Article

Causal Agent based on Large Language Model

Kairong Han [1], Kun Kuang [1,*], Ziyu Zhao [1], Junjian Ye [2] and Fei Wu [1,*]

[1] College of Computer Science and Technology, Zhejiang University, Zhejiang 310058; zju_handso@163.com (K.H.); kunkuang@zju.edu.cn (K.K.); benzhao.styx@gmail.com (Z.Z.); wufei@zju.edu.cn (F.W.)
[2] Huawei Technologies Co., Ltd; yejunjian@huawei.com
* Correspondence: wufei@zju.edu.cn; Tel.: +86-0571-87951160(F.W.); kunkuang@zju.edu.cn (K.K.)

arXiv:2408.06849v1 [cs.AI] 13 Aug 2024

Abstract: Large language models (LLMs) have achieved significant success across various domains. However, the inherent complexity of causal problems and causal theory poses challenges in accurately describing them in natural language, making it difficult for LLMs to comprehend and use them effectively. Causal methods are not easily conveyed through natural language, which hinders LLMs' ability to apply them accurately. Additionally, causal datasets are typically tabular, while LLMs excel in handling natural language data, creating a structural mismatch that impedes effective reasoning with tabular data. This lack of causal reasoning capability limits the development of LLMs. To address these challenges, we have equipped the LLM with causal tools within an agent framework, named the Causal Agent, enabling it to tackle causal problems. The causal agent comprises tools, memory, and reasoning modules. In the tools module, the causal agent applies causal methods to align tabular data with natural language. In the reasoning module, the causal agent employs the ReAct framework to perform reasoning through multiple iterations with the tools. In the memory module, the causal agent maintains a dictionary instance where the keys are unique names and the values are causal graphs. To verify the causal ability of the causal agent, we established a benchmark consisting of four levels of causal problems: variable level, edge level, causal graph level, and causal effect level. We generated a test dataset of 1.3K using ChatGPT-3.5 for these four levels of issues and tested the causal agent on the datasets. Our methodology demonstrates remarkable efficacy on the four-level causal problems, with accuracy rates all above 80%. For further insights and implementation details, our code is accessible via the GitHub repository https://github.com/Kairong-Han/Causal_Agent.

Keywords: Causal Inference; Causal Discovery; Large Language Model; Agent;

Citation: Han, K.; Kuang, K.; Zhao, Z.; Ye, J.; Wu, F. Causal Agent based on Large Language Model. *Entropy* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

Copyright: © 2024 by the authors. Submitted to *Entropy* for possible open access publication under the terms and conditions of the Creative Commons Attri- bution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, generative artificial intelligence technology has gained significant success, achieving remarkable behavior in the natural language processing field [1], image, audio synthesis, etc [84]. This advancement lays the foundation for propelling research in general artificial intelligence [85], both in terms of framework development and practical implementation. However, due to the complexity of causal problems, the causal reasoning capabilities of the LLM remain insufficient. Causal theory is difficult to describe in natural language that the LLM can understand accurately. Researchers have evaluated the pure causal reasoning abilities of the LLM and found that their pure causal reasoning is close to random [16]. Additionally, researchers believe that the current LLM are merely "causal parrots" that mimic without truly possessing causal understanding [19]. This inherent limitation severely hampers the performance of large models in tasks requiring causal reasoning. Moreover, causal datasets are typically tabular data, while large models excel in handling natural language data. When we need to draw causal conclusions based on the analysis of tabular data, LLMs that are not specifically designed cannot directly utilize tabular data and perform reasoning. This structural heterogeneity hinders LLM from effectively reasoning with tabular data. These two limitations restrict the ability of LLMs to solve causal problems effectively.

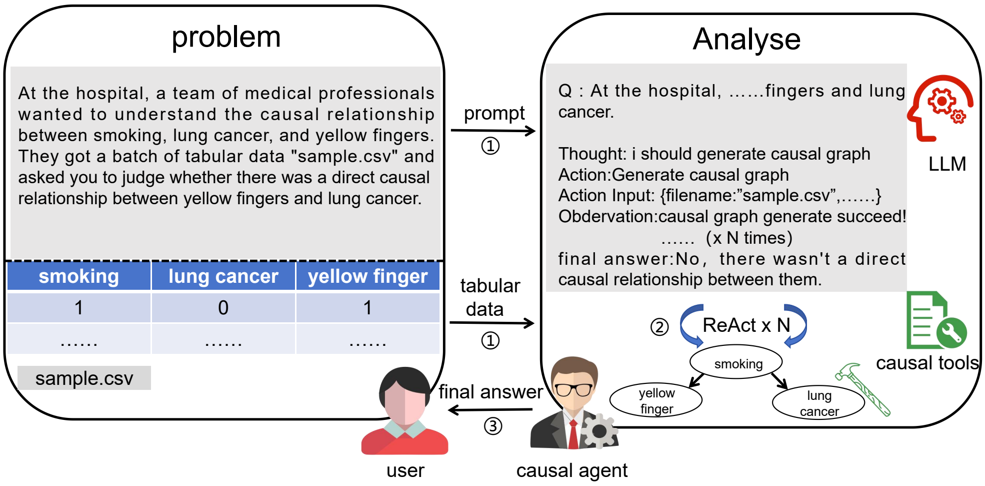

Figure 1. Working flowchart of the causal agent. In the first step, the user inputs a pair of the tabular data and the causal problem; In the second step, the causal agent invokes the causal tools (tools module) and uses the ReAct framework (plan module) to conduct multiple rounds of analysis for the tabular data, in which causal agent maintains a dictionary of causal graph names and their instantiations as memory (memory module); In the third step, the causal agent is combined with the analysis process to produce corresponding answers for the user's problems.

Several studies have emerged recently and attempted to use LLM as the core decision-making unit of intelligent agents and combine them with external tools to interact with the environment, thus achieving remarkable results in solving problems in specific fields. This type of technical strategy, which combines LLM with domain expertise, effectively improves the ability of LLM to solve problems by using external tools to empower agents in specific fields [22]. For example, ChatDev [29] proposes an end-to-end framework, where multiple agent roles communicate and collaborate through natural language conversations to complete the software development life cycle. ToolBench [38]can be used for tasks such as code auto-completion and code recommendation. However, due to the lack of relevant tools and capabilities, it is difficult for agents to solve causal problems directly.

To solve two difficult problems in sweetening the causal ability of LLM, we propose a causal problem modeling approach from the perspective of the LLM and propose a causal agent framework by guiding LLM to invoke causal tools. We model the causal problems into four levels: variable level, edge level, causal graph level, and causal effect level. The variable level focuses on the agent's judgment and understanding of correlations, the edge level focuses on the agent's examination of causal relationships between variables, the causal graph level focuses on the agent's ability to generate causal graphs, and the causal effect level focuses on the agent's estimation of causal effects between variables for quantitative expression. Based on the abovementioned causal problems, we construct a causal agent based on LLM, using ChatGPT-3.5. The causal agent is composed of tools, memory, and plan modules, as shown in Figure 1. In the tools module, the causal agent invokes the causal analysis library in Python programming tools, such as causal-learn [51] and EconML [81]. So the causal agent can receive a pair of tabular data and a causal problem description of the data as input. By invoking causal analysis tools, the tool processes the tabular data and generates natural language conclusions that the causal agent can understand. In the plan module, the causal agent utilizes its text comprehension and reasoning abilities to obtain answers to causal problems in many times iterations. In the memory module, the causal agent may need multiple tools to solve a problem. To preserve intermediate results during the planning process, the agent maintains an instantiated dictionary where the keys are names and the values are causal graphs. This special method allows the agent to retrieve the necessary causal graph using the key. On the one hand, the content of the memory is expressed more richly; on the other hand, using a data structure rather than text as memory can effectively simplify the complexity of prompt design during the reasoning process. In this way, the causal agent achieved high accuracy in answering causal problems at four level questions, with accuracy rates of over 92% in all three sub-problems for determining correlation at the variable level, over 89% in all three sub-problems at the edge level, over 81% in the causal graph level, and 93% in the causal effect estimation level.

This work contributions are summarized as follows:

- A hierarchical modeling perspective has been proposed for LLM to solve causal problems. This is a new setting and the problem is set to be data-driven, where the LLM answers causal questions about tabular data when users input a pair of tabular data and causal questions. We focus on four level questions for causal agents to solve causal problems, denoted as variable level, edge level, causal graph level, and causal effect level. We propose a test set of approximately 1.3K in size for the four levels of problems, covering nine sub-problems in total at four levels;
- The causal agent has been proposed to empower LLM with the ability to solve causal problems. In this framework, we use LLM to invoke causal tools and iterate many times to analyze and solve causal problems. Thus achieving heterogeneous data alignment between natural language input for large models and tabular data input for causal problems. The causal agent framework that empowers causal reasoning through the use of causal tools has good Interpretability and reliability;
- The causal agent achieved high accuracy in the four levels of causal problems modeled in this article. Specifically, all three sub-problems at the variable level achieved an accuracy of over 92%, all three sub-problems at the edge level achieved an accuracy of over 89%, all three sub-problems at the causal graph level achieved an accuracy of over 81%, and all two sub-problems at the causal effect level achieved an accuracy of over 93%;

2. Related Work

2.1. Causality

Causality, as a tool for data analysis, aims to accurately identify and quantify the actual effects of specific factors (causes) on outcome variables (effects) in a complex system environment [30]. It is everywhere in our daily lives. Such as statistics [3–5], economics [5], computer science [6,7], epidemiology [8,9] and psychology [10]. Different from correlation, causality explores in depth the changing pattern of how the result variable responds when the cause variable changes. Therefore, the "Ladder of Causality" theory proposed by Pearl divides causality into three progressive levels [11]: association, intervention, and counterfactual. The association focuses on discovering the correlation between variables through observation of data. However, this can only reveal the accompanying phenomena between events, and cannot indicate the causal flow between events; Intervention emphasizes when we actively change the state of an event, whether and how other related events will change accordingly; Counterfactual imagines how the current observed results would have changed if there had not been an event that had occurred. The core purpose of studying causality is to reveal the true causal chain between things and to abandon those confusing pseudo-causal relationships. Cause field problems can be briefly divided into two broad directions: causal discovery and causal inference. Causal discovery is based on directed acyclic graphs and Bayesian models, focusing on obtaining causal relationships from observation data, and methods can be divided into constraint-based methods [65–67], such as IC, PC, FCI, and function-based methods [68–70] such as LiNGAM and ANM, and hybrid methods [58] to combine the advantages of the above two methods. Common frameworks for causal inference are structural causal model[13] and potential outcome framework [12]. The potential outcome framework is also known as the Neyman–Rubin

Potential Outcomes or the Rubin Causal Model. Researchers use the structural causal model and Rubin Causal Model to model the interaction between variables, and calculate causal effect estimates such as average treatment effect (ATE) and conditional treatment effect (CATE).

2.2. LLM-based agent

Autonomous agents have long been considered a promising approach to achieving artificial general intelligence (AGI), which accomplishes tasks through autonomous planning and action [73]. In previous studies, simple and heuristic policy functions were designed for agents to learn in isolated and constrained environments [74,75]. In recent years, LLM has achieved great success in the field of natural language. Human-like intelligence has shown great potential [2,35,36] and there has been a large amount of research using LLM as the decision-making and reasoning center of agents [37–39], achieving great success in natural sciences [40,41], engineering sciences [29,42,43], and human simulation [27,44]. The LLM agent is composed of four parts, namely profile module, action module, plan module, and action module. The identity module assigns an imaginary role to the agent, such as a teacher or poet. According to different text sources. The planning module helps the agent use thinking chains to break down tasks and use different search methods to obtain solutions in the problem space, such as CoT [32], ToT [33], AoT [34], Reflexion [35], etc. The memory module is subdivided into two categories: short-term memory and long-term memory. Its specific implementation forms are diverse, depending on the data structure and technical means used. The action module is the key for the intelligent agent to take specific actions in the physical or virtual environment. The agent implements actions by using tools to change the environmental state and task process and also triggers changes in its state.

2.3. Combining LLM and causality

Since the advent of LLM, some researchers have evaluated and analyzed the causality ability of LLM. Jin et al. [16] introduced a new task CORR2CAUSE, which can infer causal relationships from correlations, to evaluate the causal inference ability of large models. This task first constructs a causal graph based on the original data and then converts it into natural language by the D-separation principle. From the experimental results, it is generally believed that the LLM with a higher version or better reasoning ability does not show positive correlation results in the causal inference task, and the performance of the LLM in the causal inference task is akin to random. Jin et al. [18] further investigated whether large language models can reason about causality and proposed a new NLP task: causal inference in natural language. Inspired by the "causal inference engine" and hypotheses proposed by Judea Pearl. They built a large dataset, CLADDER, with 10K samples: a collection (association, intervention, and counterfactual) based on causal graphs and queries. In addition, they introduce and evaluate a customized chain of thought prompting strategy CausalCOT. Gao et al. [17] presents a comprehensive evaluation of ChatGPT's causal reasoning capabilities. They found that ChatGPT is not a good causal reasoner, but is good at causal explanation and that ChatGPT has a serious problem of causal illusion, which is further exacerbated by In-Context Learning (ICL) and chain of thought techniques. Zečević et al. [19] argues that large language models cannot be causal and define a new subgroup of structure causal models, called meta-SCMs. Their empirical analysis provides favorable evidence that current LLMs are even weak "causal parrots". Long et al. [46] further investigated how imperfect expert knowledge can be used to improve the output of causal discovery algorithms. A greedy algorithm is also proposed to iteratively reject graphs from MEC while controlling the probability of excluding true graphs. They found a reduction in performance when using large models as experts. Nonetheless, their results still suggest a clear potential for LLM to help discover causal relationships. Kıcıman et al. [47] found LLM can achieve competitive performance in determining pairwise causality, with an accuracy of up to 97%, but their performance varies depending on the quality of cue word engineering. Richens and Everitt [48] provides a theoretical analysis of whether agents must learn causal models to generalize to new domains. They analytically show that any agent capable of satisfying regret bounds under a large number of distribution shifts must learn an approximate causal model of the data-generating process, and discuss the implications of this result for generative AI including transfer learning and causal inference. Nichani et al. [49] explored the causal structure learned by Transformers in attention matrices using backpropagation learning. They found that when the underlying causal graph is a tree, Gradient descent on the simplified two-layer Transformer solves this task by encoding the causal map at the first attention layer for contextual estimation of the transition distribution.

How to enhance the causal reasoning ability of large language models has become a difficult problem in current research. Solving the shortage of large language models in causal ability has great potential for solving large model illusions and promoting the development of trusted AI.

3. Materials and Methods

3.1. Modeling causal problems from the perspective of LLM

Despite the development of LLM, like ChatGPT, demonstrating strong natural language understanding and question-answering capabilities, there is still a significant gap between the problem paradigms that data-driven causal focuses on tabular data but LLM focuses on the field of natural language processing. Furthermore, LLMs struggle to truly understand and handle the intricate causal relationships inherent in complex data. The inability of LLMs to adapt to causal tasks remains a significant limitation in their development and widespread reliable use.

Therefore, it is meaningful to re-establish a causal problem framework from the perspective of the LLM. This has a significant impact on evaluating the causal ability of LLMs and enhancing their causal ability. To model causal problems within the field of LLM, We formulate our settings as follows:

Let $T \in R^{n\times c}$be a tabular data with n rows and c columns, in which each row t_i is a piece of data, and each column c_i represents a variable. So

$$T = \{t_i^c\}_{i=0}^n$$

We formalize the causal problem space as a Q and $q_i \in Q$ is one question in the form of natural language. We combine the tabular data and the problem description by Cartesian product to create the dataset D and each item $d_i \in R^{n\times c} \times Q$.So

$$D = \{d_i\} = \{(T_i, q_i) \in R^{n\times c} \times Q\}$$

The user inputs a pair of (T_i, q_i) samples from D, and then the causal agent analyses the tabular data T_i and the causal problem Q to generate a reasonable answer A. The format of answer A is not limited to the form of natural language. A can also be a causal graph or other heterogeneous non-textual data to explain the question clearly.

Table 1. Summary of the causal problem at four levels

Name	Explanation
Variable level	The ability to infer and understand correlations between variables in tabular data
Edge level	The ability to understand and analyze causal edge relationships between variables
Causal graph level	The ability to generate causal graphs
Causal effect level	The ability to quantify the causal effects between variables

Due to the complex diversity of causal problems, we simplify the problem and conduct the necessary modeling. We categorize the causal problems into four major levels, as shown in Table 1, based on the differences in problem granularity and objects: variable level, edge level, causal graph level, and causal effect level. The variable level corresponds to the first level of the causal ladder, correlation, aiming to endow LLM with the ability to infer and understand correlations between variables in tabular data. The edge level builds beyond correlation, aiming to endow LLM with the ability to understand and analyze causal edge relationships between variables. The causal graph level shifts the perspective to a more macroscopic dimension, examining the LLM's capabilities of generating causal graphs. The causal effect level aims to endow LLM with the ability to quantify the causal effects between variables. We will discuss four levels of modeling methods in detail below.

3.1.1. Variable level

At the variable level, we focus on determining the correlation between different variables, which is the first level of the causal ladder. To obtain correlation from tabular data, we transform the problem of correlation testing into independence testing. That is given variables V_i and V_j, determining whether they are independent or conditional Independence under variables $\{V_k\}_{k=1}^{N}$. If two variables are correlated, they are statistically dependent, and vice versa. Through such modeling methods, we aim to test the causal agent with the ability to analyze correlations. Specifically, we divide the problem of correlation into two subclasses: direct independence testing and conditional independence testing. The difference between them lies in whether condition variables are given when judging independence. In particular, direct independence testing can be regarded as the number of condition variables is zero. To more finely measure the model's capabilities, we further divide conditional independence testing into independence testing under a single condition and independence testing under multiple numbers of conditions, in which the difference is whether the number of conditions is one or beyond one.

3.1.2. Causal graph level

At the causal graph level, the focus is more macroscopic, examining whether the causal agent possesses the capability to generate a causal graph. The causal graph is a directed acyclic graph (DAG), and in DAG the direction of edges represents causal relationships. In this article, we choose the PC algorithm [57] as the method for generating causal graphs, which generates Markov equivalence classes of causal graphs without considering the presence and influence of unobserved variables. Modeling the capabilities of intelligence at the level of the causal graph involves two categories: generating a causal graph that includes all variables in tabular data, and the other generating a partial causal graph that includes only a subset of variables in tabular data. The capability to generate causal graphs and to reason on these graphs can effectively guide the agent to understand causal relationships and discern true causal connections amidst the fog of spurious correlations.

3.1.3. Edge level

At the edge level modeling, we still consider the relationships between variables. Instead of the associations from a statistical correlation, we focus on the deeper causal relationships between variables from a causal viewpoint. Unlike quantitative estimation of causal effects, edge level modeling provides qualitative analysis results that need to reflect the true relationships of the edges in the causal graph reconstructed from tabular data. We consider the following three types of relationships: direct causal relationship, collider relationship, and confound relationship. As discussed in Section 3.1.2, We used the PC algorithm to generate Markov equivalence classes for causal graphs, therefore we formalize three types of relationships as follows:

Denote G as a Markov equivalence class generated by the PC algorithm from tabular data, containing edges set $\{< V_i, V_j >\}$. And edge $< V_i, V_j > \in \{\rightarrow, -\}$.

We denote direct causal relationships as V_i directly causes V_j, reflected in the causal graph G as existence edge $V_i \rightarrow V_j$. we denote the collider relationship as V_i and V_j directly cause a common variable V_k , reflected in the causal graph as existence$V_i \rightarrow V_k$ and $V_j \rightarrow V_k$. We denote the confounding relationship as the presence of unobstructed backdoor paths between V_i and V_j, reflected in the causal graph as $V_i \leftarrow ... \rightarrow V_j$

3.1.4. Causal effect level

The causal effect level attempts to quantify how the outcome for an individual or system would differ if it experienced a certain intervention. Ideally, this requires controlling confounders to ensure an accurate assessment of the intervention's effect. Thus, randomized controlled trial (RCT) is the golden standard for estimating causal effects. However, in practical production scenarios, ethical constraints or high experimental costs often make it difficult to obtain results from RCT. Additionally, the sample distribution in RCT may not represent the overall population distribution due to limited sample sizes. To address the limitations of sample size and distribution bias and to balance covariates and confounding factors when estimating causal effects, researchers have proposed numerous methods based on observational data, such as IPSW [78], etc.

We expect the causal agent not only to utilize causal explanations for qualitative analysis but also to employ classical causal inference for quantitative interpretation. To simplify the problem, at the level of causal effects, we only consider the quantitative calculation of the ATE, denoted as $E(Y(T = t_1) - Y(T = t_0))$, from tabular data. Modeling at the granularity of causal effects can equip the causal agent with a more fine-grained causal perception capability.

3.2. Causal Agent Framework Based on LLM

Based on the causal modeling methods mentioned in Section 3.1, we have specifically implemented causal agents with causal reasoning capabilities for different modeling granularities. Our causal agent framework consists of three modules: tools module, memory module, and plan module. In terms of tools, to align the tabular data with natural language, we invoke causal analysis tools that can accept tabular data as input. For the output of tools, we use prompts to interpret and explain the results, enabling the causal agent to understand the output. In the planning aspect, inspired by the ReAct framework [63], through multiple rounds of reflection, we continuously invoke causal analysis tools and reflect on their output, considering whether we can derive the answer to the original question based on the agent's understanding of the causal question. If the answer to the question cannot be derived, we continue to iterate and reflect until we reach the final answer or limited iteration times. Besides, to better understand tools' usage, we use in-context learning and one-shot examples to empower the causal agent. A manual handwritten example is designed to use all tools to guide the causal agent in invoking and understanding the tool. In terms of memory, we store the output of the causal analysis tools in a dictionary in memory as short-term memory, ensuring that the agent can continuously access the causal graph before the final answer is obtained.

3.2.1. Tools

The causal agent invokes causal analysis tools to analyze tabular data, thereby compensating for the LLM's shortcomings in handling tabular data. This approach aligns tabular data with causal conclusions and enhances the LLM's causal capabilities through tool invocation. Specifically, our causal analysis tools select the library causal-learn for causal discovery and EconML for causal inference. Starting from the perspective of modeling causal problems for the LLM, we have designed specific tool functions at the variable level, edge level, causal graph level, and causal effect level. To make the tool functions easily invoked by the LLM, we have re-encapsulated the interfaces, changing the tool inputs to JSON string format, and using manual rules and handwritten prompt templates to help the

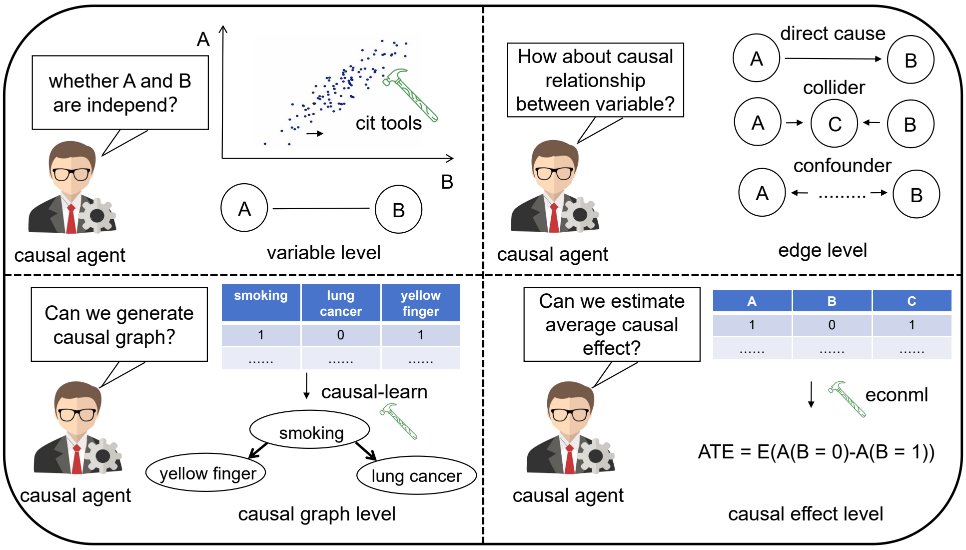

Figure 2. Causal Agent Tools Usage Diagram: Different tools are used to address causal problems at four levels. At the variable level, independence test tools are used to verify the correlation between variables. At the edge level, the relationships between causal edges are analyzed, specifically involving direct causation, collider relationships, and confounder relationships. At the causal graph level, the agent uses causal discovery algorithms to generate causal graphs. At the causal effect level, the Double Machine Learning (DML) [25] algorithm is used to obtain estimates of causal effects.

large model better understand the meaning of the tool outputs. Prompt details are in the appendix.

At the variable level, we invoke the conditional independent test class in causal-learn and use Fisherz [82] as an independent test method. At the causal graph level, since there are no unobserved variables in our data assumptions, we invoke the PC algorithm to generate the Markov equivalence class of the causal graph. It should be noted that when generating a partial causal graph, we still use the PC algorithm. However, in this case, the variables not included in arguments are unobserved variables for partial causal graphs. we think that this situation should be controlled by the user rather than the agent actively changing the causal discovery algorithm, such as the FCI [67]algorithm that can handle unobserved confounders. This design maintains the reliability of the agent's behavior and facilitates user interaction with the agent.

At the edge level, we use the tool's prompt template to guide the LLM to use the causal graph generation algorithm and obtain the Markov equivalence class of the causal graph. Then judge the relationship between the edges. For undirected edges that the PC algorithm cannot determine, the tools will categorically discuss the direction of the edge to conclude. We focus on three sub-issues at edge level: direct cause, confounding, and collider. For judging the cause relationship, we consider whether there is a directed edge directly connecting the two variables in the output G of the PC algorithm. If such a directed edge exists, the agent will determine the cause relationship based on the direction of the edge; For judging confounding, we consider whether there exist unblocked backdoor paths between the two nodes. If unblocked backdoor paths exist, the tool class will return the backdoor path. The causal agent will receive information about the presence of confounding. For judging a collider, we only consider the collider "V" structure, such as $V_i \rightarrow V_k \leftarrow V_j$.

At the level of causal effects, the causal agent invokes the LinearDML algorithm in the EconML library, where the user needs to specify which variables are covariates. The causal agent first uses the LinearDML algorithm to train the data. During the training stage, we choose the default parameters of the LinearDML algorithm in the EconML library, and the covariates are specified by the user's prompts input. After the training stage, the tool outputs an estimator of the average causal effect on the same model, using covariates consistent with those used during training. It should be emphasized that the default parameters of the LinearDML algorithm depend on the linear treatment assumption and linear heterogeneity assumption of the data. Actual test data and real-world scenario data may not meet similar assumptions, so there may be biases in the average causal effects produced. Our focus is mainly on the process of calling and understanding tools for the causal agent, while the selection of models is a more challenging aspect that can be further explored in future work.

3.2.2. Plan process

Inspired by the ReAct framework, the causal agent adopts an iterative multi-turn dialogue approach, using prompt templates to facilitate interaction and understanding between the causal agent and the tools. During the process of invoking causal analysis tools, the agent must continuously think and experiment to determine the next action. This process is crucial for the causal agent, as solving complex causal problems may require the use of causal tools more than one, as well as the observation and integration of different outputs from these tools. Therefore, the causal agent needs to repeatedly invoke these tools, observe their output, and engage in deep thinking to arrive at the solution.

Specifically, in each round of the dialogue, the user provides a problem description q and the table data T to be analyzed as input (q, T). Guided by the prompt, from the first round of each round i, the causal agent will generate a Thought O_i^t and use a tool as action a_i. The tool will accept a_i and generate observation O_i^{obs}. Based on this observation, the causal agent attempts to answer the original question description. If it is still unable to answer, it concatenates this round's observation as input to continue the next round. The input for the next round is $(q, T, O_1^t, a_1, O_1^{obs}, O_2^t, a_2, O_2^{obs}, \ldots, O_i^{obs})$, until the causal agent comes up with the final answer or reaches the maximum times of iterations.

Input prompt:
Consider four elements: age, weight, sleep time, and cancer. With the advancement of age, individuals should pay closer attention to their weight and sleep duration, as these factors can significantly impact their overall health and the risk of developing cancer. Doctors are very interested in the relationship between these variables, and therefore, they have chosen to collect a set of data through experiments. Please assist the doctors in answering whether there is a direct causal effect of weight on cancer, or whether their correlation is due to age acting as a confounding factor. Csv data stores in 'data.csv' .
Input data:data.csv

Thought: We need to determine if there is a direct causal effect of weight on cancer or if age is acting as a confounding factor. We should start by generating a causal graph to analyze the relationships between these variables.
Action: Generate Causal
Action Input: $\{"filename" : "data.csv", "analyze\_relationship" : "True"\}$

Observation: Causal graph named 'data' is generated succeed! and have written to the memory and file 'data'.
Thought: Now that we have generated the causal graph, we can use it to determine if age is a confounding factor in the relationship between weight and cancer.
Action: Determine confounder
Action Input:$\{"cg\_name" : "data", "interesting\_var" : ["weight", "cancer"]\}$

. . . . (multi times iterators)
Final Answer: There is no direct causal effect of weight on cancer.

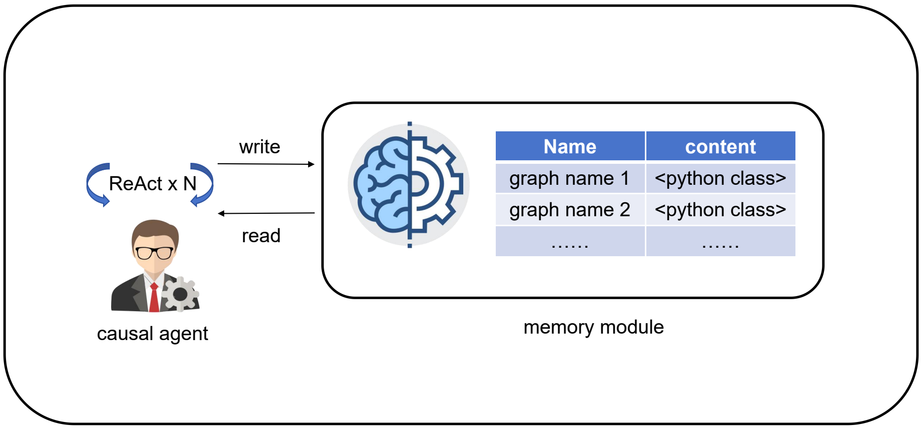

Figure 3. Causal Agent Memory Module Diagram: During the reasoning process, the causal agent maintains a memory index in its memory. The index names are in natural language form, while the index content consists of data structures such as causal graph instances containing richer information. This ensures that the memory content is not limited to text form.

The above is a concrete example where we query the causal agent with an input consisting of tabular data data.csv and a description of the problem involving "age", "weight", "sleep time", and "cancer". Upon receiving the problem input, the agent begins to think and determines that to obtain specific causal relationships, it first needs to invoke a causal graph generation tool, then generate the causal graph using the PC algorithm on tabular data by passing in specific parameters. Subsequently, the agent reasons over the causal graph and invokes the causal analysis tool to analyze the causal relationship between "weight" and "cancer". It discovers that there is no direct causal relationship between "weight" and "cancer" as described in the problem. Therefore, the specific conclusion obtained is "There is no direct causal effect of weight on cancer". Note that the tabular data in this example is synthesized and does not represent causal relationships in the real world.

3.2.3. Memory

Currently, mainstream memory mechanisms in LLM-based agents are primarily implemented in two forms: textual form and parametric form Zhang et al. [83]. Although most current memory mechanisms tend to use the textual form, parametric memory, as an emerging area of exploration, has unique application potential. Each form has its advantages and disadvantages, suitable for different application scenarios. The memory operations of an agent include three key stages: memory writing, memory management, and memory reading. These three operations interact in the agent's interaction with the environment, collectively enhancing the agent's interactive capabilities.

In this paper, the causal agent considers only short-term memory. Specifically, during the ReAct reasoning and interaction process of the causal agent, it needs to maintain the currently generated causal graph and use this graph in subsequent causal relationship judgments. Therefore, in the implementation of the causal agent's memory, the memory is not a textual form of data but the data corresponding to the causal graph Python class instance. The causal agent maintains a dictionary as memory, adding an entry and establishing a name index during memory writing, and using the index to read the corresponding causal graph information during multi-turn dialogues.

For example, some questions require invoking multiple tools. When analyzing the relationships of edges, a causal graph needs to be generated first, followed by reasoning

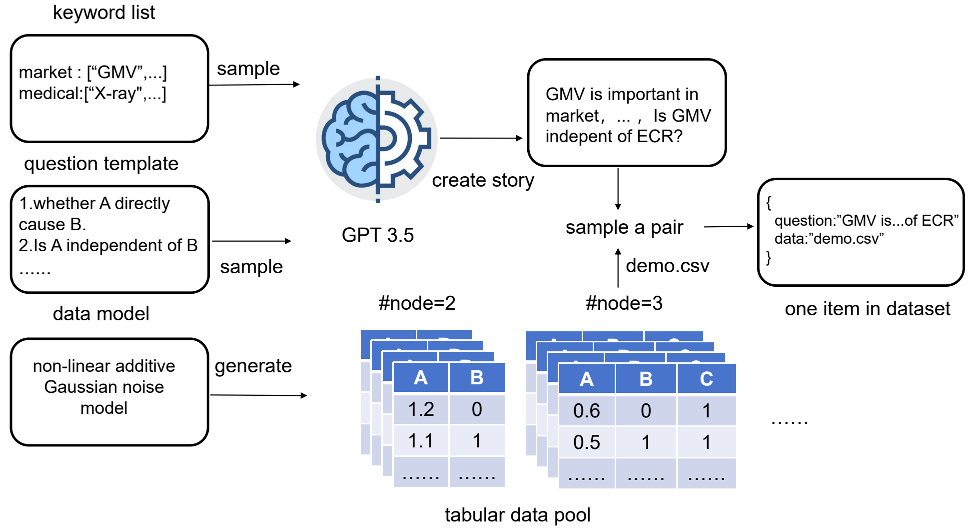

Figure 4. Overview of Data Generation Process: The final generated data consists of a pair of causal questions and tabular data. Initially, ChatGPT-3.5 samples a set of descriptions from keyword and question lists, forming a group of descriptions, and generates detailed narratives and causal question descriptions based on these. Subsequently, a non-linear additive Gaussian model is employed to generate a pool of tabular data. Data is then extracted from this pool and combined with the causal question descriptions to form a single data entry.

over the causal graph, ultimately generating the answers and results to the questions. Therefore, when answering a question, the causal agent must maintain a memory module during multi-turn reasoning dialogues, implemented as a dictionary data structure. Each time a causal graph needs to be generated, the agent assigns a name to the causal graph and stores the graph in the dictionary. If then requires analyzing the relationships between edges, the agent can use the causal graph's name as an index to retrieve and analyze the causal graph from memory.

In summary, the memory module of the causal agent differs from traditional LLM-based agent memory in that the stored data structure is not natural language but rather more informative as an abstract expression of memory.

4. Results

To test our causal intelligence agent, we start from the perspective of causal question modeling in Section 3.1 and have designed a series of question templates for variable-level, edge-level, causal graph-level, and causal effect-level respectively, details in the appendix. To obtain the data required for causal questions, we have generated tabular data using the nonlinear additive gaussian noise method [89]. In addition, for questions at the variable level, edge level, causal graph level, and causal effect level, we have constructed a dataset of size 1.3K for testing by using ChatGPT-3.5, as shown in Figure 4. Through testing, our causal agent has achieved high accuracy over four-level questions. All three sub-problems at the variable level achieved an accuracy of over 92%, all three sub-problems at the edge level achieved an accuracy of over 89%, all three sub-problems at the causal graph level achieved an accuracy of over 81%, and all two sub-problems at the causal effect level achieved an accuracy of over 93%;

4.1. Data generate process

4.1.1. tabular data

To generate the tabular data needed for the test dataset, we adopt the method used in the work by Rolland et al. [89]. Specifically, our data follows non-linear additive Gaussian noise models, $X \in R^d$ is generated using the following model:

$$X_i = f_i(pa_i(X)) + \epsilon_i$$

$i = 1, \ldots, d$, where $pa_i(X)$ selects the coordinates of X which are parents of node i in some DAG.The noise variables

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

The functions f_i are assumed to be twice continuously differentiable and non-linear in every component. That is, if we denote the parents $pa_j(X)$ of X_j by $X_{k1}, X_{k2}, \ldots, X_{kl}$, then, for all $a = 1, \ldots, l$, the function $f_j(X_{k1}, \ldots, X_{k_{a-1}}, \cdot, X_{k_{a+1}}, \ldots, X_{kl})$ is assumed to be nonlinear for some $X_{k1}, \ldots, X_{k_{a-1}}, \cdot, X_{k_{a+1}}, \ldots, X_{kl} \in R^{l-1}$

Through the aforementioned method, we have generated a series of tabular data with node counts ranging from 3 to 10. For tables with the same number of nodes, we generate a series of tabular data with different edge numbers, which range in $[0, C^2_{\#node}]$. This simulates different scenarios of sparsity and density of real causal graphs. We use all the generated tabular data as a data pool. When generating specific test samples later, we will randomly take a table from the table pool that has the same number of nodes as variables required by the question, to be a quantitative expression of the relationships between variables.

4.1.2. causal problem descriptions

To simulate causal issues in real scenarios, we generate a natural language template T_q for the four-level causal questions modeled in Section 3.1. Then we take the medical field and the market field, two common fields for causal inference, as question sources to generate questions' real-world scenes. We first used ChatGPT-3.5 to generate 100 keywords related to medical and market as list L_q . Then, we iteratively traversed through node counts from 3 to 10. For a node count of i, we randomly drew i keywords from the keyword list as seeds

$$K_1, ..., K_i \sim L_q$$

.Subsequently, we used the question template $t_q^i \sim T_q$ to allow ChatGPT-3.5 to construct a possible real scenario using the seed keywords, thus forming a piece of data.

$$description = GPT(K1, ..., K_i, t_q^i)$$

Note that the keywords that come from the list are randomly drawn, so there may be no causal relationship between them. This may result in a case in which the causal graph generated from tabular data sampled randomly in Section 4.1 is quite different from the actual scenario. This special design makes the causal agent focus on the tabular data during the process of causal analysis, leading to data-driven rather than semantic causal information between variables for causal analysis. This design also fits with counterfactual thinking, that is, the causal relationships between variables in reality and the causal relationships between variables implied by the randomly drawn tabular data in our data may be inconsistent. The causal agent may need to draw counterfactual conclusions in such scenarios.

4.2. causal problem test result

The causal agent was constructed by ChatGPT-3.5, and we set temperature parameters as 0.5 when the causal agent reasoned. We tested the causal capabilities of our causal agent. To constrain the output of the LLM and facilitate comparison with the actual ground truth during testing, we guide the model's output as format follows: For variable-level

Table 2. The test results of the causal agent on variable-level problems are represented in the table, where the values indicate accuracy, calculated as the number of correctly answered questions divided by the total number of questions. We use IT to represent the independent test, CIT to represent the conditional independent test with one variable as a conditional variable, and MULTCIT to represent a conditional independent test with beyond one variable as conditional variables.

#node	IT	CIT	MULTCIT
3	95.0	100.0	-
4	95.0	100.0	100.0
5	95.0	100.0	100.0
6	95.0	100.0	100.0
7	95.0	95.4	100.0
8	95.7	100.0	100.0
9	90.0	100.0	100.0
10	80.0	100.0	100.0
average	92.6	99.4	100.0

questions, we restricted the model's output to "yes" and "no." For instance, when a question required the causal agent to determine whether two variables were independent, the agent would output "yes" if they were independent, and "no" if they were not, based on its analysis of the question. Similarly, for edge-level questions, we limited the output to "yes", "no", and "uncertain", indicating whether the description of the relationship between the edges was correct, incorrect, or uncertain. For causal graph-level questions, the agent would generate a causal graph during the reasoning process, and we would directly assess whether the causal graph was correctly generated and return the correct name of the causal graph so that users could find it. For causal effect-level questions, we considered whether the agent's calculation of the average causal effect was accurate. Note that even though we imposed format restrictions on the model's output, the agent would still output other equivalent descriptions consistent with the answer instead of adhering strictly to the format, especially in the case of zero-shot. For example, when judging variable-level questions, we restricted the agent to output only "yes" or "no", but sometimes the agent's output included a summary of the question rather than "yes" or "no", such as "A is independent of B". In such cases, we regard it as a wrong answer because it doesn't follow the correct output format. To optimize this problem, we adopted a in contextual learning (ICL) method, using a fixed manually case to guide the causal agent to call the tool and output the correct format.

At the variable level, our results, as shown in the Table 2. We use IT to represent the independent test, CIT to represent the conditional independent test with one variable as a conditional variable, and MULTCIT represents a conditional independent test with beyond one variable as conditional variables . The causal agent achieved over 92% accuracy across the three sub-questions of the variable level. Notably, in the conditional independence test, the agent correctly utilized the tools and reached the correct conclusions on almost all questions, achieving a 99.4% in one conditional independent test and 100% in a multi-conditional independent test, indicating our causal agent performed very well in this area.

At the edge level, we tested the agent's accuracy in judging direct causal relationships (represented by CAUSE), confounding factors (represented by CONF), and colliders (represented by COLLIDER), with the results shown in Table 3. Specifically, the agent achieved 89.5% accuracy in judging direct cause relationships on average, 97.4% accuracy in judging colliders on average, and 94.6% accuracy in judging confounders. At the causal graph level, we tested the agent's ability to generate a causal graph containing all variables (represented by TOTAL) and a partial causal graph (represented by PARTIAL) containing some variables. The specific results, as shown in Table 4, were 81.8% accuracy rate for

Table 3. The test results of the causal agent on edge-level problems are represented in the table, where the values indicate accuracy, calculated as the number of correctly answered questions divided by the total number of questions. We use CAUSE to represent judging direct causal relationships, CONF to represent judging confounding factors, and COLLIDER to represent judging colliders.

#node	CAUSE	COLLIDER	CONF
3	95.0	100.0	95.5
4	95.0	95.0	100.0
5	90.0	95.0	100.0
6	80.0	95.7	100.0
7	88.9	100.0	95.5
8	83.3	94.4	86.4
9	88.9	100.0	95.5
10	94.4	100.0	86.4
average	89.5	97.4	94.6

Table 4. The test results of the causal agent on causal graph level problems are represented in the table, where the values indicate accuracy, calculated as the number of correctly answered questions divided by the total number of questions. We use TOTAL to represent generate a causal graph containing all variables, and PARTIAL to represent a partial causal graph containing some partial variables

#node	TOTAL	PARTIAL
3	90.9	-
4	63.6	95.5
5	100.0	77.3
6	90.9	95.5
7	77.3	95.5
8	86.4	86.4
9	63.6	95.5
10	81.8	95.5
average	81.8	91.6

Table 5. The test results after stratifying the answers and question domains are shown, where "yes", "no", and "uncertain" represent the three different answers in the ground truth. The values in the table indicate accuracy, calculated as the number of correctly answered questions divided by the total number of questions. Blue represents the medical domain, and red represents the market domain.

answer	IT	CIT	MULTCIT	CAUSE	CONF	COLLIDER
yes	100	100.0	100.0	72.7	90.9	87.5
	94.9	100.0	100.0	92.9	60.0	84.6
no	96.2	100	100.0	87.3	100	100
	81.8	97.6	100.0	91.22	98.5	98.5
uncertain	-	-	-	100.0	72.7	100.0
	-	-	-	100.0	75.0	100.0
average	97.5	100.0	100.0	86.11	94.5	98.6
	87.9	98.8	100.0	92.5	93.8	96.4

generating a causal graph with all nodes, and 91.6% accuracy rate for generating a partial causal graph composed of some nodes.

At the level of causal effects, the agent answered 15 out of 16 questions correctly, achieving an accuracy of 93.8%. Specifically, we have 2 instances with 3-10 nodes and 16 instances in total.These examples cover the two fields of marketing and medical, and cover 3-10 nodes. So they are representative. An example is as follow.

> As a statistician, you are working for an online retail company that has been experiencing a high rate of Cart Abandonment. The company believes that its Logo Design might be a factor influencing this issue. To test this hypothesis, the company has applied Programmatic Advertising, a method that uses automated systems to buy and sell ads in real-time. This method has been used to subtly alter the Logo Design displayed to customers, with the aim of reducing Cart Abandonment. You have collected data where the treatment variable, Programmatic Advertising, ranges from -0.46 to -0.11. The negative values indicate a decrease in the intensity of Programmatic Advertising. Now, you need to understand the causal relationship between these variables. So, your question is: Can you calculate the Average Treatment Effect (ATE) of the continuous treatment variable Programmatic Advertising on the outcome variable Logo Design, given that the treatment Programmatic Advertising change from -0.46 to -0.11?

We first use the DML algorithm and generated the ATE values as ground truth, and then test whether the causal agent can correctly call the DML tool and pass the correct parameters. When the tool is called correctly and the output of the tool can be correctly understood, the agent will output the correct answer.

Additionally, through analysis, we found that within the range where the number of nodes is between 3 and 10, the number of nodes has little impact on the correctness of the agent's use of causal tools and causal inference. Most errors were due to the agent's misunderstanding of the output from the causal tools during the inference process. This also means that the fluctuation in the correct rate does not have a significant correlation or impact based on the number of nodes. The cause of the agent's errors lies in understanding the tool's output and how to align the tool's output with the question's answer.

Therefore, we examine the correlation and impact between the ground truth of the problem and the answer's accuracy rate and explore how the types of domains involved in causal questions affect correctness. We conducted stratified exploration based on the domains involved in the problems in our test set, which are the medical domain and market domain. Through stratification, we can see the impact of the problem domain and the answer on the results, as shown in Table 5. Different problem domains lead to different complexities and different descriptions of the problems, which affects the agent's use of causal tools and the answers to the problems. In our examples, when a causal agent should analyze an independent relationship, the agent performed better in utilizing tools within the market domain, while its accuracy decreased in the medical domain. When judging edge relationships, the two domain's differences are slight. Moreover, there were noticeable differences in the agent's accuracy under different answers to the questions; compared to giving "yes" and "uncertain" conclusions, our agent was more inclined to provide negative conclusions such as "no" to the questions.

5. Discussion

In this work, we harnessed LLM to construct a causal agent by invoking causal tools, modeling causal problems across four levels in the causal domain, and endowing the large model with causal capabilities at four levels, followed by an assessment of the agent's abilities. The experimental results of the agent in solving causal problems showed that it performed particularly well at the variable level. In tasks of independence testing, accuracy rates exceeded 92% and even reached 100% in the multi-conditional independence test. This endowed the agent with the ability to leverage correlation analysis driven by tabular data. At the edge level, the agent achieved accuracy rates of over 89% in judging direct causal relationships, confounding factors, and colliders, indicating its high capability in handling causal relationships. At the causal graph level, the accuracy rates for generating complete and partial causal graphs were 81.8% and 91.6%, respectively, demonstrating the agent's potential in constructing causal relationship networks using data-driven approaches and causal tools. The agent can correctly invoke tools and generate causal graphs, which is significant for the popularization of the causal community and the automation of causal tools. Even users who are not familiar with the concept of causality can utilize the agent to produce an end-to-end process from data to causal graphs. At the causal effect level, the agent can produce the correct average causal effect, achieving an accuracy rate of 93% on our small-scale dataset.

Moreover, the use of causal tools ensures interpretability and reliability, which is of great significance for the future practical application of the causal agent.

Analyzing the agent's errors, we can find that there is a bias in the agent's understanding and application of causal tools, leading to a decrease in accuracy in some cases. However, this issue will be gradually resolved as the capabilities and generalization performance of large models improve. From this perspective, causal scientists can focus on improving the interaction efficiency and accuracy of the agent and causal tools. As the capabilities of LLM are enhanced in the future, the agent's causal inference capabilities will also increase accordingly. Additionally, the agent's performance varies across different domains (such as marketing and medical), indicating that domain-specific knowledge and further domain adaptation may help improve the agent's performance. Addressing the issue of poor robustness of the agent in different domains will greatly affect the practical application of the causal agent. Moreover, the current causal agent cannot select models and perceive data. The agent in this work only directly invokes simple causal models, such as the PC algorithm and LinearDML algorithm, but the applicability of these algorithms is limited and heavily relies on our functional assumptions about the data and the assumption of no confounding. How to endow the agent with the ability to perceive data and to have prior understanding and knowledge of tool invocation is of great significance for the agent's promotion and practical application.

Author Contributions: Conceptualization, K.H., K.K., and F.W.; methodology, K.H., K.K., and Z.Z; software, K.H.; validation, K.H.; formal analysis, K.H., K.K., J.Y.; investigation, K.H.; resources,K.K., F.W., J.Y.; data curation, K.H.; writing—original draft preparation, K.H.; writing—review and editing, K.K., Z.Z, J.Y and F.W.; project administration, K.H., F.W., K.K.; funding acquisition, F.W., K.K. and J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: We use synthetic data in this work. Specific data can be accessed via the GitHub repository link https://github.com/Kairong-Han/Causal_Agent.

Conflicts of Interest: The authors declare no conflicts of interest

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
AGI	Artificial General Intelligence
ATE	Average Treatment Effect
CATE	Conditional Average Treatment Effect
DAG	Directed Acyclic Graph
AI	Artificial Intelligence
RCT	Randomized Controlled Trial
DML	Double Machine Learning
ICL	In Context Learning
IT	Independent Test
CIT	Conditional Independent Test
MULTCIT	Multi-variables Conditional Independent Test
CAUSE	direct causal relationships
CONF	confounding factors
COLLIDER	colliders
TOTAL	causal graph containing all variables
PARTIAL	partial causal graph containing some partial variables

Appendix A. The prompts we use in causal agent

Appendix A.1. input prompt template w/o ICL

Table A1. Question input prompt template without ICL: When calling the causal agent, the question input template replaces the parameters in the following brackets with the input part and allows the agent to answer the question.

```
Answer the following questions as best you can. You have access to the following tools:

{toolsdiscription}

Use the following format:
Question: the input question you must answer
Thought: you should always think about what to do
Action: the action to take, should be one of

[toolnames]

Action Input: the input to the action
Observation: the result of the action
... (this Thought/Action/Action Input/Observation can repeat N times)
Thought: I now know the final answer
Final Answer: the final answer to the original input question
Begin!
Question:
{input}

Thought:
{agentscratchpad}
```

Table A2. Question input prompt template with ICL: When calling the causal agent, the question input template replaces the parameters in the following brackets with the input part and allows the agent to answer the question.

```
Answer the following questions as best you can. You have access to the following tools:

{toolsdiscription}

Use the following format:
Question: the input question you must answer
Thought: you should always think about what to do
Action: the action to take, should be one of

[toolnames]

Action Input: the input to the action
Observation: the result of the action
... (this Thought/Action/Action Input/Observation can repeat N times)
Thought: I now know the final answer
Final Answer: the final answer to the original input question
Begin!
##DEMO:
The doctor wants to explore the relationship between smoking, lung cancer, and yellow fingers,
so he collects a batch of data, stores it in 'data.csv', and gives the following proposition:
if all of the statements are false, output no; if one of the statements is true, output
uncertain; if all of the statements are true, output yes. The proposition is as follows:
1. Yellow fingers are a cause of lung cancer
2. There is a collider between yellow finger and lung cancer
3. There is a confounder between yellow fingers and lung cancer
4. Yellow fingers and lung cancer are independent of each other given smoking
The output is just formatted as a json string, such as {"answer":"yes"}.
Thought: I need to generate a causal diagram, and then analyze the causal and independent
relationships between variables to judge the correctness of each proposition
Action:Generate Causal
Action Input:{"filename":"data.csv","analyserelationship":"True"}
Observation:causal graph named 'data' is generate succeed! and have written to the memory.

Thought: I need to use the Determine edge directions tool to determine if the yellow finger is
the cause of smoking
Action:Determine edge directions
Action Input:{"cgname":"data","interestingvar":["yellowfingers","lungcancer"]}
Observation:There is no direct edge linking yellow finger and cancer

Thought: There is no direct causal link between yellow finger and lung cancer. The
proposition "yellow finger is the cause of lung cancer" is wrong. I need to continue to
determine whether there is a collider between yellow finger and lung cancer
Action:Determine collider
Action Input:{"cgname":"data","interestingvar":["yellowfingers","lungcancer"]}
Observation:There don't exists collider Smoking of yellow fingers and lung cancer

Thought: There is no collider between yellow finger and lung cancer, I need to continue to
determine whether there is a confounder between yellow finger and lung cancer
Action:Determine confounder
Action Input:{"cgname":"data","interestingvar":["yellowfingers","lungcancer"]}
Observation:yes, There is an unblocked backdoor path between yellow finger and lung cancer so
confounder exists. Backdoor path: yellow fingers, smoking, lung cancer

Thought: There is an unblocked backdoor path between yellow finger and lung cancer, so there
is a confounder, and it is smoking, I need to continue to determine whether yellow finger and
lung cancer are independent of each other given smoking
Action:condition independent test
Action Input:{"filename":"data.csv","interestingvar":["yellowfingers","lungcancer"],"condition":
["somking"]}
Observation:yellow fingers and lung cancer is independent under conditions: smoking

Thought:i now get the final answer. I have analyzed the four propositions, among which the
confounder between yellow fingers and lung cancer is correct, and the others are all wrong, so
the answer is uncertian

Final Answer:{"answer":"uncertain"}

##Requirement:
Answer the following questions with examples:

Question:
{input}

Thought:
{agentscratchpad}
```

Appendix A.2. tools prompt template

Table A3. Independent test tool's description prompt

```
condition independent test:  Useful for when you need to test the *** independent or d-separate
*** of variable A and variable B condition on variable C. input should be a json with the
format below
```

$$\{"filename" : ..., "interestingvar" : [...], "condition" : [...]\}$$

```
interesting var is a list of variables the user interested in.  For example, if the user wants
to test independent(d-separate) between X and Y conditions on Z, W,Q, interesting var is
```

$$["X", "Y"]$$

```
, condition is
```

$$["Z", "W", "Q"]$$

```
.  condition is
```

$$[]$$

```
if no condition is provided
```

Table A4. Causal graph generating tool's prompt template

```
Generate Causal:  Useful for when you need to generate causal graph (or partial causal graph).
input should be a json with the format below
```

$$\{"filename" : ..., "analyserelationship" : ..., "interestingvar" : [...](Optional)\}$$

```
.if you want to analyze relationship between variables( such as cause-effect, coufounder
, Collider), analyse relationship = "True" and please generate complete causal graph and
interesting var is [](which means causal graph contain all variables).if we only need
to generate **partial causal graph** (for example, generate a partial causal graph for
some variables), interesting var is used and it's values are list of variables appear in
causal graph and analyse relationship is "False".Further more, if needed, you can analyse
variables relationship in causal graph generated by this tool through these tools:  Determine
collider,Determine confounder,Determine edge direction
```

Table A5. Collider structure test tool's prompt template

```
Determine collider:  you should first generate causal graph and then use this tool.  Useful
When we are interested in whether there is a collider between two variables(ie common effect),
we use this tool and the input is
```

$$\{"cgname" : ..., "interestingvar" : [...]\}$$

```
, where interesting var is what Variable we want to test, cg name is the name of causal
generated by 'Generate Causal'.The output of the tool is yes or no or uncertainty and may be
the variable name of the collider.  Make sure the causal graph has been generated before using
this tool
```

Table A6. Confounder structure test tool's prompt

```
Determine confounder:  you should first generate causal graph and then use this tool.  Useful
When we are interested in whether there is a cofounder (ie common cause) between two variables,
we use this tool and the input is
```

$$\{"cgname" : ..., "interestingvar" : [...]\}$$

```
, where interesting var is what Variable we want to test, cg name is the name of causal
generated by 'Generate Causal'.The output of the tool is yes or no or uncertainty and the
backdoor path that may lead to the existence of the cofounder.  Make sure the causal graph
has been generated before using this tool
```

Table A7. Causal relationship direction test tool's prompt

```
Determine edge directions:  you should first generate causal graph and then use this
tool.Useful when we are interested in whether there is a direct edge between two variables
and the direction of the edge (such as determining whether A directly leads to B)., we use this
tool and the input is
```

$$\{''cgname'' = ..., ''interestingvar'' = [...]\}$$

```
, where interesting var is what Variable we want to test, cg name is the name of causal
generated by 'Generate Causal'.The output of the tool is the relationship of two variables
(ie A cause B). Make sure the causal graph has been generated before using this tool
```

Table A8. Causal effect tool's prompt

calculate CATE: Useful for when you need to calculate (conditional) average treatment effect (ATE or CATE, etc. in math function is $E(Y(T = T1) - Y(T = T0)|X = x)$ and means if we use treatment, what uplift we will get from treatment).This tool use double machine learn algorithm to calculate ate. input is a json with format

$$\{''filename'' : ..., config : \{Y : [...], T : [...], X : [...], T0 : ..., T1 : ...\}\}$$

. Y are names of outcome, T are names of treatment, X are names of covariate affect both T and Y (i.e. confounder). T1 and T0 are two different values of T that need to be calculated in ATE. you should extract each name from the description.

Appendix A.3. Data generate detail

We use the following prompt to guide ChatGPT3.5 to generate descriptions of causal problems.

> ##Requirements: Suppose you are a statistician and need to perform causal analysis on data. You need to use your imagination to compile a reasonable scene description based on the following elements, and finally ask a question Q: " ". The scenario description needs to be related to the problem and form a paragraph together with the problem. This output must end up with the question format, either directly end up with the question Q or the equivalent of the question Q. Below are all the elements you need to use to describe the scenario (including those involved in the question Q). Elements don't exist in variables listed below are not allowed.
> ##element:[]
> ##Output:

We use question template as follow:

Table A9. Independent test (IT)

```
"whether {} and {} is independent."
"Is {} independent of {}?
"Are {} and {} statistically independent?"
"Does the occurrence of {} independent on {}, or vice versa?"
"Can we assert {} and {} are independent, or are they related?"
"Can we consider {} and {} as independent events?"
"Do {} and {} independent and don't have any influence on each other?"
"Is there no statistically correlation between {} and {}?"
"test whether Are {} and {} statistically unrelated or dependent?"
"Test the independence of {} and {}."
```

Table A10. Condition independent test (CIT)

```
"whether {} and {} is independent under condition {}?"
"Is {} independent of {} given condition {}?"
"Are {} and {} statistically independent given the condition {}?"
"Does the independence of {} and {} hold true under condition {}?"
"Can we consider {} and {} as conditionally independent with respect to {}?"
"Is the independence between {} and {} maintained given the condition {}?"
"Are {} and {} conditionally independent with the presence of condition {}?"
"Can we assume that {} and {} are independent given the condition {}?"
"Is the independence of {} and {} upheld in the presence of condition {}?"
"Does the independence between {} and {} persist under the condition {}?"
```

Table A11. Mult-conditional independent test (MULTCIT)

```
"whether {} and {} is independent under conditions :  "
"Determine the independence of {} and {} given the following conditions :  "
"Examine if {} and {} are independent under the specified conditions :  "
"Assess the independence between {} and {} with the provided conditions :  "
"Investigate whether {} and {} exhibit independence given the outlined conditions :  "
"Explore the independence of {} and {} under the given circumstances :  "
"Ascertain if there is independence between {} and {} given the stated conditions :  "
"Check for independence between {} and {} based on the conditions described :  "
"Verify the independence status of {} and {} under the listed conditions :  "
"Evaluate the independence of {} and {} under the mentioned conditions :  "
"Examine whether {} and {} are independent, considering the provided conditions :  "
```

Table A12. Directly cause (CAUSE)

```
"whether {} directly cause {}."
"Assess if {} has a direct causal impact on {}."
"Examine the direct causation relationship.if {} directly cause {}?"
"Investigate whether {} directly influences {}."
"Evaluate if there exists the direct causal connection from {} to {}."
"Scrutinize if {} leads to a direct causation of {}."
"Determine whether {} is a direct cause of {}."
"Assess if there is the direct causal link of {} to {}."
"Verify if {} directly results in the causation of {}."
```

Table A13. Collider (COLLIDER)

```
"Whether there exists at least one collider (i.e., common effect) of {} and {}"
"Determine if there is at least one common effect (collider) of both {} and {}."
"Assess the presence of a shared outcome, serving as a collider, for variables {} and {}."
"Examine the potential existence of a shared consequence as a collider for {} and {}."
"Evaluate if {} and {} share a common effect (collider)."
"Analyze the presence of a common outcome serving as a collider for {} and {}."
"Verify if there exists a shared effect, acting as a collider, for both {} and {}."
"Explore whether a common consequence is a collider for variables {} and {}."
"Assess the existence of at least one common effect (collider) between {} and {}."
```

Table A14. Confounder (CONF)

```
"There exists at least one confounder (i.e., common cause) of {} and {}."
"Confirm the presence of at least one common cause (confounder) influencing both {} and {}."
"Verify whether there exists a shared factor, acting as a confounder, for variables {} and {}."
"Examine the potential existence of a common cause (confounder) impacting both {} and {}."
"Assess if {} and {} share at least one confounding factor (common cause)."
"Scrutinize the presence of a shared influencing factor, serving as a confounder, for {} and
{}."
"Investigate whether there is at least one confounder affecting both {} and {}."
"Analyze the potential impact of a common cause (confounder) on variables {} and {}."
"Verify the presence of a shared influencing factor, acting as a confounder, for {} and {}."
"Explore whether a common factor is a confounder for variables {} and {}."
"Evaluate the existence of at least one confounder (common cause) between {} and {}."
```

Table A15. Total variables' causal graph (TOTAL)

```
"please generate causal graph of the input tabular data."
"Produce a causal graph representing the relationships within the given tabular data."
"Generate a directed graph that illustrates the causal connections inherent in the provided
tabular dataset."
"Create a graphical model depicting the causality among variables in the input tabular data."
"Construct a causal diagram illustrating the interdependencies among the variables in the
tabular dataset."
"Formulate a graph that visually represents the cause-and-effect relationships present in the
input tabular information."
"Develop a graphical representation outlining the causal structure of the tabular data."
"Build a directed acyclic graph (DAG) that reflects the causal influences within the input
tabular dataset."
"Establish a graphical model showcasing the causal links between variables derived from the
tabular data."
"Design a causal graph that visually captures the cause-and-effect relationships inherent in
the tabular information."
"Construct a directed graph that visually displays the causal pathways within the given tabular
dataset."
```

Table A16. Parital variables' causal graph (PARTIAL)

```
"Please generate a partial causal diagram for some of the following variables that interest me
:  "
"Generate a subset of a causal diagram for the variables of interest :  "
"Create a partial graphical model illustrating causal relationships among selected variables :
"
"Develop a restricted causal graph focusing on specific variables from the given set :  "
"Formulate a partial directed acyclic graph (DAG) depicting causal connections for chosen
variables :  "
"Construct a limited causal diagram featuring only the variables of interest :  "
"Produce a subsection of a graphical model, emphasizing the causal links within the selected
variables :  "
"Build a causal graph subset, emphasizing relationships among the variables you find intriguing
:  "
"Develop a focused causal diagram, highlighting causal connections for the specified variables
:  "
"Form a segment of a directed graph that visually represents causal relationships among chosen
variables :  "
"Create a restricted causal network, showcasing the partial causal influences among the
variables of interest :  "
```

Table A17. Causal effect

```
calculate the Average Treatment Effect (ATE) of a continuous treatment variable {T} on an
outcome variable {Y}, given that the treatment {T} changes from {T0} to {T1}.
```

Appendix B. The detailed example of the causal agent with ICL

Answer the following questions as best you can. You have access to the following tools:
condition independent test: Useful for when you need to test the *** independent or d-separate *** of variable A and variable B condition on variable C. input should be a json with format below

$$\{''filename'' : ..., ''interestingvar'' : [...], ''condition'' : [...]\}$$

interesting var is a list of variables user interested in. for example, if user want to test independent(d-separate) between X and Y condition on Z,W,Q , interesting var is

$$[''X'', ''Y'']$$

, condition is

$$[''Z'', ''W'', ''Q'']$$

. condition is

$$[]$$

Generate Causal: Useful for when you need to generate causal graph (or partial causal graph). input should be a json with format below

$$\{''filename'' : ..., ''analyserelationship'' : ..., ''interestingvar'' : [...](Optional)\}$$

.if you want to analyse relationship between variables(such as cause effect, coufounder , Collider), analyse relationship = "True" and please generate complete causal graph and interesting var is [](which means causal graph contain all variables) .if we only need to generate **partial causal graph** (for example, generate a partial causal graph for some variables), interesting var is used and it's values are list of variables appear in causal graph and analyse relationship is "False".Further more, if needed, you can analyse variables relationship in causal graph generated by this tool through these tools : Determine collider,Determine confounder,Determine edge direction
Determine collider: you should first generate causal graph and then use this tool.Useful When we are interested in whether there is a collider between two variables(ie common effect), we use this tool and the input is

$$\{''cgname'' : ..., ''interestingvar'' : [...]\}$$

, where interesting var is what Variable we want to test, cg name is the name of causal generated by 'Generate Causal'.The output of the tool is yes or no or uncertainty and may be the variable name of the collider. Make sure the causal graph has been generated before using this tool
Determine confounder: you should first generate causal graph and then use this tool.Useful When we are interested in whether there is a cofounder (ie common cause) between two variables, we use this tool and the input is

$$\{''cgname'' : ..., ''interestingvar'' : [...]\}$$

, where interesting var is what Variable we want to test, cg name is the name of causal generated by 'Generate Causal'.The output of the tool is yes or no or uncertainty and the backdoor path that may lead to the existence of the cofounder. Make sure the causal graph has been generated before using this tool
Determine edge directions: you should first generate causal graph and then use this tool.Useful when we are interested in whether there is a direct edge between two variables and the direction of the edge (such as determining whether A directly leads to B)., we use this tool and the input is

$$\{''cgname'' = ..., ''interestingvar'' = [...]\}$$

, where interesting var is what Variable we want to test, cg name is the name of causal generated by 'Generate Causal'.The output of the tool is the relationship of two variables (ie A cause B). Make sure the causal graph has been generated before using this tool
calculate CATE: Useful for when you need to calculate (conditional) average treatment effect (ATE or CATE, etc. in math function is $E(Y(T = T1) - Y(T = T0)|X = x)$ and means if we use treatment, what uplift we will get from treatment).This tool use double machine learn algorithm to calculate ate. input is a json with format

$$\{''filename'' : ..., config : \{Y : [...], T : [...], X : [...], T0 : ..., T1 : ...\}\}$$

. Y are names of outcome, T are names of treatment, X are names of covariate affect both T and Y (i.e. confounder). T1 and T0 are two different values of T that need to be calculated in ATE. you should extract each name from the description.
Use the following format:
Question: the input question you must answer
Thought: you should always think about what to do
Action: the action to take, should be one of
[Determine collider,Determine confounder,Determine edge directions,condition independent test,Generate Causal,calculate ATE]
Action Input: the input to the action
Observation: the result of the action
... (this Thought/Action/Action Input/Observation can repeat N times)
Thought: I now know the final answer
Final Answer: the final answer to the original input question
Begin!
##DEMO:
The doctor wants to explore the relationship between smoking, lung cancer, and yellow fingers, so he collects a batch of data, stores it in 'data.csv', and gives the following proposition: if all of the statements are false, output no; if one of the statements is true, output uncertain; if all of the statements are true, output yes. The proposition is as follows:
1. Yellow fingers are a cause of lung cancer
2. There is a collider between yellow finger and lung cancer
3. There is a confounder between yellow fingers and lung cancer
4. Yellow fingers and lung cancer are independent of each other given smoking

The output is just formatted as a json string, such as $\{"answer" : "yes"\}$.

Thought: I need to generate a causal diagram, and then analyze the causal and independent relationships between variables to judge the correctness of each proposition
Action:Generate Causal
Action Input:$\{"filename" : "data.csv", "analyserelationship" : "True"\}$
Observation:causal graph named 'data' is generate succeed! and have written to the memory.

Thought: I need to use the Determine edge directions tool to determine if the yellow finger is the cause of smoking
Action:Determine edge directions
Action Input:$\{"cgname" : "data", "interestingvar" : ["yellowfingers", "lungcancer"]\}$
Observation:There is no direct edge linking yellow finger and cancer

Thought: There is no direct causal link between yellow finger and lung cancer. The proposition "yellow finger is the cause of lung cancer" is wrong. I need to continue to determine whether there is a collider between yellow finger and lung cancer
Action:Determine collider
Action Input:$\{"cgname" : "data", "interestingvar" : ["yellowfingers", "lungcancer"]\}$
Observation:There don't exists collider Smoking of yellow fingers and lung cancer

Thought: There is no collider between yellow finger and lung cancer, I need to

continue to determine whether there is a confounder between yellow finger and lung cancer
Action:Determine confounder
Action Input:$\{"cgname" : "data", "interestingvar" : ["yellowfingers", "lungcancer"]\}$
Observation:yes, There is an unblocked backdoor path between yellow finger and lung cancer so confounder exists. Backdoor path: yellow fingers, smoking, lung cancer

Thought: There is an unblocked backdoor path between yellow finger and lung cancer, so there is a confounder, and it is smoking, I need to continue to determine whether yellow finger and lung cancer are independent of each other given smoking
Action:condition independent test
Action Input:$\{"filename" : "data.csv", "interestingvar" : ["yellowfingers", "lungcancer"], "condition" : ["somking"]\}$
Observation:yellow fingers and lung cancer is independent under conditions: smoking

Thought:i now get the final answer. I have analyzed the four propositions, among which the confounder between yellow fingers and lung cancer is correct, and the others are all wrong, so the answer is uncertian

Final Answer:$\{"answer" : "uncertain"\}$

##Requirement:
Answer the following questions with examples:

Question:Consider four elements : age, weight, sleep time, cancer. With the advancement of age, individuals should pay closer attention to their weight and sleep duration, as these factors can significantly impact their overall health and the risk of developing cancer. Doctors are very interested in the relationship between these variables, and therefore, they have chosen to collect a set of data through experiments. Please assist the doctors in answering whether there is a collider of weight on cancer.if exist collider, please give it name. csv data store in '4 6 3.csv' .
Thought:

> Entering new AgentExecutor chain...
I need to first generate a causal graph to analyze the relationship between age, weight, sleep time, and cancer.
Action: Generate Causal
Action Input: $\{"filename" : "463.csv", "analyserelationship" : "True"\}$

Observation: causal graph named '4 6 3.csv 2' is generate succeed! and have written to the memory and file '4 6 3.csv 2'.
Thought:Now that I have generated the causal graph, I can use the Determine collider tool to check for a collider between weight and cancer.
Action: Determine collider
Action Input: $\{"cgname" : "463.csv2", "interestingvar" : ["weight", "cancer"]\}$

Observation: There exists at least one collider sleep time of weight and cancer

Thought:I now know that there is a collider of weight on cancer, and the collider is sleep time.
Final Answer: The collider of weight on cancer is sleep time.

> Finished chain.

References

1. Kocoń, J.; Cichecki, I.; Kaszyca, O.; Kochanek, M.; Szydło, D.; Baran, J.; Bielaniewicz, J.; Gruza, M.; Janz, A.; Kanclerz, K.; et al. ChatGPT: Jack of all trades, master of none. *Information Fusion* **2023**, *99*, 101861.
2. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models, 2023, [arXiv:cs.CL/2302.13971].
3. Chilvers, C.; Goldstein, H. Journal of the Royal Statistical Society Series A (Statistics in Society). *Journal of the Royal Statistical Society Series A: Statistics in Society* **1988**, *151*, 1–1.
4. Heckman, J.J.; Pinto, R. Causality and econometrics. Technical report, National Bureau of Economic Research, 2022.
5. Berzuini, C.; Dawid, P.; Bernardinell, L. *Causality: Statistical perspectives and applications*; John Wiley & Sons, 2012.
6. Schölkopf, B. Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl*; 2022; pp. 765–804.
7. Xu, G.; Duong, T.D.; Li, Q.; Liu, S.; Wang, X. Causality learning: A new perspective for interpretable machine learning. *arXiv preprint arXiv:2006.16789* **2020**.
8. Vlontzos, A.; Rueckert, D.; Kainz, B. A review of causality for learning algorithms in medical image analysis. *arXiv preprint arXiv:2206.05498* **2022**.
9. Gottlieb, G.; Halpern, C.T. A relational view of causality in normal and abnormal development. *Development and psychopathology* **2002**, *14*, 421–435.
10. Child, W. *Causality, interpretation, and the mind*; Clarendon Press, 1994.
11. Pearl, J.; Mackenzie, D. *The book of why: the new science of cause and effect*; Basic books, 2018.
12. Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **1974**, *66*, 688.
13. Pearl, J. Causal diagrams for empirical research. *Biometrika* **1995**, *82*, 669–688.
14. Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, 2023, [arXiv:cs.CV/2308.12966].
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need, 2023, [arXiv:cs.CL/1706.03762].
16. Jin, Z.; Liu, J.; Lyu, Z.; Poff, S.; Sachan, M.; Mihalcea, R.; Diab, M.; Schölkopf, B. Can Large Language Models Infer Causation from Correlation?, 2023, [arXiv:cs.CL/2306.05836].
17. Gao, J.; Ding, X.; Qin, B.; Liu, T. Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation, 2023, [arXiv:cs.CL/2305.07375].
18. Jin, Z.; Chen, Y.; Leeb, F.; Gresele, L.; Kamal, O.; Lyu, Z.; Blin, K.; Adauto, F.G.; Kleiman-Weiner, M.; Sachan, M.; et al. CLadder: Assessing Causal Reasoning in Language Models, 2024, [arXiv:cs.CL/2312.04350].
19. Zečević, M.; Willig, M.; Dhami, D.S.; Kersting, K. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal, 2023, [arXiv:cs.AI/2308.13067].
20. Watkins, C.J.; Dayan, P. Q-learning. *Machine learning* **1992**, *8*, 279–292.
21. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* **2013**.
22. Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432* **2023**.
23. Zhang, H.; Du, W.; Shan, J.; Zhou, Q.; Du, Y.; Tenenbaum, J.B.; Shu, T.; Gan, C. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485* **2023**.
24. Wang, L.; Zhang, J.; Chen, X.; Lin, Y.; Song, R.; Zhao, W.X.; Wen, J.R. Recagent: A novel simulation paradigm for recommender systems. *arXiv preprint arXiv:2306.02552* **2023**.
25. Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.
26. Argyle, L.P.; Busby, E.C.; Fulda, N.; Gubler, J.R.; Rytting, C.; Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis* **2023**, *31*, 337–351.
27. Fischer, K.A. Reflective linguistic programming (rlp): A stepping stone in socially-aware agi (socialagi). *arXiv preprint arXiv:2305.12647* **2023**.
28. Bran, A.M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A.; Schwaller, P. Augmenting large language models with chemistry tools. In Proceedings of the NeurIPS 2023 AI for Science Workshop, 2023.

29. Qian, C.; Cong, X.; Yang, C.; Chen, W.; Su, Y.; Xu, J.; Liu, Z.; Sun, M. Communicative agents for software development. *arXiv preprint arXiv:2307.07924* **2023**.
30. wiki. causal inference. Available online: https://en.wikipedia.org/wiki/Causal_inference.
31. Pearl, J. The Book of Why: The New Science of Cause and Effect. Available online: https://www.amazon.com/Book-Why-Science-Cause-Effect/dp/046509760X.
32. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D.; et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **2022**, *35*, 24824–24837.
33. Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems* **2024**, *36*.
34. Sel, B.; Al-Tawaha, A.; Khattar, V.; Wang, L.; Jia, R.; Jin, M. Algorithm of thoughts: Enhancing exploration of ideas in large language models. *arXiv preprint arXiv:2308.10379* **2023**.
35. Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* **2024**, *36*.
36. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
37. Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; Zhuang, Y. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems* **2024**, *36*.
38. Qin, Y.; Liang, S.; Ye, Y.; Zhu, K.; Yan, L.; Lu, Y.; Lin, Y.; Cong, X.; Tang, X.; Qian, B.; et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789* **2023**.
39. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* **2024**, *36*.
40. Boiko, D.A.; MacKnight, R.; Gomes, G. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332* **2023**.
41. Swan, M.; Kido, T.; Roland, E.; Santos, R.P.d. Math agents: Computational infrastructure, mathematical embedding, and genomics. *arXiv preprint arXiv:2307.02502* **2023**.
42. Mehta, N.; Teruel, M.; Sanz, P.F.; Deng, X.; Awadallah, A.H.; Kiseleva, J. Improving grounded language understanding in a collaborative environment by interacting with agents through help feedback. *arXiv preprint arXiv:2304.10750* **2023**.
43. Wu, H.; He, Z.; Zhang, X.; Yao, X.; Zheng, S.; Zheng, H.; Yu, B. Chateda: A large language model powered autonomous agent for eda. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **2024**.
44. Kovač, G.; Portelas, R.; Dominey, P.F.; Oudeyer, P.Y. The socialai school: Insights from developmental psychology towards artificial socio-cultural agents. *arXiv preprint arXiv:2307.07871* **2023**.
45. Long, S.; Schuster, T.; Piché, A.; de Montreal, U.; Research, S.; et al. Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279* **2023**.
46. Long, S.; Piché, A.; Zantedeschi, V.; Schuster, T.; Drouin, A. Causal discovery with language models as imperfect experts. *arXiv preprint arXiv:2307.02390* **2023**.
47. Kıcıman, E.; Ness, R.; Sharma, A.; Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050* **2023**.
48. Richens, J.; Everitt, T. Robust agents learn causal world models. *arXiv preprint arXiv:2402.10877* **2024**.
49. Nichani, E.; Damian, A.; Lee, J.D. How Transformers Learn Causal Structure with Gradient Descent. *arXiv preprint arXiv:2402.14735* **2024**.
50. Yao, L.; Chu, Z.; Li, S.; Li, Y.; Gao, J.; Zhang, A. A Survey on Causal Inference, 2020, [arXiv:stat.ME/2002.02770].
51. Zheng, Y.; Huang, B.; Chen, W.; Ramsey, J.; Gong, M.; Cai, R.; Shimizu, S.; Spirtes, P.; Zhang, K. Causal-learn: Causal Discovery in Python. *arXiv preprint arXiv:2307.16405* **2023**.
52. Chen, H.; Harinen, T.; Lee, J.Y.; Yung, M.; Zhao, Z. CausalML: Python Package for Causal Machine Learning, 2020, [arXiv:cs.CY/2002.11631].
53. Sharma, A.; Kiciman, E. DoWhy: An End-to-End Library for Causal Inference. *arXiv preprint arXiv:2011.04216* **2020**.
54. Wikipedia contributors. Vue.js — Wikipedia, The Free Encyclopedia, 2024. [Online; accessed 12-March-2024].
55. Rosa, G.J.; Valente, B.D.; de los Campos, G.; Wu, X.L.; Gianola, D.; Silva, M.A. Inferring causal phenotype networks using structural equation models. *Genetics Selection Evolution* **2011**, *43*, 1–13.
56. Ramsey, J.; Zhang, J.; Spirtes, P.L. Adjacency-Faithfulness and Conservative Causal Inference, 2012, [arXiv:cs.AI/1206.6843].
57. Spirtes, P.; Glymour, C.N.; Scheines, R. *Causation, prediction, and search*; MIT press, 2000.
58. Chickering, D.M. Optimal structure identification with greedy search. *Journal of machine learning research* **2002**, *3*, 507–554.
59. POHA, S.S. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* **2006**, *7*, 2006.
60. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems* **2022**, *35*, 22199–22213.
61. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* **2022**.
62. Wikipedia contributors. LangChain — Wikipedia, The Free Encyclopedia, 2024. [Online; accessed 12-March-2024].

63. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models, 2023, [arXiv:cs.CL/2210.03629].
64. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey, 2024, [arXiv:cs.CL/2312.10997].
65. Spirtes, P.; Glymour, C.; Scheines, R. *Causation, prediction, and search*; MIT press, 2001.
66. Tu, R.; Zhang, C.; Ackermann, P.; Mohan, K.; Kjellström, H.; Zhang, K. Causal discovery in the presence of missing data. In Proceedings of the The 22nd International Conference on Artificial Intelligence and Statistics. Pmlr, 2019, pp. 1762–1770.
67. Spirtes, P.L.; Meek, C.; Richardson, T.S. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983* **2013**.
68. Shimizu, S.; Hoyer, P.O.; Hyvärinen, A.; Kerminen, A.; Jordan, M. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* **2006**, *7*.
69. Shimizu, S.; Inazumi, T.; Sogawa, Y.; Hyvarinen, A.; Kawahara, Y.; Washio, T.; Hoyer, P.O.; Bollen, K.; Hoyer, P. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research-JMLR* **2011**, *12*, 1225–1248.
70. Hoyer, P.; Janzing, D.; Mooij, J.M.; Peters, J.; Schölkopf, B. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems* **2008**, *21*.
71. Yao, L.; Chu, Z.; Li, S.; Li, Y.; Gao, J.; Zhang, A. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **2021**, *15*, 1–46.
72. Winship, C.; Mare, R.D. Models for sample selection bias. *Annual review of sociology* **1992**, *18*, 327–350.
73. Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science* **2024**, *18*, 1–26.
74. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* **2015**.
75. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
76. Pearl, J.; et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress* **2000**, *19*, 3.
77. Rubin, D.B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **2005**, *100*, 322–331.
78. Cole, S.R.; Stuart, E.A. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *American journal of epidemiology* **2010**, *172*, 107–115.
79. Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; Robins, J. Double/debiased machine learning for treatment and causal parameters. Technical report, 2017.
80. Zheng, Y.; Huang, B.; Chen, W.; Ramsey, J.; Gong, M.; Cai, R.; Shimizu, S.; Spirtes, P.; Zhang, K. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research* **2024**, *25*, 1–8.
81. etc., K.B. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML, 2019. Version 0.x.
82. Fisher, R.A.; et al. 014: On the" Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. **1921**.
83. Zhang, Z.; Bo, X.; Ma, C.; Li, R.; Chen, X.; Dai, Q.; Zhu, J.; Dong, Z.; Wen, J.R. A Survey on the Memory Mechanism of Large Language Model based Agents, 2024, [arXiv:cs.AI/2404.13501].
84. Sakirin, T.; Kusuma, S. A Survey of Generative Artificial Intelligence Techniques. *Babylonian Journal of Artificial Intelligence* **2023**, *2023*, 10–14.
85. Ge, Y.; Hua, W.; Mei, K.; Tan, J.; Xu, S.; Li, Z.; Zhang, Y.; et al. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems* **2024**, *36*.
86. Rawte, V.; Sheth, A.; Das, A. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922* **2023**.
87. Adlakha, V.; BehnamGhader, P.; Lu, X.H.; Meade, N.; Reddy, S. Evaluating correctness and faithfulness of instruction-following models for question answering. *arXiv preprint arXiv:2307.16877* **2023**.
88. Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.t.; Koh, P.W.; Iyyer, M.; Zettlemoyer, L.; Hajishirzi, H. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251* **2023**.
89. Rolland, P.; Cevher, V.; Kleindessner, M.; Russel, C.; Schölkopf, B.; Janzing, D.; Locatello, F. Score matching enables causal discovery of nonlinear additive noise models, 2022, [arXiv:cs.LG/2203.04413].

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.