Causal Dataset Discovery with Large Language Models

Junfei Liu
University of Rochester
Rochester, New York, USA
jliu137@u.rochester.edu

Shaotong Sun
University of Rochester
Rochester, New York, USA
ssun25@u.rochester.edu

Fatemeh Nargesian
University of Rochester
Rochester, New York, USA
fnargesian@rochester.edu

ABSTRACT

Causal data discovery is crucial in scientific research by uncovering causal links among a variety of observed variables. Causal dataset discovery is the task of identifying datasets that contain columns that have causal relationships with columns in a query dataset. Discovering causal links from large-scale repositories faces three major challenges: vast scale of data, inherent sparsity of causal links, and incompleteness of variables present. Identifying causal relationships among datasets is a complex and time-intensive task, especially because it requires joining datasets, to bring all variables together, before applying causal link discovery. In this paper, we introduce the Causal Dataset Discovery problem and propose a large language model (LLM)-based framework to discover potential pairwise causal links between columns from different datasets. We heuristically improve LLM's grasp of causality through prompting and fine-tuning and prevent the extreme imbalance in causal candidate distributions due to natural sparsity of causal connections. We create benchmarks specific to this task[1], experimentally show that our framework achieves remarkable performance with GPT-3.5 and GPT-4. We summarize the distinctive behaviors of different LLM strategies, and discuss improvements for future research.

ACM Reference Format:
Junfei Liu, Shaotong Sun, and Fatemeh Nargesian. 2024. Causal Dataset Discovery with Large Language Models. In *Workshop on Human-In-the-Loop Data Analytics (HILDA 24), June 14, 2024, Santiago, AA, Chile.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3665939.3665968

1 INTRODUCTION

Causal discovery, crucial in scientific research, involves identifying causal relationships from observational data, as conducting randomized experiments is often not feasible. These methods, which have attracted substantial interest, usually try to deduce causality within a single dataset assuming no hidden confounders and missing variables [31]. They assume all key variables are captured for analysis, while this often fails in complex, multi-dataset environments where variables across different datasets may interact in unexpected ways, which can lead to erroneous conclusions about the nature and direction of causality when a sufficient set cannot be ensured [3]. On the other hand, in the era of big data, with its vast volume, variety, and velocity, traditional approaches are becoming less effective in grasping and understanding dataset interrelationships and selecting relevant datasets for analysis [4].

Data lakes have emerged as a solution to the challenges posed by big data [24]. While relationships such as joinability [40, 41] and unionability [25] have been extensively explored, there is a growing need to discover target datasets based on other implicit and semantic relationships with query datasets, such as causal relations [16, 37]. Discovering causal relations can assist in identifying causally relevant datasets within a data repository. Firstly, it enhances targeted dataset browsing for specific needs on causally linked tables. Augmenting a query dataset with causally linked tables could potentially yield more robust features for model training [28]. Additionally, the identified datasets become ideal candidates for augmenting knowledge bases. Lastly, causal relevance among datasets contributes to a better understanding of underlying structures and dependencies in complex data environments, which supports more informed decision-making and efficient data management strategies.

Maintaining numerical correspondence between the columns found with causal relations is crucial for discovering relationships among datasets in a data-driven fashion. One effective strategy for achieving this is through table joins based on common columns, which can ensure that variables from different datasets are aligned in a way that preserves the integrity of the relationships, including causality. However, join operations pose substantial computational costs, which intensifies with the volume of data and the need for precise variable alignment across numerous datasets, making pairwise or outer joins computationally expensive. There has been research on searching joinable tables and hashing schemes for efficient dataset search. For example, Santos et al. introduced a novel hashing scheme for identifying the top-k tables joinable with a query table efficiently [30], based on numerical data correlations, using a sketch-based index, which we adapt in this paper.

Large language models (LLMs) are emerging as powerful tools in various research areas, showing particular potential in causal discovery and inference [18]. Recent advancements in LLMs have demonstrated their ability to comprehend and generate complex textual content, suggesting their potential for uncovering and reasoning about causal relationships in natural language [2]. These models demonstrate promising performance across various causal tasks, including pairwise causal discovery and actual causality determination [18, 19]. This capability opens new avenues for causal dataset discovery, particularly in scenarios where traditional statistical methods may be limited by their assumptions.
In this study, we make the following contributions:

- We present the causal dataset discovery problem in data lakes.

[1] The Chicago Causal Link Discovery Benchmark is open-source and can be accessed through https://github.com/JeffLiu114514/Chicago-Causal-Link-Discovery-Benchmark.

HILDA 24 , June 14, 2024, Santiago, AA, Chile

ACM ISBN 979-8-4007-0693-6/24/06
https://doi.org/10.1145/3665939.3665968

- We propose a novel join-based technique for discovering potential causal links over join across datasets. Our technique leverages the power of LLMs in three stages of prompting and fine-tuning.
- We create an open-source benchmark designed to evaluate causal dataset discovery techniques. We evaluate our technique over this benchmark.
- We discuss findings on different LLM strategies' distinctive behaviors, summarize the limitations of this study, and suggest potential improvements.

Discovering causal links among tables has the potential to enhance the dataset browsing experience by identifying potentially causally linked datasets [28], increase the robustness of feature engineering by augmenting a query dataset with causally linked features, and allow users to make informed decisions by providing an understanding of underlying structures and dependencies in complex data environments.

2 PROBLEM DEFINITION

Let $T_Q(X, K_X, \ldots)$ and $T_C(Y, K_Y, \ldots)$ be two tables in a data lake with numerical columns $X \in T_Q$ and $Y \in T_C$ and categorical columns $K_X \in T_Q$ and $K_Y \in T_C$. We define candidate pairs of columns X, Y on the join of T_C and T_Q as follows.

$$\{(X, Y) | X \in T_Q, Y \in T_C, \exists K_X, K_Y, T_Q \bowtie_{K_X=K_Y} T_C\}$$

Definition 2.1 (Correlation Link over Join). Candidate columns X and Y have a correlation link over $T_Q \bowtie_{K_X=K_Y} T_C$, if their correlation after the join is higher than a threshold C.

As discussed in the correlation link discovery (section 3.1.3), we use mutual information as the correlation measure to effectively identify the pairs of columns that are likely to have potential causal links.

Definition 2.2 (Potential Causal Link over Join). For a given candidate pair (X, Y) a potential causal link over join $T_Q \bowtie_{K_X=K_Y} T_C$ exists if 1) there exists a correlation link between X and Y in the joined table and 2) post the application of causal inference algorithms, the link between X and Y is confirmed as causal, with its direction being either clearly established or not determined.

The outcome from a correlation link $(T_Q\langle K_X, X\rangle, T_C\langle K_Y, Y\rangle)$ can be one of the following:

A) $T_Q\langle K_X, X\rangle$ causes $T_C\langle K_Y, Y\rangle)$
B) $T_C\langle K_Y, Y\rangle$ causes $T_Q\langle K_X, X\rangle$
C) No causal relation exists between $T_Q\langle K_X, X\rangle$ and $T_C\langle K_Y, Y\rangle)$

We define two tables T_Q and T_C causally linked if there exists at least one pair of potentially causal link between columns $\langle X, Y\rangle$ from tables T_Q and T_C after join.

Definition 2.3 (Causal Dataset Discovery Problem). Given a query dataset T_Q and a data lake of datasets $\mathcal{L}$, find all datasets $T_C \in \mathcal{L}$ such that T_C and T_Q have at least one causal link.

The desired output of causal dataset discovery is a list of datasets T_C, each containing a list of column-level potential causal links. Note that dataset discovery is often formulated as a top-K [22, 40] or threshold-based search problem [41]. However, we consider the problem of finding all potentially causal links between a data lake and a query dataset. This is because the degree of causality is not necessarily quantifiable.

3 CAUSAL DATASET DISCOVERY

Our hypothesis is that large language models (LLMs) possess the potential to identify causal relationships between variables thanks to extensive training across diverse texts, including domain-specific knowledge and scientific literature [18]. Nonetheless, the challenges of time efficiency and computational demands are significant, especially given the sparsity of causal relations and the skewed distribution between positive and negative causal pairs in large data lakes [39]. Moreover, LLMs may lack the nuanced understanding of statistical and causal inference principles necessary for robust causal analysis [39]. They may struggle to distinguish between correlation and causation due to reliance on patterns in their training data rather than underlying causal mechanisms [17]. Our methodology is mainly designed to tackle these challenges. To discover the causal links, we propose a filter-verification technique. First, we minimize the number of causal candidates as input to LLMs by filtering the table pairs that are not joinable, the duplicate candidate pairs, as well as candidate pairs with low mutual information. Afterward, we verify the remaining candidates using LLMs. We employ appropriate prompting and fine-tuning to enhance LLM's understanding of causality to strengthen its performance on causal tasks.

3.1 Candidates Generation and Screening

We divide the causal pair candidates generation and screening procedure into three steps: 1) candidate pairs generation via hash join and filtering via joinability, which produces raw candidate pairs from the input data repository; 2) duplicate pairs filtering via index-based vector similarity search on pairs, which extracts unique candidate pairs from raw candidate pairs; 3) correlation link discovery through filtering by a correlation coefficient threshold, which discovers correlation links from candidate pairs.

3.1.1 Candidate Generation. Within the context of our problem scenario, the input is a repository of datasets, which defines the search space of potential causal pairs to be all combinations of two columns from different tables given they are joinable. Afterward, we exhaustively perform hash joins on every combination of categorical columns from two tables. If two tables are joinable on some columns, then all combinations of numerical columns from the two tables are considered as the raw candidate pairs. We deal with many-to-many joins by grouping tuples based on the join column values and aggregating tuples within each group by the `MEAN` operator.

3.1.2 Duplicate Filtering. The goal of this step is to refine the pool of candidate pairs by eliminating instances that represent highly similar variables, thus preventing the redundancy of nearly identical pairs in the analysis. The potential overlap of variables may convey analogous or duplicate information. For example, the column `STREET NUMBER` from datasets `CDPH_Storage_Tanks.csv` could form a candidate pair with both `Community Area Number` and `Community Area` from dataset `Chicago_Energy_Benchmarking_-_Covered_Buildings.csv`. This may mislead downstream tasks such as top-K causal dataset

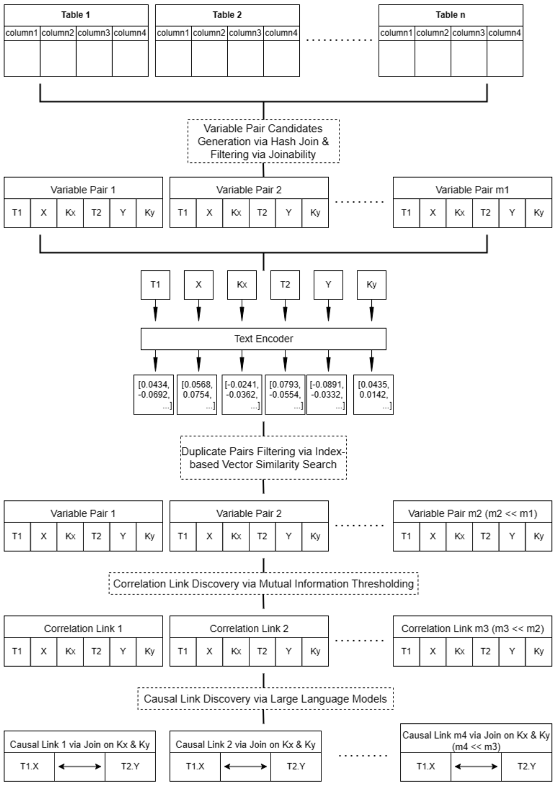
Table 1
Table 2
Table n
column1
column2
column3
column4
Variable Pair Candidates Generation via Hash Join & Filtering via Joinability
Variable Pair 1
Variable Pair 2
Variable Pair m1
T1
X
Kx
T2
Y
Ky
Text Encoder
[0.0434, -0.0692, ...]
[0.0568, 0.0754, ...]
[-0.0241, -0.0362, ...]
[0.0793, -0.0554, ...]
[-0.0891, -0.0332, ...]
[0.0435, 0.0142, ...]
Duplicate Pairs Filtering via Index-based Vector Similarity Search
Variable Pair 1
Variable Pair 2
Variable Pair m2 (m2 << m1)
Correlation Link Discovery via Mutual Information Thresholding
Correlation Link 1
Correlation Link 2
Correlation Link m3 (m3 << m2)
Causal Link Discovery via Large Language Models
Causal Link 1 via Join on Kx & Ky
Causal Link 2 via Join on Kx & Ky
Causal Link m4 via Join on Kx & Ky (m4 << m3)
T1.X
T2.Y

Figure 1: Pipeline Visualization for Causal Dataset Discovery

search. When the table `CDPH_Storage_Tanks.csv` is queried, the duplicate causal links increase unnecessary weights on the table `Chicago_Energy_Benchmarking_-_Covered_Buildings.csv`, which may undermine the ranking's effectiveness and cause more causally related datasets undiscovered. To address this concern, we apply an index-based vector similarity search on variable pairs' text embeddings to reduce the number of such duplicate pairs. We first utilize pre-trained text encoders to transform the information of each candidate pair into vector representations [5]. In our implementation, this transformation includes metadata encoding of the the names of the join (categorical) columns, and the names of the tables. Afterwards, our pipeline constructs three indices [10], one for each category of encoded vectors: numerical column names, categorical column names, and table names. These indices serve as the foundation for conducting fast approximate nearest neighbor (ANN) search. We compare each candidate pair against others within these indices: a candidate pair is deemed redundant and consequently removed from consideration if another pair exhibits high similarity across all three categories.

3.1.3 Correlation Link Discovery. The duplicate filtering step significantly narrows down the pool of candidate pairs, ensuring a focused selection of candidate pairs. The objective of this phase is to further narrow down the search by using an effective causality indicator to reduce negative pairs while retaining positive causations, thereby reducing candidates of causal links with high recall.

Correlation measures like the Pearson, Kendall's and Spearman's rank coefficients [38] quantify relationships between variables and offer insights toward causality but are limited by their need for linear or monotonic relationships or assumptions on the observational variable like normal distribution. This is problematic in cases such as the Yerkes-Dodson Law [36]. Consider the stress level as cause and performance as effect. The increase in stress level can result in first increase then decrease of performance, showing their nonlinearity, while both stress levels and cognitive performance in this scenario normally exhibit skewed and non-normal distributions.

Differing from Pearson and Spearman's rank correlation coefficients, mutual information is a measure of mutual dependence that quantifies the amount of information obtained about one variable through the other. Although Pearson correlation has been used in the existing works [13], due to sensitivity to all types of dependencies and flexibility with respect to data distributions, we believe mutual information is a more meaningful indicator of causality. Therefore, we discover the correlation links over join from all candidate pairs through a mutual information threshold to filter out candidate pairs that likely indicate a weaker or non-existent causal relationship, which empirically refines the distribution of causal relations in candidate pairs.

3.1.4 Optimized Alternative. In previous sections, we have discussed a naive pipeline for generating the correlation links over join. We noted that exhaustive table joins are extremely computationally expensive. As a result, we adapted and modified the idea of correlation sketches, by Santos et al. [30], for finding approximate join-correlations to replace table joins and correlation calculation in steps 3.1.1 and 3.1.3.

Given two tables aggregated on join columns T_X and T_Y as input, Santos et al.'s approach utilizes index sketching to emulate the correlation of the result of aggregated join. In summary, it first pairs categorical columns K_X, K_Y with a numerical column X, Y at random from aggregated table T_X, T_Y, respectively. Then, it hashes all values in the aggregated categorical column into integers without collision and then hashes integers again into real numbers, which maps distinct values randomly and uniformly to the unit interval [0, 1]. By using the bottom-K of hashed values, the correlation sketch resolves the alignment issue of the emulated joined table between the pairs and randomly takes k samples for correlation estimation. Therefore, the output would be an efficient estimation of the correlation $r_{X \bowtie Y}$ of the numerical attributes $X_{X \bowtie Y}$ and $Y_{X \bowtie Y}$ in $T_{X \bowtie Y}$ without having to compute the full join for T_X and T_Y.

3.2 Causal Link Discovery using Large Language Models

The objective of causal link discovery will be identifying the underlying potential causal relation from the correlation links obtained from the previous steps. The problem becomes a natural language classification problem if a causal direction is explicitly identified, as introduced in the definition 2.2. Because downstream tasks including causal dataset discovery problem do not require explicit directions of causal links, the outcome can be a binary response on whether there is a causal relation between $T_Q\langle K_X, X\rangle$ and $T_C\langle K_Y, Y\rangle$ only.

Common practices for improving LLMs' capabilities on highly specialized tasks include prompting and fine-tuning.

Table 1: Benchmark Statistics

	Benchmark#1	Benchmark#2	Benchmark#3
# of source tables	20	40	64
total # of pairs w/o duplicates	1095	1265	1838
# of pairs pass MI threshold	312	405	668
# of Positive causal relations	61	12	78

Prompting, also known as in-context learning, involves providing a carefully designed input to influence LLMs' output without making changes to the model's internal parameters, which is highly sensitive to the choice of examples, order, and the prompt format to retrieve the pre-trained knowledge of the model [20]. Role-prompting assigns specific roles, such as a domain-specific assistant, to LLMs as context [8]; Socratic prompting, involving the inclusion of definitions within prompts, aims to refine alignment with the intended task. These two prompting techniques encourage the model to utilize a targeted knowledge base and produce goal-aligned responses [7] and will serve as the default prompts in this study unless otherwise specified. Chain of Thought (CoT) aids multi-step reasoning by generating intermediate steps before reaching conclusions [35]. Ban et al. introduce a three-stage method for identifying causal statements with LLMs: using role-prompting for generating definitions, CoT for concluding causal relations, and error revision for verifying accuracy, addressing the complexities in identifying causal relationships [2]. Prompting templates used in this study can be found in Appendix A.1.

Fine-tuning, on the other hand, involves adjusting the model's internal parameters through additional training on a specific dataset tailored to a particular task or domain. We create the training dataset with tailored class distribution to balance among three classes, which prevents overfitting due to the skewed distribution of positive and negative causal relations, and use different benchmarks created from a disjoint set of tables separately for training and testing to prevent data leakage, where the specific dataset information can be found in Section 4.2.

4 EVALUATION

We performed extensive experiments on real-world datasets to empirically evaluate the effectiveness of our causal dataset discovery method.

4.1 Chicago Causal Link Discovery Benchmark

While there exist benchmarks for dataset discovery [32] and benchmarks for causal inference between two variables [15, 21] and for causal graph discovery within a single dataset [11, 34], to the best of our knowledge, there exists no benchmarks for causal dataset discovery problem across different tables via join.Therefeore, to evaluate algorithms for Causal Dataset Discovery Problem, we prepare a benchmark with ground truths of underlying causal links via join among columns from distinct tables.

We create a benchmark named *Chicago Causal Link Discovery* from the Chicago Data Portal [1] for evaluating the causal dataset discovery problem. Three micro-benchmarks are created independently with a disjoint set of tables. Benchmark#1 includes 20 relatively large tables, each containing on average 34 columns and 46,963 rows, designed to simulate intensive datasets with high joinability and dense potential causal links; benchmark#2 includes 40 smaller tables with on average 12 columns and 14,993 rows, designed to simulate datasets with small tables thus low joinability and sparse potential causal links; benchmark#3, the largest datasets, includes 64 tables, which contains all tables in benchmarks#1 and #2, with on average 22 columns and 42,316 rows. All three micro-benchmarks cover all public data categories to comprehensively represent diverse data sources in large data repositories.

To minimize the effort of human labellers, we apply the steps described in 3.1 to reduce the sheer number of column pairs to be annotated. To get exact join results and correlation scores over join for our benchmarks, we execute joins via hash join instead of resorting to our adaptation of correlation-join index. Performing joins on the categorical columns of all benchmark datasets and enumerating all candidate pairs yield 10041, 10490, and 13650 pairs for benchmark#1, #2, and #3, respectively. Through ANN searches among text embeddings of column metadata with a cosine similarity threshold of 0.9, we eliminate 89.09%, 87.94%, and 85.93% of these candidate pairs as highly similar duplicates. The remaining 1095, 1265, and 1838 pairs constitute the three micro-benchmarks. For each candidate pair in the benchmark, we create labels for the benchmark through a majority voting of two human labellers and the prompting result of GPT-4 [26].

4.2 Implementation and Setup

Within the duplicates filtering step, to generate text embeddings, we leverage FastText [5]. It excels in capturing the nuances of text data by considering subword information, which enables it to effectively understand and represent the semantics of column and table names, even when dealing with abbreviations, acronyms, or words not seen during training. To implement similarity search efficiently, we employ Faiss [10], an optimized library designed specifically for similarity search and clustering of dense vectors, which enables the high-speed execution of ANN searches across a large amount of candidate pairs. The large language model involved in this study is implemented through API requests to GPT models provided by OpenAI, specifically gpt-3.5-turbo-0125 when refered as GPT-3.5 [6] and gpt-4-0125-preview when referred as GPT-4 [26].

We conduct prompting and finetuning experiments and present results on benchmarks#1, #2, and #3. Prompting tests include default, Chain of Thought, and three-stage prompts introduced in Section 3.2 and specific prompts can be found in Appendix A.1. To avoid data leakage, fine-tuned models' performance on benchmarks #1 or #2 are tested on the other benchmark to ensure that training and testing data do not overlap, i.e., the test results of GPT-3.5 fine-tuned model shown in benchmark#2 section is trained on benchmark #1 and vice versa. The model fine-tuned on benchmark #3 is trained and tested through 8-2 split of the entire benchmark, which may include overlapping variables but no identical causal candidate pairs in train and test datasets. The training set's distribution is tailored such that each class's percentage does not exceed the sum of the other two classes to ensure that choice C (No causal relation exists) does not dominate and cause overfitting.

4.3 Experimental Results

We empirically find the best mutual information (MI) threshold C by searching over a threshold space with minimum granularity γ

Table 2: Causal Links Discovery Results

	Benchmark#1				Benchmark#2				Benchmark#3			
Measurement	**precision**	**recall**	**accuracy**	**F-1 Score**	**precision**	**recall**	**accuracy**	**F-1 Score**	**precision**	**recall**	**accuracy**	**F-1 Score**
GPT-3.5 + default prompts	73.97	75.27	74.26	74.26	41.64	52.50	66.67	66.67	59.37	70.16	68.00	68.00
GPT-3.5 + CoT prompts	54.18	54.46	41.58	41.58	36.41	29.79	23.53	23.53	40.67	45.45	27.00	27.00
GPT-3.5 + 3-stage prompts	85.54	75.14	80.20	81.00	77.54	49.52	80.39	82.00	84.69	56.11	78.00	80.41
GPT-3.5 finetuned	81.43	78.68	80.20	80.20	75.16	66.76	86.27	86.27	84.11	75.73	88.00	88.00
GPT-4 + default prompts	**97.08**	**93.07**	**95.05**	**95.05**	**91.71**	**83.81**	**94.12**	**94.12**	**97.75**	**88.94**	**95.00**	**95.00**

Table 3: Causal Links Identification Results

	Benchmark#1				Benchmark#2				Benchmark#3			
Measurement	**precision**	**recall**	**accuracy**	**F-1 Score**	**precision**	**recall**	**accuracy**	**F-1 Score**	**precision**	**recall**	**accuracy**	**F-1 Score**
GPT-3.5 + default prompts	80.0	89.80	84.16	84.62	47.62	83.33	74.51	60.61	58.00	**93.55**	77.00	71.60
GPT-3.5 + CoT prompts	52.75	**97.96**	56.44	68.57	25.58	**91.67**	35.29	40.00	32.95	**93.55**	39.00	48.74
GPT-3.5 + 3-stage prompts	**100**	66.67	84.00	80.00	**100**	33.33	84.00	50.00	**100**	37.93	81.44	55.00
GPT-3.5 finetuned	90.70	79.59	86.14	84.78	88.89	66.67	90.20	76.19	95.83	74.19	91.00	83.64
GPT-4 + default prompts	**100**	89.80	**95.05**	**94.62**	**100**	83.33	**96.08**	**90.91**	**100**	83.87	**95.00**	**91.23**

with the highest F-1 score in terms of finding positive causal links among candidate pairs. The optimal MI threshold is at 0.501 where correlation link discovery extracts 312, 405, and 668 correlation links from 1095, 1265, and 1838 pairs while maintaining a high recall of 80.32%, 100.00%, and 83.56% for three benchmarks, respectively.

We present the experiments of causal link discovery with separate criteria. Table 2 shows the average precision, recall, accuracy, and F-1 score for the 3-class causal discovery problem which requires the causal direction being explicitly determined, while table 3 shows the results for causal determination, as introduced in Section 3.2. The observation is that GPT-4 performs dominantly in 3-class classification, which indicates its capability at both identifying causal links and directions. GPT-4's superior performance on most academic benchmarks has shown its comprehensive capabilities [26], which explains its higher proficiency at causal tasks. Chain of thoughts (CoT) gives strong bias to respond with positive causal links but is very inaccurate; 3-stage approach performs in an opposite manner against CoT as it is precise, especially at identifying causal links, while giving low recall. This sharp comparison is mainly contributed by the self-revision step in the 3-stage approach. We speculate that an ensemble of CoT and 3-stage approach may lead to better overall performance. Finally, fine-tuning gives the best general performance among all GPT-3 based approaches. Given non-overlapping training and testing data with distinctive data distribution, we believe the boost in performance through fine-tuning demonstrates LLMs' potential to learn and generalize causality.

Causal links discovered are then used for testing the performance downstream tasks, the causal dataset discovery, where results can be found in Table 4. We evaluate the average precision and recall of each tables' query results given the causal links discovered within the benchmark. The general observation aligns with causal link discovery results but with more polarized performance, which is because causal connections between tables are naturally skewed. Large amount of causal links could concentrate between specific table pairs, for example, table `Selected Socioeconomic Indicators` have 38 causal links with `Selected Public Health Indicators` in benchmark #1, which is already more than half of the positive causal relations in the benchmark, while most table pairs are not causally linked. Returning tables when there is no causally linked tables to the query table or failing to return causally linked tables will largely affect the precision and recall.

5 RELATED WORK

Dataset Discovery Dataset discovery identifies datasets that meet certain informational needs, supported by tools that improve search and navigation effectiveness and scalability[27]. Dataset organization and discovery are essential for efficient data navigation, retrieval, and analysis, which involves techniques for finding joinable, or semantically similar datasets with approaches ranging from metadata analysis to multi-dimensional similarities. Relevant dataset discovery includes DeepJoin[9] which offers an embedding-based method using Pre-trained Language Models and ANN Search to efficiently locate joinable datasets[14], and data lake organization defined by Nargesian et al. as identifying optimal structure for efficiently locating the required dataset within a data lake[23].

Causal Discovery Causal discovery aims to uncover causal relationships from observational data to create Directed Graphical Causal Models (DGCMs). These models focus on d-separation and Markov Equivalence Classes (MECs) essential for defining the Causal Faithfulness Assumption[12]. Most causal discovery methods focus on causal relations within a single dataset. The foundational PC Algorithm assumes i.i.d. sampling and iterates to reduce from a fully-connected graph to a causal DAG's MECs, with its correctness hinging on the causal Markov and faithfulness conditions. The FCI Algorithm, extending PC's framework, accommodates undirected edges and latent confounders. GES, in contrast, starts with no connections and incrementally adds edges based on metrics like Bayesian information criterion(BIC), mapping final models to their MECs.[12][31]. Huang et al. introduced CD-MiNi, a method for

Table 4: Causal Dataset Discovery Results

	Benchmark#1		**Benchmark#2**		**Benchmark#3**	
Measurement	**avg precision**	**avg recall**	**avg precision**	**avg recall**	**avg precision**	**avg recall**
GPT-3.5 + default prompts	50.83	81.25	43.89	76.67	65.71	83.33
GPT-3.5 + CoT prompts	37.69	**100**	32.89	**100**	53.33	**100**
GPT-3.5 + 3-stage prompts	**100**	43.75	**100**	38.33	**100**	12.50
GPT-3.5 finetuned	**100**	43.75	82.14	55.00	**100**	12.50
GPT-4 + default prompts	**100**	62.50	**100**	85.00	**100**	58.33

identifying causal relationships over the complete set of variables that are non-identical from multiple datasets using two estimation approaches, showed theoretical identifiability of causal structures, and extended applications to confounding and cyclic scenarios[16]. Salimi et al. proposed a declarative language for causal queries in relational domains, which involves constructing causal DAGs without exogenous variables[29].

Causality and LLMs Advances in LLMs have notably improved their performance on generating viable responses and insights from textual schemas.[18] Immanuel Trummer's work shows that even smaller LLMs can accurately predict data correlations using just column names, applicable across various data correlation metrics[33]. Kıcıman et al. highlight the potential of LLMs in causal reasoning, showing their effectiveness in pairwise causal discovery tasks, suggesting their use to complement traditional causal analysis[18]. Long et al. see LLMs as imperfect experts in causal discovery, demonstrating how LLMs can refine outputs of causal discovery algorithms, though results vary by dataset and model[19]. However, it has been arguable regarding LLMs' capabilities on causal tasks. Jin et al. challenge the capability of LLMs in causal tasks, showing through the CORR2CAUSE task that even fine-tuned LLMs struggle with causal inference from correlation, generally failing in out-of-distribution generalization[17]. Similarly, Zečević et al. argue LLMs lack genuine reasoning in causal tasks and only mimic training data[39].

6 DISCUSSION AND FUTURE WORK

The current study, while pioneering in its approach to discovering causal links over join in complex datasets, encounters several limitations that could impact the robustness and applicability of its findings.

A primary challenge arises from the size of the Chicago Causal Link Discovery Benchmark, where the construction of this benchmark is an intensive process due to the inherent rarity of causal relations and therefore the intensive manual labeling efforts at identifying positive causal links. Given the constraints of time and resources, our ability to develop a more expansive benchmark, enriched with high-quality labels, was restricted. This limitation is particularly pertinent as we envision the application of our methodology in expansive data repositories such as data lakes, where the scope and scale of data significantly exceed the benchmark's current capacity. Future work should, therefore, focus on expanding the benchmark to better represent the vastness and variety of real-world data, enhancing the generalizability and effectiveness in production environments.

Moreover, an alternative data-driven method is desired as there is an absence of suitable approaches to supplement LLMs for causal relation identification. Determining causal relations among enormous datasets across various fields will inevitably involve important missing variables, It is anticipated to have crucial covariates missing from observational datasets across various fields, and thus typical statistical measures are unable to take place without strong assumptions on missing variables, usually by disregarding them. Besides, data-driven methods are likely to conclude that similar or identical variables are causally linked, for example, "% of employees with Debt" could be causally linked with "# of employees with Debt", where we do not wish to see such redundancy. In addition, it may also be arguable to include column values in the prompts such that LLMs respond with reference to data. However, we empirically find LLMs perform worse with a sample of column values, but it is open for future study to see if integrating column values with other strategies could work better.

Meanwhile, we currently use exhaustive iteration to find joinable tables by attempting to join every combination of tables on every categorical column, while algorithms that discover joinable tables could potentially simplify this process by efficiently filtering joinable columns and tables. Besides, our methodology focuses on numerical column pairs joined over categorical columns due to tractability concerns, but it could be extended to include categorical columns and use numerical columns for joining with advanced optimization techniques.

Last but not least, the effectiveness of LLMs in understanding causality is debated, as they may overly rely on causal facts from training data without a nuanced grasp of causality[39], possibly limiting their generalization in real-world causal dataset discovery tasks. Potential substitutes to LLMs may include encoder models for semantics and contextual understanding and a classifier for causal link determination, which also benefits from lower computational costs. Future research is encouraged to explore the capabilities of LLMs for causality more deeply and the possibilities of alternative approaches.

REFERENCES

[1] 2024. Chicago Data Portal. https://data.cityofchicago.org/

[2] Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023. From Query Tools to Causal Architects: Harnessing Large Language Models for Advanced Causal Discovery from Data. arXiv:2306.16902 [cs.AI]

[3] Elias Bareinboim and Judea Pearl. 2016. Causal inference and the data-fusion problem. https://doi.org/10.1073/pnas.1510507113

[4] Alex Bogatu, Alvaro A. A. Fernandes, Norman W. Paton, and Nikolaos Konstantinou. 2020. Dataset Discovery in Data Lakes. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE. https://doi.org/10.1109/icde48307.2020.00067

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[7] Edward Y. Chang. 2023. Prompting Large Language Models With the Socratic Method. arXiv:2303.08769 [cs.LG]

[8] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. arXiv:2310.14735 [cs.CL]

[9] Yuyang Dong, Chuan Xiao, Takuma Nozawa, Masafumi Enomoto, and Masafumi Oyamada. 2023. DeepJoin: Joinable Table Discovery with Pre-trained Language Models. arXiv:2212.07588 [cs.DB]

[10] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). arXiv:2401.08281 [cs.LG]

[11] Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang. 2022. Deep End-to-end Causal Inference. *arXiv preprint arXiv:2202.02195* (2022).

[12] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics* 10 (2019). https://doi.org/10.3389/fgene.2019.00524

[13] Yue Gong, Sainyam Galhotra, and Raul Castro Fernandez. 2023. Nexus: Correlation Discovery over Collections of Spatio-Temporal Tabular Data.

[14] Rihan Hai, Christos Koutras, Christoph Quix, and Matthias Jarke. 2023. Data Lakes: A Survey of Functions and Systems. *IEEE Transactions on Knowledge and Data Engineering* 35, 12 (Dec. 2023), 12571–12590. https://doi.org/10.1109/tkde.2023.3270101

[15] Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. CauseNet: Towards a Causality Graph Extracted from the Web. In *CIKM*. ACM.

[16] Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. 2020. Causal Discovery from Multiple Data Sets with Non-Identical Variable Sets. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 06 (Apr. 2020), 10153–10161. https://doi.org/10.1609/aaai.v34i06.6575

[17] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023. Can Large Language Models Infer Causation from Correlation? arXiv:2306.05836 [cs.CL]

[18] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. arXiv:2305.00050 [cs.AI]

[19] Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. 2023. Causal Discovery with Language Models as Imperfect Experts. arXiv:2307.02390 [cs.AI]

[20] Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided Few-shot Prompting for Large Language Models. arXiv:2303.13217 [cs.CL]

[21] Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2015. Distinguishing cause from effect using observational data: methods and benchmarks. arXiv:1412.3773 [cs.LG]

[22] Pranay Mundra, Jianhao Zhang, Fatemeh Nargesian, and Nikolaus Augsten. 2023. KOIOS: Top-k Semantic Overlap Set Search. arXiv:2304.10572 [cs.DB]

[23] Fatemeh Nargesian, Ken Q. Pu, Bahar Ghadiri Bashardoost, Erkang Zhu, and Renée J. Miller. 2020. Data Lake Organization. arXiv:1812.07024 [cs.DB]

[24] Fatemeh Nargesian, Erkang Zhu, Renée J. Miller, Ken Q. Pu, and Patricia C. Arocena. 2019. Data lake management: challenges and opportunities. *Proc. VLDB Endow.* 12, 12 (aug 2019), 1986–1989. https://doi.org/10.14778/3352063.3352116

[25] Fatemeh Nargesian, Erkang Zhu, Ken Q. Pu, and Renée J. Miller. 2018. Table union search on open data. *Proc. VLDB Endow.* 11, 7 (mar 2018), 813–825. https://doi.org/10.14778/3192965.3192973

[26] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[27] Norman W. Paton, Jiaoyan Chen, and Zhenyu Wu. 2023. Dataset Discovery and Exploration: A Survey. *ACM Comput. Surv.* 56, 4, Article 102 (nov 2023), 37 pages. https://doi.org/10.1145/3626521

[28] Audrey Poinsot and Alessandro Leite. 2023. A Guide for Practical Use of ADMG Causal Data Augmentation. arXiv:2304.01237 [cs.LG]

[29] Babak Salimi, Harsh Parikh, Moe Kayali, Sudeepa Roy, Lise Getoor, and Dan Suciu. 2020. Causal Relational Learning. arXiv:2004.03644 [cs.DB]

[30] Aécio Santos, Aline Bessa, Fernando Chirigati, Christopher Musco, and Juliana Freire. 2021. Correlation Sketches for Approximate Join-Correlation Queries. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD/PODS '21)*. ACM. https://doi.org/10.1145/3448016.3458456

[31] Peter Spirtes and Richard Scheines. 1993. *Causation, Prediction, and Search*. Vol. 81. https://doi.org/10.1007/978-1-4612-2748-9

[32] Kavitha Srinivas, Julian Dolby, Ibrahim Abdelaziz, Oktie Hassanzadeh, Harsha Kokel, Aamod Khatiwada, Tejaswini Pedapati, Subhajit Chaudhury, and Horst Samulowitz. 2023. LakeBench: Benchmarks for Data Discovery over Data Lakes. arXiv:2307.04217 [cs.DB]

[33] Immanuel Trummer. 2023. Can Deep Neural Networks Predict Data Correlations from Column Names? arXiv:2107.04553 [cs.DB]

[34] Ruibo Tu, Kun Zhang, Bo Christer Bertilson, Hedvig Kjellström, and Cheng Zhang. 2019. Neuropathic Pain Diagnosis Simulator for Causal Discovery Algorithm Evaluation. arXiv:1906.01732 [cs.LG]

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL]

[36] Robert M Yerkes and John D Dodson. 1908. The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology and psychology* 18, 5 (1908), 459–482.

[37] Brit Youngmann, Michael Cafarella, Babak Salimi, and Anna Zeng. 2023. Causal Data Integration. arXiv:2305.08741 [cs.DB]

[38] Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of Biostatistics* 7 (2005).

[39] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. arXiv:2308.13067 [cs.AI]

[40] Erkang Zhu, Dong Deng, Fatemeh Nargesian, and Renée J. Miller. 2019. JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes. In *Proceedings of the 2019 International Conference on Management of Data* (Amsterdam, Netherlands) *(SIGMOD '19)*. Association for Computing Machinery, New York, NY, USA, 847–864. https://doi.org/10.1145/3299869.3300065

[41] Erkang Zhu, Fatemeh Nargesian, Ken Q. Pu, and Renée J. Miller. 2016. LSH ensemble: internet-scale domain search. *Proc. VLDB Endow.* 9, 12 (aug 2016), 1185–1196. https://doi.org/10.14778/2994509.2994534

A APPENDIX

A.1 LLM Prompting Templates

Default Prompts with Role-prompting and Definitions

{
"messages": [
{
"role": "system",
"content": "You are a helpful assistant for causal reasoning. Include answer A, B, or C within the tags <Answer>A/B/C</Answer>."
},
{
"role": "user",
"content": "Causality is an influence by which one event, process, state, or object contributes to the production of another event, process, state, or object where the cause is partly responsible for the effect, and the effect is partly dependent on the cause.Choose the correct causal relation between {var1} and {var2}: A. {var1} causes {var2}; B. {var2} causes {var1}; C. no causal relation between {var1} and {var2}." }] }

Chain of Thought(CoT) prompt

{
"messages": [
{
"role": "system",
"content": "You are a helpful assistant for causal reasoning. Include answer A, B, or C within the tags <Answer>A/B/C</Answer>."
},
{
"role": "user",
"content": "Causality is an influence by which one event, process, state, or object contributes to the production of another event, process, state, or object where the cause is partly responsible for the effect, and the effect is partly dependent on the cause.Choose the correct causal relation between {var1} and {var2}: A. {var1} causes {var2}; B. {var2} causes {var1}; C. no causal relation between {var1} and {var2}.
Let's work this out in a step by step way to be sure that we have the right answer. Provide your final answer within the tags <Answer>A/B/C</Answer>." }] }

Three-stage prompt

{
"messages": [
{
"role": "system",
"content": "You are an expert in analyzing public data."
},
{
"role": "user",
"content": "You are investigating the cause-and-effect relationships between {var1} and {var2}. Please understand the real meaning of each variable, and explain them in order." },
{
"role": "assistant",
"content": {step-1 response}
},
{
"role": "user",
"content": "Causality is an influence by which one event, process, state, or object contributes to the production of another event, process, state, or object where the cause is partly responsible for the effect, and the effect is partly dependent on the cause. Based on {step-1 response}, analyze the cause-and-effect relationships between them. Choose the correct causal relation between {var1} and {var2}: A. {var1} causes {var2}; B. {var2} causes {var1}; C. no causal relation between {var1} and {var2}.
Let's work this out in a step by step way to be sure that we have the right answer. Provide your final answer within the tags <Answer>A/B/C</Answer>." },
{
"role": "assistant",
"content": {step-2 response}
},
{
"role": "user",
"content": "Based on your explanation, check whether the causal statement {step-2 response} between A and B is correct, and give the reasons. If you believe the causal statement is true, respond with <Answer>True</Answer>. If you believe the causal statement is false, analyze the cause-and-effect relationships between them: Which cause-and-effect relationship is more likely?" },
] }