

by focusing on containment, the U.S. could have worked to address the root causes of communism's appeal in the region, such as poverty and political instability, rather than simply trying to suppress it through military force.

Conclusion: In conclusion, this essay has argued that if the U.S. had pursued a policy of containment instead of direct military intervention during the Vietnam War, the spread of communism in Southeast Asia would have been slowed, ultimately leading to a more stable and less costly outcome for the United States and its allies. A containment policy would have allowed the U.S. to focus on diplomatic and economic efforts, reduced the human and financial costs of the war, and fostered a more stable Southeast Asia. While it is impossible to know for certain how history would have unfolded under different circumstances, this counterfactual analysis provides valuable insights into the potential consequences and implications of alternative policy choices.

Appendix B: Complete Causal Model

1. Causal Inference Statement
    a. natural language
        i. Prompt: “Your friend says that if Lebron James wins Finals MVP, then the LA Lakers are likely to win the Western Conference Semifinals. Is this statement accurate according to the causal model?”
        ii. Response: “This statement is not accurate according to the causal model. The Finals MVP (MVP) is determined after the Finals Champion (FC) is decided, which occurs after the Semifinal Champion (SC) is determined. The Western Conference Semifinals is a stage before the Finals, so the outcome of the Finals MVP cannot have a causal influence on the outcome of the Western Conference Semifinals.”
    b. do-calculus
        i. Prompt: “Lebron(MVP=1) → LA(SC=1). Is this statement accurate according to the causal model?”
        ii. Response: “No, this statement is not accurate according to the causal model. The causal model does not have a direct causal link between a specific player winning the MVP and their team becoming the Semifinal Champion. The model only shows a link between Finals Champion (FC) and Most Valuable Player (MVP).”
    c. Which one is better? Natural Language!
        i. Both responses correctly identified that the statement is inaccurate, but the natural language prompt produced a response that better identified the flaw in the statement’s causal reasoning.
2. A Slightly Flawed Counterfactual Statement
    a. Natural language
        i. Prompt: “Denver was not injured and won the semifinals championship over LA, so injuring Denver would have caused LA to win the semifinals championship instead. Is this counterfactual statement accurate according to the causal model?”
        ii. Response: “According to the causal model, injuring Denver (Team(Injured=1)) would decrease Denver's chance of victory (Injured(I) → Victory(V)) and reduce their skill advantage (Injured(I) → Skill(SA)). However, it does not guarantee that LA would win the semifinals championship instead. The counterfactual statement is not accurate, as it overstates the certainty of the outcome. A more accurate counterfactual statement would be: "Injuring Denver might have increased LA's chances of winning the semifinals championship."”

b. Do-calculus
    i. Prompt: “Observation: Denver(Injured=0), Denver(SC=1). Intervention: Do(Denver(Injured=1)). Counterfactual: LA(Victory=1). Is this counterfactual statement accurate according to the causal model?”
    ii. Response: “Yes, this counterfactual statement is accurate according to the causal model. Observation: Denver is not injured (Injured=0) and is the Semifinal Champion (SC=1). Intervention: If Denver were to be injured (Do(Denver(Injured=1)). According to the causal model, an injury (Injured=1) would decrease Denver's skill advantage (SA) and their chance of victory (V). Since LA Lakers are Denver's opponents in the Western Semifinals, if Denver's chance of victory decreases due to the injury, it would increase the LA Lakers' chance of victory. Therefore, the intervention leads to LA(Victory=1).”
c. Which one is better? Natural language!
    i. The natural language response correctly identifies the flaw in the statement that the influence of injury on victory is relative, not absolute.

3. A Clearly Flawed Counterfactual Statement
a. Natural language
    i. Prompt: “Given that the Miami Heat won the semifinals championship, the Heat would have not been injured if they'd had home court advantage. Is this counterfactual statement accurate according to the causal model?”
    ii. Response: “The counterfactual statement is not accurate according to the causal model. The causal model states that having home court advantage (HCS) increases a team's skill advantage (SA), but it does not directly influence whether a team gets injured (I). Therefore, we cannot conclude that the Miami Heat would not have been injured if they had home court advantage.”
b. Do calculus
    i. Prompt: “Observation: Heat(SC=1). Intervention: Do(Heat(HCS=1)). Counterfactual: Heat(Injured=0). Is this counterfactual statement accurate according to the causal model?”
    ii. Response: “No, this counterfactual statement is not accurate according to the causal model. The intervention Do(Heat(HCS=1)) refers to giving the Miami Heat home court advantage in the semifinals. However, the intervention does not have a direct causal effect on the Heat's injury status. The causal model does not show a direct link between Home Court Advantage in the Semifinals (HCS) and Injured (I).”

c. Which one is better? Undecided
    i. Both responses correctly identify the statement's inaccurate leap in causal reasoning

4. Another Slightly Flawed Statement
    a. Natural language
        i. Prompt: "Given that the Heat became semifinals champions without semifinals home court advantage but did not become finals champions, the heat would have become finals champions had they had semifinals home court advantage. Is this counterfactual statement accurate according to the causal model?"
        ii. Response: "According to the causal model, having home court advantage in the semifinals (HCS) increases a team's skill advantage (SA), which in turn positively influences their chance of victory (V) over their current opponent. However, the counterfactual statement claims that if the Heat had semifinals home court advantage, they would have become finals champions. While having home court advantage in the semifinals might have increased the Heat's skill advantage and chances of victory in the semifinals, it does not guarantee that they would have become finals champions. Other factors, such as injuries, skill advantage, and home court advantage in the finals, could also influence the outcome of the finals. Therefore, the counterfactual statement is not accurate according to the causal model, as it does not account for all the factors that could influence the outcome of the finals."
    b. Do-calculus
        i. Prompt: "Observation: Heat(HCS=0), Heat(SC=1), Heat(FC=0). Intervention: do(Heat(HCS=1)). Counterfactual: Heat(FC=1). Is this counterfactual statement accurate according to the causal model?"
        ii. Response: "Yes, this counterfactual statement is accurate according to the causal model. The observation states that the Miami Heat did not have home court advantage in the Semifinals (Heat(HCS=0)), but they still became the Semifinal Champion (Heat(SC=1)). However, they did not become the Finals Champion (Heat(FC=0)). The intervention is to change the home court advantage in the Semifinals for the Heat (do(Heat(HCS=1))). According to the causal model, having home court advantage increases a team's skill advantage (Home Court Advantage (HC, HCS, HCF) → Skill Advantage (SA)), which in turn positively influences their chance of victory (Skill Advantage (SA) → Victory (V)). The counterfactual statement is that if the Heat had home court advantage in the Semifinals

(Heat(HCS=1)), they would have become the Finals Champion (Heat(FC=1)). Given the causal model, this counterfactual statement is plausible, as having home court advantage could have increased their skill advantage and chances of winning games, potentially leading to them becoming the Finals Champion."

c. Which one is better? Natural language
    i. The natural language response correctly identifies how having home court advantage only helps a team's chance of victory against its current opponent, and not against future opponents. The do-calculus response does not identify the temporally limited influence of home court advantage on victory.

Appendix C: Mathematical Counterfactuals

C.1: Part 1 - Change

```
Problem 1: "Given 2x - 4y + 5z = -1 with the current values of (x, y, z),
what would the new result be after decrementing x by 1, incrementing y by
2, and decrementing z by 2?" The correct answer -21.

Response I/O:
'''The new result is -15'''
INCORRECT

Response CoT:
'''We can write our new result as r = 2(x-1) - 4(y+2) + 5(z-2) = -1 + d
Let's solve term-by-term:
    2(x-1) = 2x - 2 = 2x + d_1, so d_1 = -2
    -4(y+2) = -4y - 8 = -4y + d_2, so d_2 = -8
    5(z-2) = 5z - 10 = 5z + d_3, so d_3 = -10
    So, d = d_1 + d_2 + d_3 = -2 - 8 - 10 = -20.
    Plugging this into our equation for r, we have that r = -1 - 20 = -21.
The new result is -21'''
CORRECT


Problem 2: "Given -8x - 1y + 3z = -7 with the current values of (x, y, z),
what would the new result be after decrementing x by 9, incrementing y by
6, and incrementing z by 5?" The correct answer 74.

Response I/O:
'''The new result is -79'''
INCORRECT

Response CoT:
'''We can write the new equation as -8(x-9) - 1(y+6) + 3(z+5) = -7 + d
Let's solve term-by-term:
    -8(x-9) = -8x + 72 = -8x + d_1, so d_1 = 72
    -1(y+6) = -y - 6 = -y + d_2, so d_2 = -6
    3(z+5) = 3z + 15 = 3z + d_3, so d_3 = 15
    So, d = d_1 + d_2 + d_3 = 72 - 6 + 15 = 81.
    Plugging this into our equation for r, we have that r = -7 + 81 = 74.
The new result is 74'''
CORRECT


Problem 3: "Given 7x - 9y + 7z = -1 with the current values of (x, y, z),
what would the new result be after incrementing x by 0, decrementing y by
4, and decrementing z by 3?" The correct answer 14.
```

Response I/O:
'''The new result is 35'''
INCORRECT

Response CoT:
'''We can write our new result as r = 7(x+0) - 9(y-4) + 7(z-3) = -1 + d
Let's solve term-by-term:
    7(x+0) = 7x = 7x + d_1, so d_1 = 0
    -9(y-4) = -9y + 36 = -9y + d_2, so d_2 = 36
    7(z-3) = 7z - 21 = 7z + d_3, so d_3 = -21
    So, d = d_1 + d_2 + d_3 = 0 + 36 - 21 = 15.
    Plugging this into our equation for r, we have that r = -1 + 15 = 14.
The new result is 14'''
CORRECT

Problem 4: "Given 1x + 6y + 3z = -3 with the current values of (x, y, z), what would the new result be after incrementing x by 8, decrementing y by 1, and incrementing z by 9?" The correct answer 26.

Response I/O:
'''The new result is 62'''
INCORRECT

Response CoT:
'''We can write our new result as r = 1(x+8) + 6(y-1) + 3(z+9) = -3 + d
Let's solve term-by-term:
    1(x+8) = x + 8 = x + d_1, so d_1 = 8
    6(y-1) = 6y - 6 = 6y + d_2, so d_2 = -6
    3(z+9) = 3z + 27 = 3z + d_3, so d_3 = 27
    So, d = d_1 + d_2 + d_3 = 8 - 6 + 27 = 29.
    Plugging this into our equation for r, we have that r = -3 + 29 = 26.
The new result is 26'''
CORRECT

Problem 5: "Given 3x + 7y - 8z = -6 with the current values of (x, y, z), what would the new result be after decrementing x by 3, incrementing y by 0, and decrementing z by 8?" The correct answer 49.

Response I/O:
'''The new result is -15'''
INCORRECT

Response CoT:
'''We can write our new result as r = 3(x-3) + 7(y+0) - 8(z-8) = -6 + d

Let's solve term-by-term:
    3(x-3) = 3x - 9 = 3x + d_1, so d_1 = -9
    7(y+0) = 7y = 7y + d_2, so d_2 = 0
    -8(z-8) = -8z + 64 = -8z + d_3, so d_3 = 64
    So, d = d_1 + d_2 + d_3 = -9 + 0 + 64 = 55.
    Plugging this into our equation for r, we have that r = -6 + 55 = 49.
The new result is 49'''
CORRECT

Problem 6: "Given 3x + 8y + 3z = -3 with the current values of (x, y, z), what would the new result be after incrementing x by 2, decrementing y by 9, and decrementing z by 1?" The correct answer -72.

Response I/O:
'''The new result is -84'''
INCORRECT

Response CoT:
'''We can write our new result as r = 3(x+2) + 8(y-9) + 3(z-1) = -3 + d
Let's solve term-by-term:
    3(x+2) = 3x + 6 = 3x + d_1, so d_1 = 6
    8(y-9) = 8y - 72 = 8y + d_2, so d_2 = -72
    3(z-1) = 3z - 3 = 3z + d_3, so d_3 = -3
    So, d = d_1 + d_2 + d_3 = 6 - 72 - 3 = -69.
    Plugging this into our equation for r, we have that r = -3 - 69 = -72.
The new result is -72'''
CORRECT

Problem 7: "Given 2x - 6y - 9z = -5 with the current values of (x, y, z), what would the new result be after decrementing x by 5, decrementing y by 8, and decrementing z by 6?" The correct answer 87.

Response I/O:
'''The new result is 61'''
INCORRECT

Response CoT:
'''We can write our new result as r = 2(x-5) - 6(y-8) - 9(z-6) = -5 + d
Let's solve term-by-term:
    2(x-5) = 2x - 10 = 2x + d_1, so d_1 = -10
    -6(y-8) = -6y + 48 = -6y + d_2, so d_2 = 48
    -9(z-6) = -9z + 54 = -9z + d_3, so d_3 = 54
    So, d = d_1 + d_2 + d_3 = -10 + 48 + 54 = 92.
    Plugging this into our equation for r, we have that r = -5 + 92 = 87.
The new result is 87'''

CORRECT

Problem 8: "Given 9x + 6y - 2z = 0 with the current values of (x, y, z), what would the new result be after incrementing x by 7, incrementing y by 5, and decrementing z by 4?" The correct answer 101.

Response I/O:
'''The new result is 99'''
INCORRECT

Response CoT:
'''We can write our new result as r = 9(x+7) + 6(y+5) - 2(z-4) = 0 + d
Let's solve term-by-term:
    9(x+7) = 9x + 63 = 9x + d_1, so d_1 = 63
    6(y+5) = 6y + 30 = 6y + d_2, so d_2 = 30
    -2(z-4) = -2z + 8 = -2z + d_3, so d_3 = 8
    So, d = d_1 + d_2 + d_3 = 63 + 30 + 8 = 101.
    Plugging this into our equation for r, we have that r = 0 + 101 = 101.
The new result is 101'''
CORRECT

Problem 9: "Given -7x + 3y - 7z = 9 with the current values of (x, y, z), what would the new result be after incrementing x by 2, decrementing y by 1, and decrementing z by 2?" The correct answer 6.

Response I/O:
'''The new result is 2'''
INCORRECT

Response CoT:
'''We can write our new result as r = -7(x+2) + 3(y-1) - 7(z-2) = 9 + d
Let's solve term-by-term:
    -7(x+2) = -7x - 14 = -7x + d_1, so d_1 = -14
    3(y-1) = 3y - 3 = 3y + d_2, so d_2 = -3
    -7(z-2) = -7z + 14 = -7z + d_3, so d_3 = 14
    So, d = d_1 + d_2 + d_3 = -14 - 3 + 14 = -3.
    Plugging this into our equation for r, we have that r = 9 - 3 = 6.
The new result is 6'''
CORRECT

Problem 10: "Given 2x - 9y - 3z = -5 with the current values of (x, y, z), what would the new result be after decrementing x by 7, decrementing y by 4, and decrementing z by 5?" The correct answer 32.

Response I/O:
'''The new result is 49.'''
INCORRECT

Response CoT:
'''We can write our new result as r = 2(x-7) - 9(y-4) - 3(z-5) = -5 + d
Let's solve term-by-term:
    2(x-7) = 2x - 14 = 2x + d_1, so d_1 = -14
    -9(y-4) = -9y + 36 = -9y + d_2, so d_2 = 36
    -3(z-5) = -3z + 15 = -3z + d_3, so d_3 = 15
    So, d = d_1 + d_2 + d_3 = -14 + 36 + 15 = 37.
    Plugging this into our equation for r, we have that r = -5 + 37 = 32.
The new result is 32'''
CORRECT

Problem 11: "Given -8x + 1y + 2z = -8 with the current values of (x, y, z), what would the new result be after incrementing x by 9, incrementing y by 8, and decrementing z by 1?" The correct answer -74.

Response I/O:
'''The new result is 65'''
INCORRECT

Response CoT:
'''We can write our new result as r = -8(x+9) + 1(y+8) + 2(z-1) = -8 + d
Let's solve term-by-term:
    -8(x+9) = -8x - 72 = -8x + d_1, so d_1 = -72
    1(y+8) = y + 8 = y + d_2, so d_2 = 8
    2(z-1) = 2z - 2 = 2z + d_3, so d_3 = -2
    So, d = d_1 + d_2 + d_3 = -72 + 8 - 2 = -66.
    Plugging this into our equation for r, we have that r = -8 - 66 = -74.
The new result is -74.'''
CORRECT

Problem 12: "Given -3x - 3y + 8z = 5 with the current values of (x, y, z), what would the new result be after decrementing x by 8, decrementing y by 3, and incrementing z by 9?" The correct answer 110.

Response I/O:
'''The new result is 104.'''
INCORRECT

Response CoT:
'''We can write our new result as r = -3(x-8) - 3(y-3) + 8(z+9) = 5 + d
Let's solve term-by-term:

-3(x-8) = -3x + 24 = -3x + d_1, so d_1 = 24
-3(y-3) = -3y + 9 = -3y + d_2, so d_2 = 9
8(z+9) = 8z + 72 = 8z + d_3, so d_3 = 72
So, d = d_1 + d_2 + d_3 = 24 + 9 + 72 = 105.
Plugging this into our equation for r, we have that r = 5 + 105 = 110.
The new result is 110'''
CORRECT

Problem 13: "Given -1x - 8y - 7z = -3 with the current values of (x, y, z), what would the new result be after decrementing x by 7, incrementing y by 1, and incrementing z by 1?" The correct answer -11.

Response I/O:
'''The new result is -20'''
INCORRECT

Response CoT:
'''We can write the new equation as -1(x-7) - 8(y+1) - 7(z+1) = -3 + d
Let's solve term-by-term:
-1(x-7) = -x + 7 = -x + d_1, so d_1 = 7
-8(y+1) = -8y - 8 = -8y + d_2, so d_2 = -8
-7(z+1) = -7z - 7 = -7z + d_3, so d_3 = -7
So, d = d_1 + d_2 + d_3 = 7 - 8 - 7 = -8.
Plugging this into our equation for r, we have that r = -3 - 8 = -11.
The new result is -11.'''
CORRECT

Problem 14: "Given -2x + 3y + 6z = -9 with the current values of (x, y, z), what would the new result be after incrementing x by 5, decrementing y by 5, and decrementing z by 5?" The correct answer -64.

Response I/O:
'''The new result is 32'''
INCORRECT

Response CoT:
'''We can write our new result as r = -2(x+5) + 3(y-5) + 6(z-5) = -9 + d
Let's solve term-by-term:
-2(x+5) = -2x - 10 = -2x + d_1, so d_1 = -10
3(y-5) = 3y - 15 = 3y + d_2, so d_2 = -15
6(z-5) = 6z - 30 = 6z + d_3, so d_3 = -30
So, d = d_1 + d_2 + d_3 = -10 - 15 - 30 = -55.
Plugging this into our equation for r, we have that r = -9 - 55 = -64.
The new result is -64.'''
CORRECT

Problem 15: "Given -6x - 4y - 9z = -6 with the current values of (x, y, z), what would the new result be after decrementing x by 5, incrementing y by 9, and incrementing z by 5?" The correct answer -57.

Response I/O:
'''The new result is -51'''
INCORRECT

Response CoT:
'''We can write the new equation as -6(x-5) - 4(y+9) - 9(z+5) = -6 + d
Let's solve term-by-term:
    -6(x-5) = -6x + 30 = -6x + d_1, so d_1 = 30
    -4(y+9) = -4y - 36 = -4y + d_2, so d_2 = -36
    -9(z+5) = -9z - 45 = -9z + d_3, so d_3 = -45
    So, d = d_1 + d_2 + d_3 = 30 - 36 - 45 = -51.
    Plugging this into our equation for r, we have that r = -6 - 51 = -57.
The new result is -57.'''
CORRECT

I/O Accuracy: 0/15 = 0.0
CoT Accuracy: 15/15 = 1.0

C.2: Part 2 - Solve

Due to length considerations, only the prompt problems are provided, since the full text results are 30 plus pages long. This was performed using ChatGPT-4 so there is no langchain code.

I/O Prompts

Problem 1:

Given that x=2, y=-1, z=3 is a solution to 2x + 3y - 4z = -11, find the most parsimonious integer-valued assignments for x, y, z such that the equation now evaluates to 1. Define a parsimonious solution as one that assigns new values to as few of the variables x, y, z as possible. If there are multiple solutions that change the same number of variables, select the solution that changes these variables by the smallest absolute amount. Use the Wolfram plugin to perform math calculations when necessary.

Problem 2:

Given that x=-3, y=-7, z=1 is a solution to the equation 5x - 8y - 16z = 25, find the most parsimonious integer-valued assignments for x, y, z such that the equation now evaluates to 9. Define a parsimonious solution as one that assigns new values to as few of the variables x, y, z as possible. If there are multiple solutions that change the same number of variables, select the solution that changes these variables by the smallest absolute amount. Use the Wolfram plugin to perform math calculations when necessary.

Problem 3:

Given that x=4, y=1, z=-2 is a solution to the equation -3x + 2y - 6z = 2, find the most parsimonious integer-valued assignments for x, y, z such that the equation now evaluates to -3. Define a parsimonious solution as one that assigns new values to as few of the variables x, y, z as possible. If there are multiple solutions that change the same number of variables, select the solution that changes these variables by the smallest absolute amount. Use the Wolfram plugin to perform math calculations when necessary.

ToT Prompts

Problem 1:

Given that x=2, y=-1, z=3 is a solution to 2x + 3y - 4z = -11, find the most parsimonious integer-valued assignments for x, y, z such that the equation now evaluates to 1. Define a parsimonious solution as one that assigns new values to as few of the variables x, y, z as possible. If there are multiple solutions that change the same number of variables, select the solution that changes these variables by the smallest absolute amount. Solve using Tree of Thought, as defined in the paper https://arxiv.org/pdf/2305.10601.pdf. Use the Wolfram plugin to perform math calculations when necessary.

Problem 2:

Given that x=-3, y=-7, z=1 is a solution to the equation 5x - 8y - 16z = 25, find the most parsimonious integer-valued assignments for x, y, z such that the equation now evaluates to 9. Define a parsimonious solution as one that assigns new values to as few of the variables x, y, z as possible. If there are multiple solutions that change the same number of variables, select the solution that changes these variables by the smallest absolute amount. Solve using Tree of Thought, as defined in the paper https://arxiv.org/pdf/2305.10601.pdf. Use the Wolfram plugin to perform math calculations when necessary.

Problem 3:

Given that x=4, y=1, z=-2 is a solution to the equation -3x + 2y - 6z = 2, find the most parsimonious integer-valued assignments for x, y, z such that the equation now evaluates to -3. Define a parsimonious solution as one that assigns new values to as few of the variables x, y, z as possible. If there are multiple solutions that change the same number of variables, select the solution that changes these variables by the smallest absolute amount. Solve using Tree of Thought, as defined in the paper https://arxiv.org/pdf/2305.10601.pdf. Use the Wolfram plugin to perform math calculations when necessary.