Journal of Machine Learning Research 27 (2026) 1-95 Submitted 9/25; Revised 8/26; Published 10/26

Discovering and Reasoning of Causality in the Hidden World with Large Language Models

Chenxi Liu[*] [1] CSCXLIU@COMP.HKBU.EDU.HK
Yongqiang Chen[*] [2, 3] YQCHEN24@GMAIL.COM
Tongliang Liu [4, 2] TONGLIANG.LIU@SYDNEY.EDU.AU
Mingming Gong [5, 2] MINGMING.GONG@UNIMELB.EDU.AU
James Cheng [6] JCHENG@CSE.CUHK.EDU.HK
Bo Han[‡] [1] BHANML@COMP.HKBU.EDU.HK
Kun Zhang [3, 2] KUNZ1@CMU.EDU

[1] *TMLR Group, Hong Kong Baptist University, Kowloon, Hong Kong SAR*
[2] *Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE*
[3] *Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[4] *School of Computer Science, The University of Sydney, NSW 2008, Australia*
[5] *School of Mathematics and Statistics, The University of Melbourne, Victoria 3010, Australia*
[6] *The Chinese University of Hong Kong, Shatin, Hong Kong SAR*

Editor: Ilya Shpitser

Abstract

Revealing hidden causal variables alongside the underlying causal mechanisms is essential to the development of science. Despite the progress in the past decades, existing practice in causal discovery (CD) heavily relies on high-quality measured variables, which are usually given by human experts. In fact, the *lack of well-defined high-level variables behind unstructured data* has been a longstanding roadblock to a broader real-world application of CD. This procedure can naturally benefit from an automated process that can suggest potential hidden variables in the system. Interestingly, Large language models (LLMs) are trained on massive observations of the world and have demonstrated great capability in processing unstructured data. To leverage the power of LLMs, we develop a new framework termed **C**ausal representati**O**n **A**ssistan**T** (COAT) that incorporates the rich world knowledge of LLMs to propose useful measured variables for CD with respect to high-value target variables on their paired unstructured data. Instead of directly inferring causality with LLMs, COAT constructs feedback from intermediate CD results to LLMs to refine the proposed variables. Given the target variable and the paired unstructured data, we first develop COAT-MB that leverages the predictivity of the proposed variables to iteratively uncover the Markov Blanket of the target variable. Built upon COAT-MB, COAT-PAG further extends to uncover a more complete causal graph, i.e., Partial Ancestral Graph, by iterating over the target variables and actively seeking new high-level variables. Moreover, the reliable CD capabilities of COAT also extend the debiased causal inference to unstructured data by discovering an adjustment set. We establish theoretical guarantees for the CD results and verify that COAT can effectively and reliably make the most use of

[*] These authors contributed equally to this work. [‡] Corresponding author.

©2026 Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-2137.html.

LLM knowledge for CD across 7 realistic benchmarks and real-world case studies, including analysis of human reviews and diagnosis of neuropathic and brain tumors.

Keywords: causal discovery, large language models, causal representation learning, active causal representation elicitation, partial ancestral graph, Markov blanket

1. Introduction

Science originates along with identifying important variables and revealing their causal relations (Hanson, 1958; Kuhn, 1962). Despite the progress in the past decades, existing causal discovery (CD) approaches mainly rely on high-quality measured variables, which are usually given by human experts (Spirtes et al., 1993, 2018; Vowels et al., 2022). However, the desired variables and their measurements are usually unavailable in a wide range of real-world applications. For example, Amazon sellers who want to analyze the factors related to user ratings only have raw user reviews, which are written according to the underlying user preferences for certain product characteristics. The high-level variables, such as the preferred product characteristics, reflected from the low-level raw text reviews, are useful for causal analysis on how to improve purchase rates of the products. Therefore, *the lack of high-quality high-level variables* has been a longstanding impediment to broader real-world applications of CD or causality-inspired methods (Schölkopf et al., 2021).

The recently emerged Large Language Models (LLMs) (Brown et al., 2020; OpenAI, 2022; Touvron et al., 2023; OpenAI, 2023) offer a new opportunity to mitigate the gap (Kiciman et al., 2023). Trained from massive observations of the world, LLMs demonstrate impressive capabilities in *comprehending unstructured inputs*, and combining the rich knowledge learned from pretraining to resolve a variety of general tasks (Bubeck et al., 2023). Early studies suggest that incorporating LLMs as a *direct reasoner* can effectively answer commonsense causal questions, demonstrating the great potential in resolving the bottlenecks in causal learning and reasoning (Zhang et al., 2023a; Kiciman et al., 2023; Abdulaal et al., 2024). Nevertheless, the reliability of LLMs in directly reasoning the causal structure behind any specific data-generating process remains a debate (Zečević et al., 2023; Jin et al., 2023a,b; Zhang et al., 2023a) due to a series of drawbacks of LLMs (Zhang et al., 2023b; Cui et al., 2023; Berglund et al., 2023). The other works also try to leverage the prior knowledge of LLMs to assist causal discovery with respect to *a fixed set of given variables* (Ban et al., 2023; Abdulaal et al., 2024), while overlooking the valuable insight from unstructured data, e.g., image or text. In fact, the existing use of LLMs for CD still sticks to the traditional paradigm of CD, while surprisingly overlooking the identifiability of causal structure that is essential for the faithfulness of CD results (Spirtes et al., 1993, 2018; Vowels et al., 2022). Furthermore, it is also unclear how LLMs can be integrated into causal inference tasks (Peters et al., 2017; Bareinboim et al., 2022). Hence, it raises a challenging research question:

How can LLMs reliably assist with discovering and reasoning the causality behind the observed world?

To answer the question, we consider a general setting ubiquitous in real-world applications, where the data consists of a set of high-value numerical variables $\mathcal{U}$, such as customer ratings and medical diagnoses, and their paired unstructured data $\boldsymbol{X}$, such as customer reviews or

	Content	Clarification
Method	COAT-MB (Algorithm 1; Section 3.3)	inherited
	COAT-PAG (Algorithm 2; Section 3.4)	**new**
	COAT-ADJ (Algorithm 3; Section 3.5)	**new**
Theoretical Analysis	Identifiability on Markov Blanket (Section 4.1)	inherited
	Identifiability on Partial Ancestral Graphs (Section 4.2)	**new**
	Identifiability on Adjustment Sets (Section 4.3)	**new**
Experiments	Apple Gastronome (setting 1) (Section 5.1)	inherited
	Neuropathic data set (Section 5.3)	inherited
	Three case studies (Section 6)	inherited
	Apple Gastronome (setting 2) (Section 5.2)	**new**
	Hillstrom and Retail Hero data sets (Section 5.4)	**new**
	Apple Gastronome constructed with different LLMs	**new**
	Error Propagation Stress Test and Analysis	**new**
	Empirical comparison between COAT-MB, -PAG, and -ADJ	**new**

Table 1: Comparison of main results with conference version (2024).

medical images. Built upon the setting, we further decompose and approach the general question into actionable sub-questions:

Q1. As the first step, we begin by identifying factors that are useful for predicting the high-value target variable, which are immediately valuable for a variety of applications. Specifically, given one scalar target variable $Y \in \mathcal{U}$, how can we reliably acquire a set of high-level factors from unstructured data $\boldsymbol{X}$ that is a Markov Blanket of Y?

Q2. Furthermore, we extend our analysis to reveal more causal insights on the underlying high-level factors. Specifically, given a set of variables $\mathcal{U}$, how can we maximally reveal the causal information in the Partial Ancestral Graph (PAG) $\mathcal{P}_{\mathcal{U}}$ by extending the nodes with factors from unstructured data $\boldsymbol{X}$?

Q3. Built upon the high-level factors identified from low-level observations data, we then explore another important application of causality in real-world applications, i.e., causal effect estimation: How can the elicitation of causal variables facilitate treatment effect estimation?

To answer the aforementioned questions, we investigate a new interplay between CD and LLMs, and propose a different paradigm termed **C**ausal representati**O**n **A**ssistan**T** (COAT): rather than directly reasoning for causality on existing variables, we harness LLMs to actively augment data sets by proposing and parsing useful high-level variables from unstructured data, which we term as *Active Causal Representation Elicitation*. In general, the COAT framework employs two key components. (i) LLMs: these models can understand unstructured input $\boldsymbol{X}$ like text and images. Therefore, they are capable of proposing useful candidate factors $\mathcal{W}$ from the underlying possible factors $\mathcal{O}$ based on observation and instruction, and can parse out actual factor values for each unstructured

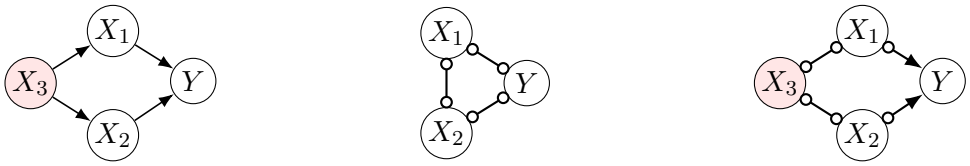


Figure 1: An example where PAG is less informative when limited in $\mathrm{MB}(Y)$. (a) is the true causal structure indicating causal directions; (b) and (c) are the partial ancestral graphs. The circle mark ($\circ$) means it is undetermined to be arrow head or tail. The node out of $\mathrm{MB}(Y)$ is colored.

observation. (*ii*) Guaranteed CD Methods for structured data: these numerical algorithms have rigorous identifiability about when and how the causal information can be revealed. We utilize them to parse and construct feedback to refine the candidate factors $\mathcal{W} \subseteq \mathcal{O}$ proposed by LLMs. The feedback is the essential drive for Active Causal Representation Elicitation, with which COAT is able to combine the best of LLMs and traditional CD methods to uncover the desired causality from the unstructured data.

For the sub-question **Q1**, we introduce COAT-MB as a reformulation of the previous work (Liu et al., 2024) on revealing Markov Blanket (MB) factors. As stated in Algorithm 1, in each iteration COAT-MB first uses LLMs to propose a set of candidate factors, and then uses CD methods to construct feedback based on the predictivity of the target variable. The feedback filters out samples where the target variable can not be well explained by existing factors discovered in previous rounds. We also establish conditions on when LLMs can effectively identify the subset $\mathcal{W} \subseteq \mathcal{O}$ as a Markov Blanket of the given target variable $Y \in \mathcal{U}$ (Section 4.1).

For the sub-question **Q2**, we extend COAT framework with a new approach beyond the previous Markov Blanket setting. Essentially, Markov Blanket factors only cover limited causal information of the high-level variables. Shown as in Figure 1, the skeleton between two Markov Blanket factors $X_1, X_2 \in \mathrm{MB}(Y)$ and their arrow heads towards Y, cannot be revealed without $X_3 \notin \mathrm{MB}(Y)$. Therefore, it motivates us to develop COAT-PAG to tackle **Q2**. COAT-PAG can elicit helpful factor set $\mathcal{W}$ from unstructured input $\boldsymbol{X}$ to further reveal causal structure (in the partial ancestral graph) among the input variable set $\mathcal{U}$. In Section 4.2, we show that the skeleton and arrow-head information within any subset $\mathcal{U}' \subseteq \mathcal{V} \triangleq \mathcal{W} \cup \mathcal{U}$ (where $\mathcal{W}$ is produced by COAT-PAG) can be maximally revealed if $\mathrm{MB}(\mathcal{U}') \subseteq \mathcal{V}$.

For the sub-question **Q3** , we further extend COAT framework to causal inference. Specifically, we propose COAT-ADJ that finds a PAG $\mathcal{P}_{\mathcal{V}}$ where the adjustment sets within any subset $\mathcal{U}' \subseteq \mathcal{V}$ are preserved if $\mathrm{MB}(\mathcal{U}') \subseteq \mathcal{V}$, as stated in Algorithm 3. In this paper, we employ the Generalized Adjustment Criterion (Perković et al., 2018), a sound and complete graphical characterization for adjustments sets, to process the PAG returned by COAT-ADJ. By this approach, we show that COAT-ADJ can achieve human-level unbiased treatment effect estimation from unstructured data (Section 4.3).

This paper builds upon the preliminary conference work (Liu et al., 2024), which focuses on identifying a set of Markov Blanket factors (Section 3.3 and 4.1). However, as summarized in Table 1, the present paper is significantly extended in several aspects.

- In Section 3.4, we further extend the original framework to refine causal information about the input structured variable set $\mathcal{U}$ with unstructured data $\boldsymbol{X}$. We propose a new approach COAT-PAG, as stated in Algorithm 2, to augment the node set $\mathcal{U}$ by proposing factors $\mathcal{W}$ from $\boldsymbol{X}$. In Section 4.2, we show that COAT-PAG can produce an augmented node set $\mathcal{V}$ where the skeleton and arrow heads among any subset $\mathcal{U}' \subseteq \mathcal{V}$ are maximally revealed if $\mathrm{MB}(\mathcal{U}') \subseteq \mathcal{V}$ (Proposition 8). In Corollary 9, we show that COAT-PAG can uncover information about discriminating paths. Furthermore, we also provide additional evaluation for this setting in Section 5.2.
- In Section 3.5, we extend the original framework to construct adjustment sets from unstructured data $\boldsymbol{X}$ and the input variable set $\mathcal{U}$ for causal effect estimation from a treatment T to an effect Y. We propose a novel approach COAT-ADJ, as stated in Algorithm 3, to augment the node set $\mathcal{U}$ by proposing factors $\mathcal{W}$ from $\boldsymbol{X}$. In Section 4.3, through Proposition 16 and other related propositions, we show that on the partial ancestral graph over the augmented node set $\mathcal{V}$, all valid adjustment sets within any subset $\mathcal{U}' \subseteq \mathcal{V}$ can be read off by graphical criterion if $\mathrm{MB}(\mathcal{U}') \subseteq \mathcal{V}$.

This paper is organized as follows. In Section 2, we briefly review related work, and elaborate the novelty and motivation of our research. In Section 3, we introduce our COAT framework. We first explain the key components, then describe the variant and extensions including COAT-PAG and COAT-ADJ. In Section 4, we establish the identifiability results for the proposed methods. In Section 5, we present comprehensive evaluation, including synthetic data for eliciting the Markov Blanket and extending Partial Ancestral Graphs, ablation studies as well as the results on several realistic benchmarks. In Section 6, we perform case studies of COAT on the real-world unstructured data including images, database that requires coding, and sequential news data.

2. Related Work

2.1 Causal Methods for Structured Data

Causal discovery This topic aims to uncover the underlying causal structure from observational data (Spirtes et al., 1993). It is critical to both real-world applications and scientific discoveries (Greenland et al., 1999; Pearl and Mackenzie, 2018). Such structures are usually represented by Directed Acyclic Graphs, where each arrow starts from a cause to an effect.

Classical methods include constrained-based methods (Spirtes et al., 2018, 1993, 1995) that incorporate conditional independence tests to find Markov Equivalence class. There are also different methods with different assumptions, like constrained functional (Shimizu et al., 2006; Zhang and Hyvärinen, 2012; Hoyer et al., 2008), multiple domain data (Huang et al., 2020; Yang et al., 2018; Brouillard et al., 2020; Mooij et al., 2020; Perry et al., 2022), and non-stationary data (Malinsky and Spirtes, 2019; Huang et al., 2019, 2020; Liu and Kuang, 2023). There are also some works aiming to address realistic challenges,

like structure models with arbitrary distribution (Chen et al., 2022b), sub-sampled time series (Wu et al., 2024), and latent hierarchical causal structure (Huang et al., 2022; Dong et al., 2023; Jiang and Aragam, 2023; Li et al., 2024).

Despite recent theoretical and empirical improvements (Glymour et al., 2019; Vowels et al., 2022), most existing causal discovery approaches rely on well-structured data with human-crafted factors as inputs. For example, the identification of causal direction can still be challenging when there are confounders not included in data (Montagna et al., 2023). In this work, we show how to improve causal identification by introducing new variables from the paired unstructured data.

Causal effect estimation This topic aims to estimate the causal influence of one variable on another. The causal effect can be formulated as the average treatment effect in the potential outcome framework (Rubin, 1974; Rosenbaum and Rubin, 1983; Holland, 1986), or represented by an interventional distribution in the structural causal model (Goldszmidt and Pearl, 1992; Pearl and Verma, 1995; Pearl, 2009). While the random control experiment is the gold standard for the inference of causal effect (Rubin, 1978), it is appealing to conduct estimation through nonexperimental data (Pearl, 1995). One approach is to utilize instrumental variables that influence the effect variable only through the treatment variable (Angrist et al., 1996; Angrist and Pischke, 2009). For instance, Mendelian randomization analyzes causal effects among clinical factors with genetic variants (Burgess and Thompson, 2015). Another approach is to estimate the interventional distribution by conditional distributions through a set of covariates satisfying certain graphical criterion (Pearl, 1993; Shpitser et al., 2012). In this paper, we employ the Generalized Adjustment Criterion (Perković et al., 2018) as a showcase to demonstrate the usefulness of the proposed COAT framework in constructing valid adjustment sets with unstructured data.

2.2 Causal Methods for Unstructured Data

Causal representation learning This topic aims to develop approaches to identifying latent high-level variables (like location and color of an object) from low-level observations such as images (Schölkopf et al., 2021; Zhang et al., 2024b). Such identification can be feasible with certain conditions, like auxiliary information (Hyvärinen et al., 2019; Khemakhem et al., 2020; Locatello et al., 2020; Kong et al., 2022; Zhang et al., 2024b), sparsity (Klindt et al., 2021; Lachapelle et al., 2022; Ahuja et al., 2022; Zheng et al., 2022), or interventional data (Ahuja et al., 2023; Zhang et al., 2024a; von Kügelgen et al., 2023; Squires et al., 2023; Buchholz et al., 2023). Recent works also generalize the causal representation learning to graph-structured data (Chen et al., 2022a, 2023a, 2024), and discuss the optimization challenges in realizing causal representation learning (Chen et al., 2023c,b). In this work, we open up a different setting with LLMs, which we termed as *Active Causal Representation Elicitation*, to iteratively augment tabular data with constructed interpretable factors to improve causal structure identification.

Causal reasoning with LLMs This topic aims to analyze the causal relations implied by the semantics given a text. One research line is *text mining of causal relations* (Drury et al., 2022; Cui et al., 2024; Hosseini et al., 2021). In particular, some works focus on the identification of causal relations among events specified in documents (Gao et al., 2019;

Weber et al., 2020; Liu et al., 2021). Another line is to *understand the causal reasoning ability of LLMs*. Kiciman et al. (2023) find that LLMs can recover the pairwise causal relations based on their textual description. Lyu et al. (2022) find that it is crucial to incorporate prompts aligned with the underlying causal story for LLMs to do pairwise causal relation inference. Willig et al. (2022); Zečević et al. (2023) find that LLMs can not understand causality but simply retell the causal knowledge contained in the training data.

Recent works explore causal reasoning with LLMs in trustworthy, explainable, and high-stake scenario (Han et al., 2025; Zhou et al., 2025; Li et al., 2025b). Chi et al. (2024) propose goal-oriented causal reasoning with retrieval-augmented generation system to involve external reliable knowledge. Khatibi et al. (2025) incorporate causal insights to refine the query iteratively for multi-hop reasoning. In this work, instead of reasoning about existing causal knowledge, COAT focuses on learning novel causal insights from data, which is complementary to these approaches.

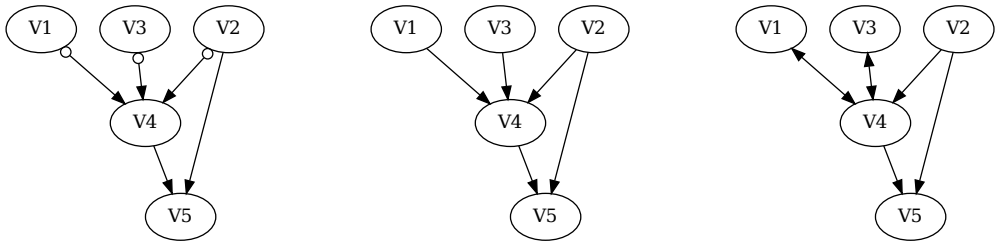
Causal learning with LLMs This topic aims to draw novel causal insights on data with LLMs. Preliminary studies (Zhang et al., 2023a; Jin et al., 2023b,a; Zečević et al., 2023) find that LLMs can hardly provide satisfactory answers for discovering new knowledge or decision-making tasks. Wu et al. (2025) suggest that LLMs should be a non-decisional support in causal discovery. One promising direction is to incorporate the causal discovery results by LLMs as a prior or constraint (Choi et al., 2022; Long et al., 2023; Ban et al., 2023; Abdulaal et al., 2024) to improve the performance of data-driven causal discovery algorithms. Vashishtha et al. (2023) find that causal order is a more stable interface for utilizing imperfect expert knowledge from LLMs. There are also some active learning approaches. Lampinen et al. (2023) shows that transformer-based agents can learn causal strategies passively if intervention is allowed during tests. Li et al. (2025a) incorporates LLMs to augment numerical causal discovery by intervention targeting. Wang et al. (2025) develops the Causal Copilot framework to enable rigorous causal analysis for non-experts.

Despite promising advances in integrating LLMs with causal discovery, current approaches still focus on artificially curated structured data with a fixed variable set, and the potential of paired unstructured data in each sample is less explored. In this work, COAT focuses on actively extending the variables from unstructured data. As we shall show later, this approach can reliably improve the structure identification with guarantees.

3. The Causal representatiOn AssistanT Framework

In this section, we present the COAT framework, which leverages an LLM to serve as a representation assistant and actively elicits the desired causal variables on unstructured data. The key drive for the Active Causal Representation Elicitation is based on the feedback constructed from the intermediate causal discovery results with CD methods. We will begin with revealing the underlying Markov Blanket for a given target variable, and then extend to Partial Ancestral Graph, and the desirable Adjustment Set for treatment effect estimation.

Notation In this paper, random variables are denoted by capital Roman letters (e.g., $U_1, U_2, \ldots, U_m$); random vectors are expressed by bold capital Roman letters (e.g., $\mathbf{U}$); and sets of random variables are written by calligraphic letters (e.g., $\mathcal{U}$). Consistency



(a) partial ancestral graph (b) mixed graph (c) mixed graph

Figure 2: Examples of Partial Ancestral Graph and two Markov-equivalent mixed graphs.

in the base letter reflects their relationship: for instance, $\mathbf{U} = [U_1, U_2, \ldots, U_m]^\top$ and $\mathcal{U} = \{U_1, U_2, \ldots, U_m\}$. We summarize the symbols with special meanings in this paper in Appendix A.

For presenting edges in PAGs, this paper follows the practice in Zhang (2008a) and Perković et al. (2018). Each end of one edge is associated with one of three possible marks: tail ($-$), arrow ($>$), and circle ($\circ$). For example, edges can be bi-directed $\leftrightarrow$, directed $\rightarrow$, partially directed $\circ\!\!\rightarrow$, and undirected $\circ\!\!-\!\!\circ$. An asterisk ($*$) would be used in a rule if it can be applied to any of the three marks.

3.1 Preliminary and Overview

In this paper, we aim to reveal three types of causal representations that are of practical value for discovery from unstructured data: Markov Blanket, Partial Ancestral Graph, and Adjustment Sets, respectively. In these settings, the observed unstructured data is denoted as $\boldsymbol{X} \in \mathcal{X}$ (e.g., natural language, images) and an LLM serves as a transformation function $\Psi : \mathcal{X} \to \mathcal{Z}$, parameterized by a set of prompts. Let $\boldsymbol{Z} := \Psi(\boldsymbol{X}; \mathbf{p})$ be the structured vector converted from an unstructured data $\boldsymbol{X}$ according to prompt $\mathbf{p}$.

Markov Blanket The *Markov Blanket* (MB) of a target variable Y (here Y is an arbitrary variable of interest), denoted $\mathbf{MB}(Y)$, is the minimal subset of factors such that Y is conditionally independent of others:

$$Y \perp (\mathcal{Z} \setminus \mathbf{MB}(Y)) \mid \mathbf{MB}(Y). \tag{1}$$

Equivalently, $\mathbf{MB}(Y)$ comprises the parents, children, and spouses (parents of common children) of Y in the underlying causal DAG.

COAT-MB leverages the LLM to iteratively analyze data $\mathcal{D} \subseteq \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, and suggests new factors by itself. The goal is to ensure $\mathbf{MB}(Y) \subseteq \boldsymbol{Z}$ such that no factor outside the set of the discovered factors provides additional information on Y.

Partial Ancestral Graphs A *Maximal Ancestral Graph* (**MAG**) is typically used to represent the causal structure among a set of observed variables $\mathbf{O}$ when there are possibly unobserved latent variables $\mathbf{L}$. As shown in Figure 2(b) and 2(c), it has following edge types: *directed* edge ($\rightarrow$) indicates clear causal relations; *bi-directed* edge ($\leftrightarrow$) indicates latent confounder between two variables. A *Partial Ancestral Graph* (**PAG**) $\mathcal{P}_{\boldsymbol{U}}$ over variables $\mathbf{O}$ is used to represent a Markov equivalent class of MAGs that share the same set of

conditional independencies. It has two additional edge types: *partially directed* edge ($\circ\!\!\rightarrow$) for possible causal relations, i.e., could be directed or bi-directed; *non-directed* edge ($\circ\!\!-\!\!\circ$) for the uncertain relations. The circle ($\circ$) mark means that there exist some Markov equivalent MAGs with different edge marks. For example, one can find the circle mark at $V1$ at Figure 2(a) is due to the inconsistent marks in the two MAGs. When given the numerical samples of $\mathbf{O}$, the Fast Causal Inference (Spirtes et al., 1995) algorithm constructs $\mathcal{P}$ by performing conditional independence tests.

In COAT-PAG, given an initial structured variables $\boldsymbol{U}$, the algorithm extends the node set to $\boldsymbol{U}' := \boldsymbol{U} \cup \boldsymbol{Z}$ by proposing new factors, where each factor can be treated as a function of unstructured data $\boldsymbol{X}$. The goal is to maximally reveal the causal information among $\boldsymbol{U}$. For example, some non-directed edge in $\mathcal{P}_{\boldsymbol{U}}$ could be directed in $\mathcal{P}_{\boldsymbol{U}'}$.

Adjustment Sets One important question in causal inference is to estimate the causal effect from a treatment variable T to an outcome variable Y given a causal graph $\mathcal{G}$ over a node set $\boldsymbol{V}$ with $\{T, Y\} \subseteq \boldsymbol{V}$. The association between T and Y can be confounded by other variables, e.g., a confounder V_1 that is a shared causal parent of both T and Y (i.e., $T \leftarrow V_1 \rightarrow Y$). In such cases, instead of estimating the raw association, we condition on an *adjustment set* to block non-causal paths while preserving causal pathways. For instance, adjusting for V_1 can be conducted by *inverse probability weighting* (Robins et al., 2000): one models the propensity score $\pi(V_1) = P(T = 1 \mid V_1)$ and reweights each observation by $w = \frac{T}{\pi(V_1)} + \frac{1-T}{1-\pi(V_1)}$ to create a pseudo-population in which T is independent of V_1. The average treatment effect (ATE) can then be estimated as

$$\hat{\tau}_{\text{ATE}} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i Y_i}{\pi(V_{1i})} - \frac{(1 - T_i) Y_i}{1 - \pi(V_{1i})} \right),$$

which provides a consistent estimator of $\mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)]$ under correct propensity score specification and positivity.

In COAT-ADJ, given $(T, Y) \subseteq \boldsymbol{U}$ and unstructured data $\boldsymbol{X}$, we discover factors $\boldsymbol{Z}$ from $\boldsymbol{X}$ to construct valid adjustment sets within $\boldsymbol{V} := \boldsymbol{U} \cup \boldsymbol{Z}$. Based on the estimated causal structure $\mathcal{P}_{\boldsymbol{V}}$, one could check identifiability conditions (e.g., the absence of unobserved confounding along back-door paths) and select adjustment sets using generalized graphical criteria. In this paper, we focus on the *Generalized Graphical Criteria* (Perković et al., 2018), please see Definition 12 for details.

3.2 Active Causal Representation Elicitation

Data We are given *unstructured data* (e.g., text or images) as a high-dimensional random vector $\boldsymbol{X}$, and *structured data* (e.g., stars rated by a customer or the tumor type of a patient) as a set of random variables of interest $\mathcal{U} = \{U_1, \ldots, U_m\}$. The input data set $\mathcal{D} = \{\boldsymbol{x}^{(j)}, \boldsymbol{u}^{(j)}\}_{j=1}^{n}$ contains n samples independently drawn from the joint distribution over $(\boldsymbol{X}, \boldsymbol{U})$, where $\boldsymbol{U}$ is an m-dim random vector that consists of variables in $\mathcal{U}$. There is no prior assumption on the relations between $\boldsymbol{X}$ and $\boldsymbol{U}$.

LLM as a representation assistant Given the unstructured data $\boldsymbol{X}$, we aim to make use of the rich knowledge of LLMs to assist in converting meaningful unstructured information

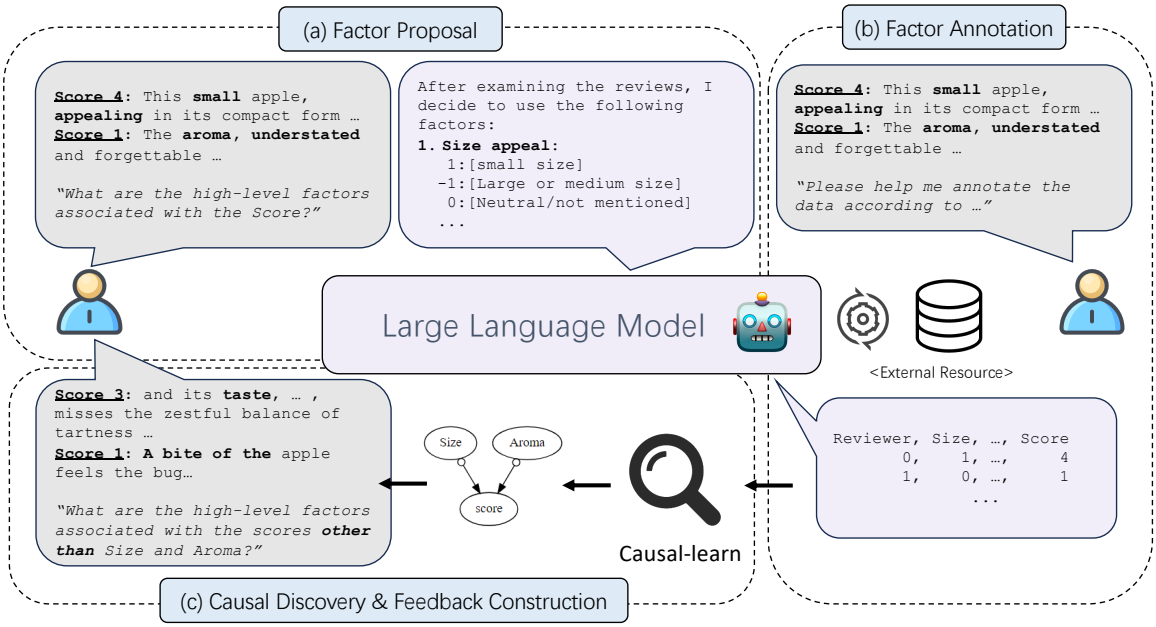


Figure 3: Illustration of COAT framework in eliciting the Markov Blanket of the rating score from the paired text reviews. COAT first (a) adopts an LLM to read, comprehend, and relate the rich knowledge about reviews on apples. The LLM needs to propose a series of candidate factors such as apple sizes and smells, along with some meta-information such as annotation guidelines. Based on the candidate factors, COAT then (b) prompts another LLM to annotate the unstructured review into structured data. (c) The CD algorithm then finds causal relations among the factors, and constructs feedback based on samples where the ratings can not be well explained by the existing factors. By looking into the new samples, the LLM is expected to use more related knowledge to uncover the desired causal factors.

in a structured form. Formally speaking, we seek a mapping $h : \boldsymbol{X} \mapsto \boldsymbol{W}$ which serves as a *structured* representation $\boldsymbol{W} = h(\boldsymbol{X})$. We present the following definition:

Definition 1 (Representation Assistant) *A* ***Representation Assistant*** *for the unstructured data $\boldsymbol{X}$ with domain $\mathcal{X}$ is defined by a set of high-level factors $\mathcal{W} \triangleq \{W_1, \ldots, W_k\}$. Each factor W_i is induced by a prompt instruction (e.g., a natural language query), which specifies a deterministic mapping $\boldsymbol{w}_i : \mathcal{X} \to \mathcal{C}_i$ via an LLM Ψ. Here, $\mathcal{C}_i \subseteq \mathbb{R}$ is a predefined numerical set (discrete or continuous), and $W_i = \boldsymbol{w}_i(\boldsymbol{X})$ is a random variable with range $\mathcal{C}_i$. The structured representation is the concatenation:*

$$h(\boldsymbol{X}) = \boldsymbol{W} = \Big(W_1, W_2, \cdots, W_k \Big).$$

Each factor is defined by a specific prompt instruction so that an LLM can decide a value based on a sample $\boldsymbol{x}$ (drawn from $\boldsymbol{X}$). For example, given a customer review on

an apple $\boldsymbol{x}^{(j)}$, an LLM-proposed factor $\boldsymbol{w}_i$ with $\mathcal{C} = \{-1, 0, 1\}$ could be described as: $\boldsymbol{w}_i(\boldsymbol{x}^{(j)}) = 1$ if the customer appreciates the sweetness of the apple; $\boldsymbol{w}_i(\boldsymbol{x}^{(j)}) = -1$ if the customer is disappointed about the sweetness; and $\boldsymbol{w}_i(\boldsymbol{x}^{(j)}) = 0$ if the sweetness has not been mentioned. In addition, a factor instruction can require LLM to utilize external tools for further processing.

Active causal representation elicitation In this work, we propose **C**ausal representati**O**n **A**ssistan**T** (COAT) framework to actively elicit representations consisting of factors $\mathcal{W}$ to reveal causality in different settings.

- Given a target variable $Y \in \mathcal{U}$, COAT-MB (Section 3.3) finds factors $\mathcal{W}$ that serves as a Markov Blanket of Y with respect to $\boldsymbol{X}$, i.e., $Y \perp\!\!\!\perp \boldsymbol{X} \mid \mathcal{W}$. That is, $\mathcal{W}$ captures all relevant information about Y from $\boldsymbol{X}$;
- Furthermore, based on COAT-MB, COAT-PAG (Section 3.4) extends the structured input set to $\mathcal{V} \triangleq \mathcal{W} \cup \mathcal{U}$ so that the causal information, i.e., skeleton or arrow heads, among $\mathcal{U}$ can be maximally revealed in the partial ancestral graph $\mathcal{P}_{\mathcal{V}}$ with the node set $\mathcal{V}$;
- In causal inference setting, for a treatment-effect pair $T, Y \in \mathcal{U}$, COAT-ADJ (Section 3.5) extends structured input set to $\mathcal{V} \triangleq \mathcal{W} \cup \mathcal{U}$ so that valid adjustment sets within $\mathcal{U}$ for the causal effect on $T \to Y$ can be found by graphical criterion on the partial ancestral graph $\mathcal{P}_{\mathcal{V}}$ with the node set $\mathcal{V}$.

To introduce the key components of COAT framework, we take the illustration in Figure 3 as a running example, where COAT aims to elicit useful factors to serve as a Markov Blanket of a target variable Y (rating score) with respect to text reviews through multiple rounds of iteration. We use the superscript t to refer to the input and output of LLM at the t-th round. We denote the union of results from the first t rounds by the superscript "$\leq t$".

Factor proposal To obtain useful high-level factors from the rich world knowledge of LLMs, COAT employs a prompt $\boldsymbol{p}$ and a few samples $\widehat{\mathcal{D}} \subseteq \mathcal{D}$. Specifically, in Figure 4, $\boldsymbol{p}$ contains three components: *samples*, *instructions*, and *format control*. To encourage LLM to focus on the information related to the target variable Y, *samples* are grouped by the value of Y. The *instruction* requires Ψ to give each proposed factor $\boldsymbol{w}_i$ a concrete description of the mapping $\boldsymbol{w}_i(\cdot)$, like how to decide the factor values. In addition, the metadata about the task, such as the task description and context, can also be incorporated if available. The prompt $\boldsymbol{p}$ essentially imitates the typical process where human experts abstract and define high-level variables. The set of resulting factors in the t-th round, defined with natural language by the LLM Ψ, is denoted as $\mathcal{W}^t = \Psi(\boldsymbol{p}^t, \widehat{\mathcal{D}}^t)$. We merge them with factors proposed in the previous rounds to update the set of all factors $\mathcal{W}^{\leq t} = \mathcal{W}^1 \cup \cdots \cup \mathcal{W}^t$.

Factor parsing Once we obtain the candidate factors, we then collect the values of the factors from the unstructured observations. In prior works, they are usually collected from human experts according to the given factors (Spirtes et al., 1993). To do so, another LLM Ψ_s is instructed to read the annotated guidelines of each variable $\boldsymbol{w}_i$ and parse the unstructured observations into structured or tabular data, with its entity at the i-th column

```
You are an excellently helpful AI assistant
for analyzing, abstracting, and processing
data. Now try your best.

Data
group with 'score' = -1.0
- This apple variety impresses ...
(more samples)
group with 'score' = 0.0
(more samples)
group with 'score' = 1.0
(more samples)
....

Your output should contain parts described
as follows.

Part 1: Consideration.
In this part, feel free to write down the
process of your considerations.
Hint: You need to abstract, identify, and
design suitable factors with corresponding
criteria, and each factor is only allowed
to take value from [-1, 0, 1].
```

```
Part 2: Factor Filtration.
You should decide whether to use each of
the proposed factors by following criteria:
- Each new factor should be helpful to
distinguish the groups.
- Each Factor should focus on one concrete
aspect and try to avoid overlap.

Part 3: Final Output.
In this part, you are required to report
the factors that you decided to use.
- Each factors' value should from [-1, 0, 1]
- You should give specific criterion for
each value of each factor.
- Keep the criterion for value 0 fixed as
"Otherwise; or not mentioned".

Report the factors in following template:

```
Factor Name
- 1: [Positive Criterion]
- 0: [Otherwise; or not mentioned].
- -1: [Negative Criterion]
```
```

Figure 4: Illustration of the prompt template for factor proposal in COAT.

and the j-th row:

$$\boldsymbol{w}_i(\,\boldsymbol{x}^{(j)}\,) := \Psi_s(\,\boldsymbol{x}^{(j)},\, \boldsymbol{w}_i,\, \boldsymbol{p}_p\,), \tag{2}$$

where $\boldsymbol{p}_p$ refers to the additional instruction to prompt Ψ_s to parse the observed data.

When the data curation of the proposed factors requires additional domain-specific knowledge/skills (e.g., intervening on the external environments) that the LLMs do not acquire, we could fetch $\boldsymbol{w}_i(\boldsymbol{x}^{(j)})$, the value of each factor $\boldsymbol{w}_i$ for each sample $\boldsymbol{x}^{(j)}$, through some external process (Schick et al., 2023; Xi et al., 2023). For example, studying the causes of a disease requires annotating relevant symptoms from diagnosis records and conducting additional medical checks (Tu et al., 2019). Our experiments show that COAT can effectively extract the hidden factors under both schemes.

Causal discovery Given the candidate factors $\mathcal{W}^{\leq t}$ with their associated factor values on data set $\mathcal{D}$, a CD algorithm $\mathcal{A}$ (e.g., FCI (Spirtes et al., 1995)) is used to reason about the causal structure based on the parsed data:

$$\mathcal{G}^t = \mathcal{A}\Big(\,\mathcal{W}^{\leq t} \cup \{Y\}\,\Big), \tag{3}$$

where $\mathcal{G}^t$ is the discovered causal structure. In general, any CD approach with suitable theoretical assumptions could be used for $\mathcal{A}$. More detailed theoretical discussions are presented in Section 4.1. In this work, we demonstrate the idea of COAT via the FCI algorithm (Spirtes et al., 2018) as it is flexible with respect to different functional classes of the underlying generation process, allowing for the existence of latent confounders (Spirtes et al., 1995), which aligns well with our objective.



(a) Parent (b) Child (c) Confounded (d) Spouse

Figure 5: Illustration of variables that could be discovered with COAT. Let Y be the target variable, and W be a factor that has been discovered, and also assume a latent variable $U = \widehat{\boldsymbol{w}}(\boldsymbol{X}) \in \mathrm{MB}(Y)$. Conditioning on W facilitates the discovery of U.

Improving factor proposal with causal feedback LLMs require proper prompts to fully unlock their capabilities (Wei et al., 2022; Huang et al., 2023; Qiao et al., 2023). When it comes to factor proposing, it is also hard for LLMs to propose all factors at once. Nevertheless, from the causal discovery results, we could find useful information and thus provide feedback to further improve the factor proposal:

$$\Big(\widehat{\mathcal{D}}^{t+1},\, \boldsymbol{p}^{t+1}\Big) = \mathcal{F}\Big(\mathcal{G}^t,\, \mathcal{D},\, \boldsymbol{p}^t\Big), \tag{4}$$

where $\mathcal{F}$ samples specific examples from $\mathcal{D}$ and constructs new prompts according to the results of $\mathcal{A}$ for the next round of factor proposal. For example, FCI is able to imply the existence of latent confounders, from which we could refine $\boldsymbol{p}$ to prompt Ψ to focus on the corresponding factors.

3.3 COAT for Markov Blanket

In this section, we restate our previous result (Liu et al., 2024) to identify a set of *structured* high-level factors $\mathcal{M}$ from *unstructured* observations $\boldsymbol{X}$ that serve as a Markov Blanket of the given target variable Y. The basic idea is intuitive: we iteratively verify the proposed factors from LLMs, and, in the meantime, encourage LLMs to propose diverse factors based on their observations. All proposed factors will be put into the factor pool $\mathcal{W}$ to serve as proxies of the unstructured variable $\boldsymbol{X}$ in statistical testing. After each iteration, we update the choice of Markov Blanket $\mathcal{M} \subseteq \mathcal{W}$ with respect to the current factor pool, i.e., $Y \perp\!\!\!\perp \mathcal{W} \setminus \mathcal{M} \mid \mathcal{M}$.

The concrete procedure is stated in Algorithm 1. In the t-th loop, we first conduct clustering on the current Markov Blanket set. The intuition is to strengthen the correlation between Y and the potential factors. If there exists such a potential factor, say, $\widehat{\boldsymbol{w}}$, that means:

$$H(\,Y \mid \mathcal{M}\,) > H\Big(Y \mid \mathcal{M},\, \widehat{\boldsymbol{w}}(\boldsymbol{X})\Big), \tag{5}$$

where $H(\cdot)$ refers to the entropy. This is because Y should be independent of other factors from $\boldsymbol{X}$ conditioning on the Markov Blanket. As shown in Figure 5, finding factors satisfying Inequality 5 progressively expands the discovered factors and pushes $\mathcal{M}$ to a valid Markov Blanket. Therefore, to find the desired factor, we are motivated to select a suitable $\widehat{\mathcal{D}}^t$ for

Algorithm 1 COAT for Markov Blanket (COAT-MB)

Input: Data set $\mathcal{D} = \{ (\boldsymbol{x}^{(j)}, y_j) \}_{j=1}^{n}$; LLM for factor proposal Ψ; Model for factor parsing Ψ_s; causal discovery algorithm $\mathcal{A}$; Maximal rounds t_{Max}.

1. Let $\mathcal{M} = \emptyset$ be the initial Markov Blanket. Let $\mathcal{W} = \left\{ W_1, \cdots, W_m \right\}$ be the set of all the proposed factors from $\boldsymbol{X}$. Initially, $m = 0$.
2. Repeat with loop index t starting from 1, stop when $t = t_{\text{Max}}$ or $\mathcal{M}$ is unchanged:
   (a) Do clustering on samples with $\mathcal{M}$; use random partition if $\mathcal{M}$ is empty.
   (b) Take $\widehat{\mathcal{D}}^t$ as the cluster with highest conditional entropy w.r.t. Y.
   (c) LLM Ψ propose new factors $W_{m+1}, \cdots, W_{m'}$, then update $\mathcal{W}$ and set $m = m'$.
   (d) For each new factor, LLM Ψ_s parse out its values on $\{\boldsymbol{x}^{(j)}\}_{j \in [n]}$.
   (e) if $\mathcal{M}$ is not Y's Markov Blanket w.r.t. $\mathcal{W}$, update $\mathcal{M}$ with $\mathcal{A}$.
3. Get causal graph $\mathcal{G}$ over $\{Y\} \cup \mathcal{W}$ with $\mathcal{A}$.

Output: Factors $\mathcal{W}$, Markov Blanket set $\mathcal{M}$, causal graph $\mathcal{G}$.

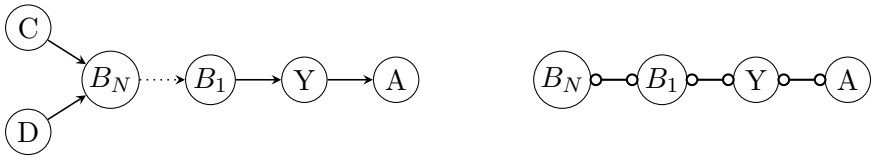
the current iteration such that $\widehat{\mathcal{D}}^t = \arg\max_{\widehat{\mathcal{D}} \subseteq \mathcal{D}} H_{\widehat{\mathcal{D}}}(Y|\mathcal{M})$, where $\mathcal{M}$ can not well explain Y. This problem can be interpreted as selecting the hardest data set $\widehat{\mathcal{D}}^*$ for predicting Y with existing factors. In our experiments, we implement the classification via clustering with respect to $\mathcal{M}$. The clustering elicits C groups $\widehat{\mathcal{D}}_c := \{\boldsymbol{x}_i \text{ for } i \in \mathcal{I}_c\}$: $\mathcal{I}_1, \cdots, \mathcal{I}_C = \text{K-Means}(\mathcal{M})$. We then take the group of samples with the largest conditional entropy to construct the feedback. In the rest of the t-th loop, we utilize the COAT components stated in Section 3.2 to instruct LLMs to propose candidate factors and then parse out their values on each observation. The Markov Blanket can be obtained by statistical testing or the causal graph from $\mathcal{A}$.

In practice, many factors, such as the LLM capabilities, data faithfulness, and prompt templates, could affect the success of Algorithm 1. Therefore, in Section 4.1, we formalize the required assumptions on the LLMs' ability, and establish a theoretical guarantee on the factor identification for Markov Blanket.

3.4 COAT for Partial Ancestral Graph

The causal information may not be maximally revealed due to insufficient node set. For example, as we have seen in Figure 1, we could further reveal the skeleton and arrow heads among $\text{MB}(Y)$ if we can include one additional factor $X_3 \notin \text{MB}(Y)$.

One straightforward approach is to apply COAT-MB recursively to the discovered variables to obtain $\text{MB}^{(2)}(Y)$, i.e., the Markov Blanket of the Markov Blanket of Y. Although it helps revealing the skeleton information, the arrow heads information can still be missing. Essentially, for any positive integer N, there exists a causal graph whose arrow heads cannot be revealed by obtaining $\text{MB}^{(N)}(Y)$: as shown in Figure 6(a), simply considering a chain structure $B_N \to \cdots B_1 \to Y \to A$ with a V-structure at the end away from Y:



(a) Ground-truth causal graph

(b) PAG over $\mathrm{MB}^{(N)}(Y)$

Figure 6: An example where arrow heads cannot be revealed within $\mathrm{MB}^{(N)}(Y)$.

Algorithm 2 COAT for Partial Ancestral Graph (COAT-PAG)

Input: Data set $\mathcal{D} = \{ (\boldsymbol{x}^{(j)}, \boldsymbol{u}^{(j)}) \}_{j=1}^{n}$; LLM for factor proposal Ψ; Model for factor parsing Ψ_s; causal discovery algorithm $\mathcal{A}$; Maximal rounds t_{Max}.

1. Initialize $\mathcal{W}$ with $\mathrm{MB}(\mathcal{U}) \setminus \mathcal{U}$ by performing Algorithm 1 with Ψ, Ψ_s, $\mathcal{A}$, and t_{Max}.
2. Let $\mathcal{P}_{\mathcal{V}}$ be the partial ancestral graph over $\mathcal{V} \triangleq \mathcal{U} \cup \mathcal{W}$, and initialize $\mathcal{S}_{\mathrm{mark}} = \mathcal{U}$.
3. Let $\mathcal{S}_{\mathrm{target}} = \left\{ \alpha \in \mathcal{V} \setminus \mathcal{S}_{\mathrm{mark}} \mid \mathcal{P}_{\mathcal{V}} \text{ contains edges like } \alpha \circ\!\!-\!\!* \beta \text{ for some } \beta \in \mathcal{V} \right\}$
4. Repeat with loop when $\mathcal{S}_{\mathrm{target}}$ is not empty:
   (a) Take one arbitrary element α from $\mathcal{S}_{\mathrm{target}}$.
   (b) Extend $\mathcal{W}$ with $\mathrm{MB}(\alpha) \setminus \mathcal{V}$ by performing Algorithm 1 on α.
   (c) Put α into $\mathcal{S}_{\mathrm{mark}}$, then update $\mathcal{P}_{\mathcal{V}}$ and $\mathcal{S}_{\mathrm{target}}$.

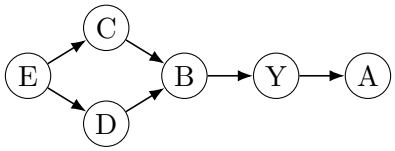
Output: Factor set $\mathcal{W}$, causal graph $\mathcal{P}_{\mathcal{V}}$ on $\mathcal{V} \triangleq \mathcal{U} \cup \mathcal{W}$.

$C \to B_N \leftarrow D$. When only $\mathrm{MB}^{(N)}(Y)$ are observed, as shown in Figure 6(b), C and D are latent variables; and thus there is no V-structure in the PAG restricted on $\mathrm{MB}^{(N)}(Y)$. Consequently, only the skeleton can be identified.

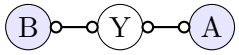
Motivated by the observation above, we propose a method called COAT-PAG to adaptively decide on which node we should apply COAT-MB to expand the current node set by its Markov Blanket. As we stated in Algorithm 2, the purpose of COAT-PAG is to ensure both the skeleton and the arrow heads within the initial factor set can be maximally revealed in the PAG on the output factor set. In the first step, it simply expands the factor set $\mathcal{U}$ to $\mathrm{MB}(\mathcal{U})$ to ensure the correctness of the skeleton. In the following steps, it conducts Breadth-First Search to apply COAT-MB on the nodes that are adjacent to an undetermined edge head ($\circ$).

The procedure of COAT-PAG is illustrated via the example in Figure 7. The skeleton is identified at the beginning; however, the arrow heads are present only after factor E is included. The presence of E will introduce a V-structure $C \to B \leftarrow D$ and thus several edges can be oriented.

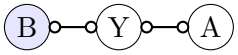
In Section 4.2, we formally show that for any subset $\mathcal{U}'$ of the extended factor set $\mathcal{V}$ from Algorithm 2, the skeleton and arrow heads will be maximally revealed if $\mathrm{MB}(\mathcal{U}') \subseteq \mathcal{V}$.



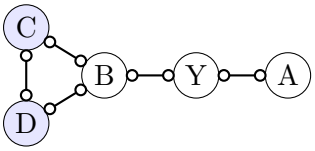
(a) Ground-truth causal graph



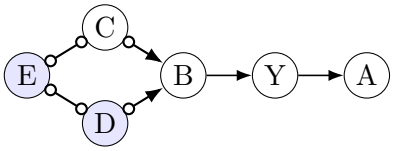
(b) Loop 0: $\mathrm{S}_{\text{target}} = \{A, B\}$



(c) Loop 1: $\mathrm{S}_{\text{target}} = \{B\}$



(d) Loop 2: $\mathrm{S}_{\text{target}} = \{C, D\}$



(e) Loop 3: $\mathrm{S}_{\text{target}} = \{D, E\}$

Figure 7: An example where COAT-PAG is conducted in several loops. $\mathrm{S}_{\text{target}}$ are colored.

For arrow tails, we show that they can be maximally revealed if it is in a discriminating path within $\mathcal{U}'$.

3.5 COAT Extension on Adjustment Set

With the reliable causal representation elicitation frameworks established in the previous sections, we are able to further apply COAT approach to causal inference. Specifically, we consider two variables $T, Y \in \mathcal{U}$, where T is the treatment variable, Y is the effect variable, and $\boldsymbol{X}$ is the associated unstructured data. The goal is to propose a set of factors from $\boldsymbol{X}$ to serve as an adjustment set for the total causal effect estimation from T to Y, when it is possible. We seek an algorithm to output a factor set $\mathcal{W}$ so that an adjustment set for (T, Y) can be constructed by applying existing graphical criterion over the causal graph $\mathcal{P}_{\mathcal{V}}$ on $\mathcal{V} \triangleq \mathcal{W} \cup \mathcal{U}$.

The construction of an adjustment set is not always feasible even when all factors are given. For example, according to the sound and complete graphical characterization by Perković et al. (2018), if a PAG does not satisfy the *amenability* condition, then no adjustment set can be found in this graph. Therefore, one inevitable assumption is the existence of such adjustment sets that meet the graphical criterion.

COAT-ADJ is given in Algorithm 3: COAT-ADJ first includes all the important factors in the graph (step 1 to 3), and then further extends the node set to reveal the causal structure (step 4). The most important part is to maintain the $\mathcal{S}_{\text{path}}$ defined in step 2. The purpose of this set is to track and include the intermediate nodes on the path between T and Y, as well as their possible descendants. As we shall show in Section 4.3, including the

Algorithm 3 COAT for Adjustment Sets (COAT-ADJ)

Input: Data set $\mathcal{D} = \{ (\boldsymbol{x}^{(j)}, \boldsymbol{u}^{(j)}) \}_{j=1}^{n}$; specifications of $T, Y \in \mathcal{U}$; LLM Ψ for factor proposal; a factor parsing model Ψ_s; a causal discovery algorithm $\mathcal{A}$; and maximal rounds t_{Max}.

1. Initialize $\mathcal{W}$ with $\text{MB}(\mathcal{U}) \setminus \mathcal{U}$ by performing Algorithm 1 with Ψ, Ψ_s, $\mathcal{A}$, and t_{Max}; and initialize $\mathcal{S}_{\text{mark}}$ by $\mathcal{U}$. $\mathcal{P}_{\mathcal{V}}$ is the causal graph over $\mathcal{V} \triangleq \mathcal{U} \cup \mathcal{W}$ estimated by $\mathcal{A}$.
2. Let $\mathcal{S}_{\text{path}} = \left\{ \alpha \in \text{PossDe}(\beta, \mathcal{P}_{\mathcal{V}}) \setminus \mathcal{S}_{\text{mark}} \mid \beta \text{ is on a path between } T \text{ and } Y \text{ in } \mathcal{P}_{\mathcal{V}} \right\}$
3. Repeat with loop when $\mathcal{S}_{\text{path}}$ is not empty:
   (a) Take one arbitrary element α from $\mathcal{S}_{\text{path}}$.
   (b) Extend $\mathcal{W}$ with $\text{MB}(\alpha) \setminus \mathcal{V}$ by performing Algorithm 1 on α.
   (c) Put α into $\mathcal{S}_{\text{mark}}$, then update $\mathcal{P}_{\mathcal{V}}$ and $\mathcal{S}_{\text{path}}$.
4. Update $\mathcal{W}$ and $\mathcal{P}_{\mathcal{V}}$ by applying Algorithm 2 treating $\mathcal{V}$ as structured data, i.e., using the augmented data set $\mathcal{D}' \triangleq \{ (\boldsymbol{x}^{(j)}, \boldsymbol{v}^{(j)}) \}_{j=1}^{n}$.

Output: Factor set $\mathcal{W}$, causal graph $\mathcal{P}_{\mathcal{V}}$.

intermediate nodes can preserve the amenability, and including their possible descendants can help to exclude invalid nodes. In addition, Step 1 provides a necessary initialization for constructing $\mathcal{S}_{\text{path}}$; Step 3 is to exhaust $\mathcal{S}_{\text{path}}$ iteratively; and Step 4 reveals the causal structure among those factors.

After applying COAT-ADJ, the adjustment sets can be constructed by searching subsets that meet the generalized adjustment criterion (Perković et al., 2018). In Section 4.3, for the output factor set $\mathcal{W}$ (and also $\mathcal{V} \triangleq \mathcal{W} \cup \mathcal{U}$), we show that all the valid adjustment sets relative to (T, Y) in $\mathcal{U}' \subseteq \mathcal{V}$ will be preserved if $\text{MB}(\mathcal{U}') \subseteq \mathcal{V}$.

3.6 Practical Discussions

Supplementary to the aforementioned methodology discussion, in this section, we elaborate more on the practical considerations of COAT.

Instruction following In practice, strict compliance with output format specifications is critical for pipeline reliability. The violation of output format can be (1) proposing a factor in HTML format instead of the markdown format as required in Figure 4; or (2) passing unsupported measurements when using external tools as shown in Listing 1 in Section 6. These violations make parts of the output invalid, and consequently influence the overall quality. To alleviate this issue, one may need to provide more examples on valid format and documents about external tools, or employ stronger LLMs to handle complex operations.

Modern causal discovery We use FCI in this paper in order to illustrate the idea. To attain identifiability better than the Markov equivalent class, one can choose more advanced CD methods under different assumptions, see Section 2 if interested. We also discuss some cases where we need to handle them properly in practice with LLMs.

Factor filtering LLMs may output several factors with similar semantics or exhibit multicollinearity in the annotated data, which will hinder the causal discovery process. To mitigate the issue, one could do factor filtering that adopts PCA or early conditional independence tests given the currently discovered variables in the Markov Blanket to detect and eliminate these variables.

Factor pool LLMs may discover useful factors in early rounds while being discarded. For example, the underlying spouse variables of the target label Y may be independent of Y without conditioning on their common children variables. To resolve the issue, we could introduce a factor pool that stores the candidate variables proposed in the past, and replay the variables that have not been passed by conditional independence tests with existing variables in the Markov Blanket for a double check.

Cost analysis The key component in the cost of COAT is the number of candidate factors that are proposed by LLMs from unstructured data. For each proposed candidate factor, the cost, $c_{\text{candidate}}$, is on evaluating the factor values on each of the N samples, i.e., $c_{\text{candidate}} = N \cdot c_{\text{sample}}$. For example, if the text data in each sample has an average length of 1K tokens (including the instruction prompt), the complete annotation of $N = 1000$ samples will cost 1M tokens (the number of output tokens can be ignored with a suitable prompt). This will cost about \$0.15 USD with GPT-4o-mini, or \$0.04 USD with Qwen-turbo.

The required number for candidate factors is determined by both LLMs' capacity and the nature of the underlying causal structure. Ideally, if all the proposed factors are valid, according to the first step in Algorithm 2, the minimal factor number is $|\text{MB}(\mathcal{U})| - |\mathcal{U}|$ for extending a PAG with original node set $\mathcal{U}$. These nodes in $\text{MB}(\mathcal{U})$ are necessary for refining the skeleton.

The upper bound of the number for candidate factors is influenced by the *long directed chains* in causal graph where each node on it is only connected with its unique predecessor and unique successor. For example, in the special case demonstrated by Figure 6, to reveal the arrow head from Y to A, COAT will extend the node set by $B_1, \cdots, B_N$, and then C and D. However, such a directed chain can be arbitrarily long, which will increase the cost of the algorithm.

Bootstrap Uncertainty Estimation We discuss how to provide uncertainty estimation for COAT-MB algorithm based on stability (Meinshausen and Bühlmann, 2010). The same idea can be applied to COAT-PAG and COAT-ADJ.

At each round t, with current numerical data $\mathcal{D}$, we perform B bootstrap resampling to estimate the *selection frequency* of each candidate factor $W_i \in \mathcal{W}$. Specifically, for the b-th bootstrap sample $\mathcal{D}^{(b)} \sim \text{Bootstrap}(\mathcal{D})$, we obtain an estimated Markov Blanket $\mathcal{M}^{(b)} = \mathcal{A}(\mathcal{D}^{(b)}, Y, \mathcal{W})$. The selection frequency for factor W_i is defined as:

$$f(W_i) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left[W_i \in \mathcal{M}^{(b)}\right], \tag{6}$$

where $\mathbb{I}[\cdot]$ is the indicator function. The frequency vector $\mathcal{F}_t = [f(W_1), \cdots, f(W_m)]^\top$ serves as an *uncertainty measure*, and the final Markov Blanket at round t is determined by a threshold $\tau \in (0, 1]$:

$$\mathcal{M}_t = \{W_i \in \mathcal{W} : f(W_i) \geq \tau\}. \tag{7}$$

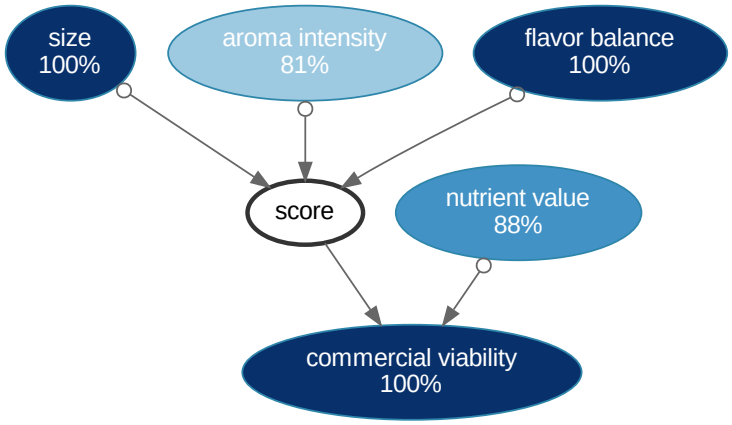


Figure 8: An illustration of COAT-MB with uncertainty estimation. Deeper colors of discovered factors indicate lower uncertainty.

In Figure 8, we present the proposed factors and their uncertainty at the final round. We apply COAT-MB on AppleGastronome with $B = 16$, $\tau = 0.8$, and $|\mathcal{D}^{(b)}|/|\mathcal{D}| = 0.5$. The bootstrap procedure relies only on the sample values of the target ("score") and the proposed factors, e.g., "size". The computation load depends on the method for finding Markov Blanket.

Extensions to improve efficiency To reduce the required candidate factor number, and to also control the cost budget, we provide the following practical suggestions:

- **Incorporate multi-view unstructured data.** This will enlarge the pool of all possible factors that can be proposed by LLMs, and can reduce the length of the *long directed chains* mentioned above.
- **Utilize priori knowledge.** Sometimes it is possible to acquire reliable prior knowledge. For example, if we know that the unstructured data is generated before some nodes in the original PAG are generated, then these factors cannot be causes of the proposed factors.
- **Cost budget.** The loop at step 4 in Algorithm 2 can be terminated when the cost budget (on resources or time) is exhausted. The early stop will leave some arrow heads not oriented in the output PAG.

Extensions to experiment design In existing causal discovery literature, there are methods that leverage data sets from different domains or node sets (Triantafillou and Tsamardinos, 2015; Mooij et al., 2020; Huang et al., 2020), as well as strategies for optimal experimental design (Agrawal et al., 2019; Tigas et al., 2022; Toth et al., 2022).

Assumption	Notes
I. LLM Capability Assumptions	
Propose valid factors, i.e., $p_\Psi > 0$ and $C_\Psi > 0$ (Definition 3).	Validated in Figure 10 (c) and Table 15
Accurate annotation with independent noise (Assumption 1).	Validated in Figure 10 (a-b) and Table 5
II. Causal Discovery Assumptions (Spirtes et al., 1995)	
Faithfulness and Markov condition	Depends on the specific algorithm (e.g., FCI, LiNGAM)
Validity of conditional independence tests	

Table 2: Summary of Core Assumptions for COAT Methods.
Note: These assumptions are shared across COAT-MB, COAT-PAG, and COAT-ADJ.

In the experimental-design methods, the causal structure can be learned by actively collecting samples from an intervened distribution. For example, in Figure 7(b), one can apply an intervention on node Y to distinguish the causal directions in the two undetermined edges. Instead of collecting new samples, COAT proposes three new factors: C, D, and E from the paired unstructured data in existing samples, and therefore the causal directions can be revealed in the extended causal graph in Figure 7(e). The COAT framework can integrate with these approaches in new perspectives:

1. By expanding the subgraph around an intervention target, COAT can resolve undetermined edges using unstructured data, reducing the need for redundant experiments.

2. With the expanded node set and the refined causal structure, COAT delivers richer input to methods for experimental design, and ultimately yields more valuable experimental data sets.

4. Identifiability Analysis

In this section, we provide theoretical discussions and analyze the identifiability of different variants of COAT in the corresponding settings. As shown in Table 2, we summarize the key assumptions required by the three COAT methods, which are grouped into LLM capability assumptions and causal discovery assumptions.

4.1 On Markov Blanket

Given a new factor $\boldsymbol{w}_{k+1}$, with the current representation as

$$h_{[k]}(\boldsymbol{X}) = \Big(\boldsymbol{w}_1(\boldsymbol{X}), \boldsymbol{w}_2(\boldsymbol{X}), \cdots, \boldsymbol{w}_k(\boldsymbol{X}) \Big),$$

and COAT tests:

$$Y \not\perp\!\!\!\perp \boldsymbol{w}_{k+1}(\boldsymbol{X}) \mid h_{[k]}(\boldsymbol{X}). \tag{8}$$

COAT also requires the following usual condition about distribution and causal graph:

Assumption 1 (Faithful and Markov conditions, from Spirtes et al. (1993)) *For any disjoint non-empty subsets $A, B, C \subset \mathcal{W}^{\leq t} \cup \{Y\}$, A and B are d-separated by C on the causal graph iff $A \perp\!\!\!\perp B \mid C$ on the factors' distribution. All conditional independencies are preserved after factor parsing.*

The annotation from a poor model could introduce an additional "measurement error" on the true factor values, disturbing the true distribution and thus violating the above assumption. The measurement error injected through LLM-based parsing may be of independent interest to the literature of causal discovery (Glymour et al., 2019; Vowels et al., 2022). Empirically, as one shall see in Section 5.1, predominant LLMs can conduct effective factor annotation. If Assumption 1 holds, the conditional mutual information between Y and $\boldsymbol{X}$ given the desired factors decreases:

Proposition 2 *Under Assumption 1, if condition 8 holds, then for Markov Blanket $\mathcal{S} \subseteq [k+1]$ of Y, i.e., $Y \perp\!\!\!\perp h_{[k+1]\setminus\mathcal{S}}(\boldsymbol{X}) \mid h_{\mathcal{S}}(\boldsymbol{X})$, we have the following about conditional mutual information:*

$$I\Big(Y; \boldsymbol{X} \mid h_{\mathcal{S}}(\boldsymbol{X})\Big) = I\Big(Y; \boldsymbol{X} \mid h_{[k+1]}(\boldsymbol{X})\Big) < I\Big(Y; \boldsymbol{X} \mid h_{[k]}(\boldsymbol{X})\Big). \tag{9}$$

We provide an initial exploration under what conditions LLMs can identify target-related factors and how the ability of an LLM influences this procedure.

Definition 3 (Ability of LLMs) *Given a suitable prompt about current factors and data, the LLM Ψ has non-zero probability $p_\Psi > 0$ to propose a new factor $\boldsymbol{w}_{k+1}$ that satisfies condition 8 and*

$$\frac{I(Y; \boldsymbol{X} \mid h_{[k+1]}(\boldsymbol{X}))}{I(Y; \boldsymbol{X} \mid h_{[k]}(\boldsymbol{X}))} < 1 - C_\Psi, \tag{10}$$

for some positive constant C_Ψ whenever $I\left(Y; \boldsymbol{X} \mid h_{[k]}(\boldsymbol{X})\right) > 0$. Note that $h_{[0]}(\boldsymbol{X}) = \phi$, hence we also use $I\left(Y; \boldsymbol{X} \mid h_{[0]}(\boldsymbol{X})\right)$ to refer to $I(Y; \boldsymbol{X})$. We use p instead of p_Ψ when the context is clear.

We further explain the intuition behind Def 3: the *Perception Score* p captures the LLM's responsiveness to the given prompts and the feedback; the *Capacity Score* C_Ψ captures the quality of the factors proposed by the LLM. Empirically, the two scores are used to estimate the abilities of the predominant LLMs in Section 5.1. Theoretically, we use them to characterize the influence of prompt templates, the LLM responsiveness, and the quality of factors on the performance of COAT:

Proposition 4 (Characterization for Factor Identification Process) *With Assumption 1, for any small number $\epsilon, \delta \in (0, \frac{1}{2})$, perception score $p > 0$, capacity score $C_\Psi > 0$, with t COAT rounds that*

$$\sqrt{t} > \frac{|z_\delta|\sqrt{1-p}}{2\sqrt{p}} \left(1 + \sqrt{1 + \frac{4 \log \epsilon}{z_\delta^2 (1-p) \log (1 - C_\Psi)}}\right), \tag{11}$$

where z_δ is the δ-quantile of the standard normal distribution, we have

$$\Pr\left(\frac{I(Y; \boldsymbol{X} \mid h_{\leq t}(\boldsymbol{X}))}{I(Y; \boldsymbol{X})} < \epsilon\right) \geq 1 - \delta. \tag{12}$$

The proof is given in Appendix B. Prop. 4 gives a guarantee on identifying a Markov Blanket. Intuitively, Prop. 4 also characterizes the influence of prompt templates, the LLM responsiveness, and the quality of factors on the performance of COAT via the two proposed measures: p and C_Ψ. When both of them are positive, COAT will converge exponentially:

Proposition 5 (Rate of Convergence) *With Assumption 1, for any small number $\epsilon, \delta \in (0, \frac{1}{2})$, perception score $p > 0$, capacity score $C_\Psi > 0$, after t COAT rounds, the following inequality holds with probability at least $1 - \delta$:*

$$\frac{I(Y; \boldsymbol{X} \mid h_{\leq t}(\boldsymbol{X}))}{I(Y; \boldsymbol{X})} \leq \left(\frac{1}{1 - C_\Psi}\right)^{-tp - z_\delta \sqrt{tp(1-p)}} \tag{13}$$

Finite-Sample Error Control. The previous discussion assumes perfect conditional independence oracles. Here we analyze the Type-I error rate with a finite sample of size n. Assume that W is the candidate factor to be tested, $\mathcal{M}$ is the current choice of Markov blanket of the target variable Y. The following proposition quantifies the sample size required to simultaneously control the Type I error (false positive, probability α) and the Type II error (false negative, probability β) under Gaussianity.

Proposition 6 (Sample size analysis for Markov blanket) *Assume that the joint distribution of $(Y, \mathcal{M}, W)$ is multivariate Gaussian. Let ρ be the true partial correlation between target Y and candidate factor W given the current choice of Markov blanket $\mathcal{M}$. $\rho \neq 0$ indicates that W carries additional information about Y beyond $\mathcal{M}$. Suppose a conditional independence test based on the Fisher z-transform (Fisher, 1921) of ρ is performed at significance level $\alpha \in (0, 1)$. Then, to detect the dependence with power $1 - \beta$, the sample size n must satisfy*

$$n \;\geq\; \frac{4\,(z_{1-\alpha/2} + z_{1-\beta})^2}{\left[\ln\left(\frac{1+\rho}{1-\rho}\right)\right]^2} \;+\; |\mathcal{M}| \;+\; 3, \tag{14}$$

where z_q denotes the q-quantile of the standard normal distribution and $|\mathcal{M}|$ is the cardinality of the conditioning set.

The proof is given in Appendix B. The above bound reveals that the required sample size grows linearly with the cardinality of the current Markov blanket $|\mathcal{M}|$. Intuitively, conditioning on more variables consumes additional degrees of freedom, which increases the variance of the estimation. In Section 3.6, we provide the bootstrap stability selection procedure and show the stability of COAT.

Causal structure identification It is clear that LLMs are not involved in the causal discovery process, which is mainly executed by causal discovery methods such as FCI. Therefore, the CD guarantees the identifiability of the final causal graph over the LLM-proposed factors. The concrete assumptions required for identifiability depend on the specific CD used in COAT. For instance, the FCI algorithm requires faithfulness of the data distribution with respect to the true causal graph (Spirtes et al., 1995). In our experiments, we also verify that the structured data annotated by LLMs has a high accuracy and little noise, which is friendly to the CD assumption. In general, one could switch to another CD in COAT, while

using different CD may require different assumptions. For example, LiNGAM (Shimizu et al., 2006) requires the relations among variables to be linear and non-Gaussian models. Empirically, we find that COAT with LiNGAM works very well (Appendix D).

4.2 On Partial Ancestral Graphs

In this section, we analyze how COAT-PAG improves causal structure identification by leveraging unstructured data $\boldsymbol{X}$. As introduced in Section 3.4, COAT-PAG extends the original structured variables $\mathcal{U}$ by defining a new set of variables $\mathcal{W}$ from $\boldsymbol{X}$. The augmented node set enables more conditional independence test involving nodes in $\mathcal{U}$, and therefore yields a more informative partial ancestral graph (PAG). We will demonstrate: in the PAG $\mathcal{P}_{\mathcal{V}}$ over the extended node set $\mathcal{V} \triangleq \mathcal{W} \cup \mathcal{U}$, the information in $\boldsymbol{X}$ will maximally refine the causal structure among variables from $\mathcal{U}$.

We start with the following necessary formulation. Let $\mathcal{O}$ be the set of all possible factors that can be defined based on $\boldsymbol{X}$, and we have $\mathcal{W} \subseteq \mathcal{O}$. Then, the maximally observable node set is $\mathcal{E} \triangleq \mathcal{O} \cup \mathcal{U}$. We expect that the causal structure among $\mathcal{U}$ in $\mathcal{P}_{\mathcal{E}}$ will be preserved in $\mathcal{P}_{\mathcal{V}}$. We formalize this idea in the following definition:

Definition 7 *For three vertex sets $\mathcal{U} \subseteq \mathcal{V} \subseteq \mathcal{E}$, and let $\mathcal{P}_{\mathcal{V}}$ and $\mathcal{P}_{\mathcal{E}}$ be the maximally informative PAG for $\mathcal{V}$ and $\mathcal{E}$ respectively. We say arrow heads (or arrow tails) in $\mathcal{P}_{\mathcal{E}}$ are preserved by $\mathcal{P}_{\mathcal{V}}$ within $\mathcal{U}$, if for $\alpha, \beta \in \mathcal{U}$:*

1. *α and β are adjacent in $\mathcal{P}_{\mathcal{V}}$ if and only if they are adjacent in $\mathcal{P}_{\mathcal{E}}$; and*
2. *$\alpha \leftarrow\!* \beta$ (or $\alpha -\!* \beta$) occurs in $\mathcal{P}_{\mathcal{V}}$ if and only if it occurs in $\mathcal{P}_{\mathcal{E}}$.*

In the next proposition, we state under what condition, one can have arrow head preservation in a node subset $\mathcal{U}$ in the PAG augmented by Algorithm 2.

Proposition 8 *Let $\mathcal{W}$ be the node set returned by applying Algorithm 2 on $(\boldsymbol{X}, \mathcal{U})$. If a set $\mathcal{U}'$ satisfies MB($\mathcal{U}'$) $\subseteq \mathcal{V}$, then arrow heads in $\mathcal{P}_{\mathcal{E}}$ are preserved by $\mathcal{P}_{\mathcal{V}}$ within $\mathcal{U}'$.*

As a special case, by the first step in Algorithm 2, one can see that the original structured variable set $\mathcal{U}$ satisfies MB($\mathcal{U}$) $\subseteq \mathcal{V}$, therefore, by Proposition 8, arrow heads in $\mathcal{P}_{\mathcal{E}}$ are preserved by $\mathcal{P}_{\mathcal{V}}$ within $\mathcal{U}$. The following result will be used later. It says that the edges in discriminating path in desired $\boldsymbol{U}$ can be clearly oriented.

Corollary 9 *Let $\mathcal{W}$ be the node set returned by applying Algorithm 2 on $(\boldsymbol{X}, \mathcal{U})$, and MB($\mathcal{U}'$) $\subseteq \mathcal{V}$. Let u be a discriminating path between θ and γ in $\mathcal{P}_{\mathcal{E}}$ within $\mathcal{U}'$, then for any node α between θ and γ that is adjacent to γ, the edge is either $\alpha \rightarrow \gamma$ or $\alpha \leftrightarrow \gamma$ in $\mathcal{P}_{\mathcal{V}}$.*

Proof According to Proposition 8, the arrow heads in $\mathcal{P}_{\mathcal{E}}$ are preserved by $\mathcal{P}_{\mathcal{V}}$ within $\mathcal{U}'$. If the corollary doesn't hold, we have $\alpha \circ\!\!\rightarrow \gamma$ in $\mathcal{P}_{\mathcal{V}}$, then either of the FCI orientation rules $\mathcal{R}1$ and $\mathcal{R}4$ from Zhang (2008a) can be applied to orient $\alpha \circ\!\!\rightarrow \gamma$. However, this contradicts with that $\mathcal{P}_{\mathcal{V}}$ is maximally informative. ■

4.3 On Adjustment Sets

To begin with, we introduce the essential concept and definition:

Definition 10 (Visibility (Zhang, 2008b; Perković et al., 2018)) *Let X, Y be nodes in a mixed graph $\mathcal{G}$.*

- *If $\mathcal{G}$ is a DAG or a CPDAG, any directed edge $X \to Y$ is said to be **visible**.*
- *If $\mathcal{G}$ is a MAG or a PAG, a directed edge $X \to Y$ is **visible** if there exists a node C (where $C \neq X$ and $C \neq Y$) such that:*
  1. *C is not adjacent to Y; and*
  2. *there is an edge between C and X with an arrowhead pointing into X (i.e., $C \to X$ or $C \leftrightarrow X$), **or** there is a collider path from C to X where:*
     - *every edge on the path has an arrowhead pointing into X; and*
     - *every non-endpoint vertex on this path is a parent of Y.*
- *Otherwise, the directed edge $X \to Y$ is called **invisible**.*

Definition 11 (Amenability (Perković et al., 2018)) *For $T, Y \in \mathcal{V}$, $\mathcal{P}_{\mathcal{V}}$ is **amenable** relative to (T, Y) if every proper possibly directed path from T to Y in $\mathcal{P}_{\mathcal{V}}$ starts with a visible edge out of T.*

We consider the Generalized adjustment criterion (Perković et al., 2018) defined as follow:

Definition 12 (Generalized adjustment criterion (Perković et al., 2018)) *Let $\mathcal{Z} \subseteq \mathcal{V} \setminus \{T, Y\}$ be a node set in $\mathcal{P}_{\mathcal{V}}$. $\mathcal{Z}$ satisfies the generalized adjustment criterion if the following three conditions hold:*

*(**Amenability**) $\mathcal{P}_{\mathcal{V}}$ is amenable relative to (T, Y); and*

*(**Forbidden Set**) $\mathcal{Z} \cap Forb(T, Y, \mathcal{P}_{\mathcal{V}}) = \emptyset$; and*

*(**Blocking**) all proper definite status non-causal paths from T to Y are blocked by $\mathcal{Z}$.*

Note that $Forb(T, Y, \mathcal{P}_{\mathcal{V}}) := \cup_{\alpha \in S(T,Y,\mathcal{P}_{\mathcal{V}})} PossDe(\alpha, \mathcal{P}_{\mathcal{V}})$, and $S(T, Y, \mathcal{P}_{\mathcal{V}})$ is the set of all nodes in a proper directed path from T to Y in $\mathcal{P}_{\mathcal{V}}$ (exclude T itself).

For the first key condition, i.e., Amenability, we show the Algorithm 3 can produce a PAG $\mathcal{P}_{\mathcal{V}}$ that is amenable to (T, Y) when $\mathcal{P}_{\mathcal{E}}$ is amenable to (T, Y). If $\mathcal{P}_{\mathcal{E}}$ is not amenable, that means we cannot find an adjustment set even if all possible factors from given unstructured data are given.

Proposition 13 *Let $\mathcal{W}$ be the node set returned by Algorithm 3 on T and Y. If $\mathcal{P}_{\mathcal{E}}$ is amenable relative to (T, Y), then $\mathcal{P}_{\mathcal{V}}$ is amenable relative to (T, Y).*

The next proposition will be used in later proof. It says that we can identify some special arrow tails in the PAG $\mathcal{P}_{\mathcal{V}}$ produced by Algorithm 3.

Proposition 14 *Let $\mathcal{W}$ be the node set returned by Algorithm 3 on T and Y. If α and γ are two adjacent nodes in a possibly directed path from T to Y in $\mathcal{P}_{\mathcal{V}}$, then the arrow tails in $\mathcal{P}_{\mathcal{E}}$ are preserved in $\mathcal{P}_{\mathcal{V}}$ within $\{\alpha, \gamma\}$.*

With the Proposition 14, we can identify definite status paths between T and Y. A definite status path consists of colliders and definite non-colliders, where a definite non-collider means there is at least one edge out of this node. Therefore, we need to identify arrow tails by Proposition 14.

Corollary 15 *Let $\mathcal{W}$ be the node set returned by Algorithm 3 on T and Y. Let $\boldsymbol{u}$ be a path from T to Y in $\mathcal{P}_{\mathcal{E}}$. $\boldsymbol{u}$ is a definite status in $\mathcal{P}_{\mathcal{V}}$ if and only if it is definite status in $\mathcal{P}_{\mathcal{E}}$.*

Proof We have $\boldsymbol{u} \subseteq \mathrm{MB}^{(2)}(\boldsymbol{u}) \subseteq \mathcal{V}$ by the Loop at step 3 in Algorithm 3. By Proposition 8 and 14, its arrow tails and heads are preserved. ■

With above preparation, in the following proposition, we show that if a node set $\mathcal{U}'$ have its Markov Blanket in the PAG $\mathcal{P}_{\mathcal{V}}$ produced by Algorithm 3, an adjustment set can be found if it is within the node set $\mathcal{U}'$. As a special case, by the first step in Algorithm 3, one can see that the original structured variable set $\mathcal{U}$ satisfies $\mathrm{MB}(\mathcal{U}) \subseteq \mathcal{V}$, therefore, it must satisfy the condition described in Proposition 16. In practice, one can keep applying Algorithm 3 multiple times to expand $\mathcal{U}$ for searching possible adjustment sets.

Proposition 16 (Adjustment Sets Preservation) *Let $\mathcal{W}$ be the node set returned by Algorithm 3 on (T, Y). If $\mathcal{Z} \subseteq \mathcal{U}'$ s.t. $MB(\mathcal{U}') \subseteq \mathcal{V}$, then $\mathcal{Z}$ is an adjustment set relative to (T, Y) in $\mathcal{P}_{\mathcal{V}}$, if and only if, it is an adjustment set relative to (T, Y) in $\mathcal{P}_{\mathcal{E}}$.*

5. Empirical Analysis of COAT

We evaluate the proposed COAT framework in multiple settings with different data sets.

- For the identifiability assumptions related to LLM capability (Definition 3), we use synthetic data (Section 5.1) to examine the capabilities of 10 predominant LLMs including GPT-4o (OpenAI, 2024), Claude-3-Opus (Anthropic, 2024), LLaMA3-70b (MetaAI, 2024), and Mistral-Large (Jiang et al., 2024). The results are shown in Figure 10(c).
- For COAT-MB method (Algorithm 1), we conduct experiments on one synthetic benchmark (Section 5.1) to evaluate the quality of the proposed Markov blanket factors from textual reviews; furthermore, we use one realistic clinical data set (Section 5.3) to draw Neuropathic factors using external tools and symptom-level diagnosis.
- For COAT-PAG method (Algorithm 2), we conduct experiments on one synthetic benchmark (Section 5.2) to determine the causal direction between two correlated variables. The direction is not identifiable within the partial ancestral graph on the input node set unless suitable factors are proposed.

- For COAT-ADJ method (Algorithm 3), we conduct experiments on two realistic data sets about marketing treatment with customers' social media posts (Section 5.4). The quality of adjustment sets are evaluated by the adjusted causal effect.

- We evaluate the annotation error in Section 5.1 (Figure 10(a) and 10(b)); and the hyperparameter sensitivity in Section 5.5.

5.1 Eliciting Markov Blanket

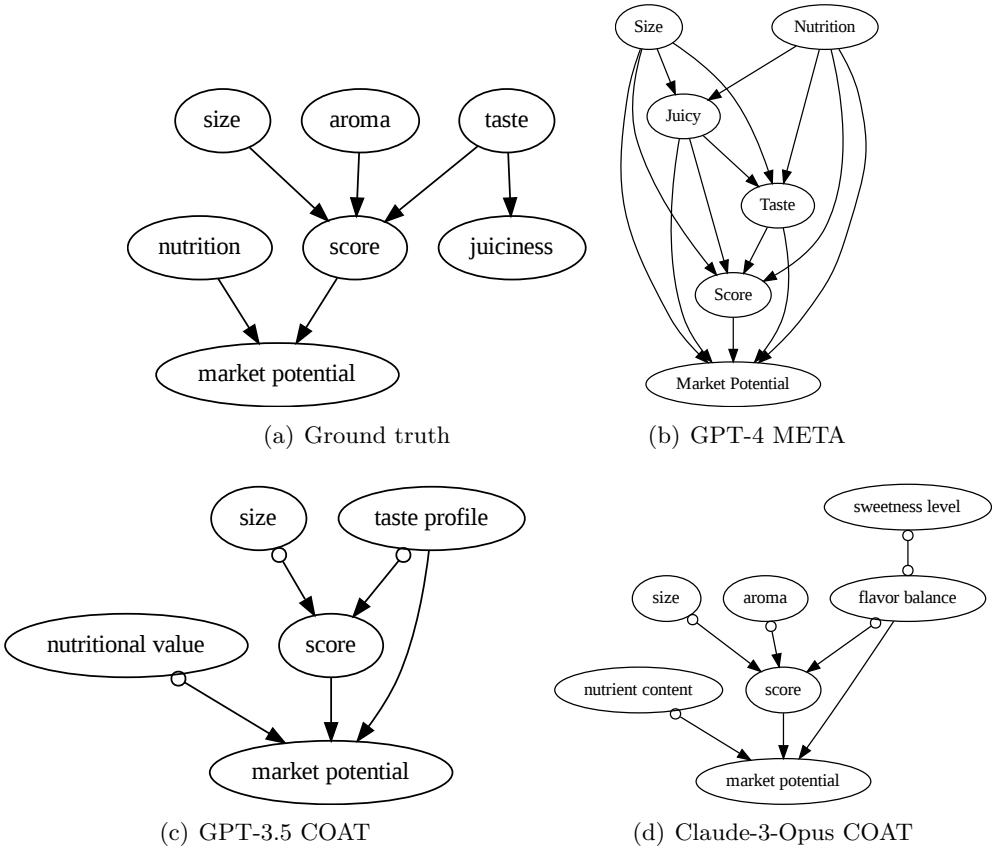
We consider the scenario where a rating score Y will be assigned to each apple by gastronomes. Each apple has its own attributes, including size, aroma, and taste. Each gastronome pays attention to a subset of the above three attributes. They will write a review according to their preference and give the rating score. As shown in Figure 9(a), we prepare different high-level factors: 3 parents of Y (size, aroma, and taste), one child of Y (market potential), and one spouse of Y (nutrition value). These factors form a Markov blanket of Y. In addition, we also prepared one disturbing factor, juiciness, that is related to Y but not a part of this blanket. A good method is expected to propose the five high-level factors (up to semantical meanings) and exclude the disturbing factor.

Setup and Benchmark Construction For each sample, we first generate the numerical values of each factor according to Figure 9(a). These are ground-truth values and will not be accessible to the algorithm. Then, we generate the according review using GPT-4 by feeding the predefined interpretation of each factor value. For example, taste $= 1$ means the apple is sweet, taste $= -1$ means the apple is sour, and taste $= 0$ means the gastronome doesn't care the taste. We generated 200 samples for LLMs' analysis and annotation. In appendix, we show the concrete prompt template for this procedure in Figure 28, and several review examples in Figure 29.

Baselines and Evaluation In this setting, we mainly employ two different uses of LLMs as the baselines: **META** is the zero-shot factor proposal given only the context to LLMs; and **DATA** additionally gives some samples of raw observations, which is an ablation of COAT without the *feedback* module, i.e., only one COAT round. For causal relation recovery, we follow Kiciman et al. (2023) that prompt LLMs to reason for the causal direction of each pair of the discovered variables by **DATA**.

We evaluate the ability on factor proposal based on three metrics: *MB*, *NMB*, and *OT*. *MB* means the desired factor forming the Markov Blanket of Y. The ground truth of the benchmark is given in Figure 9(a). The *MB* factors include `size`, `aroma`, `taste`, `market potential`, and `nutrition`. *NMB* means the undesired factor relevant to data but not in *MB*. In this benchmark, it is `juiciness`. *OT* means the unexpected factors irrelevant to data. They are factors that are outside of the six factors mentioned. For example, an LLM may propose `color` as a factor whose information can be rarely observed in samples. We also present the corresponding recall, precision, and F1 with respect to MB(Y).

Finding 1: COAT reliably elicits Markov Blanket Empirically, LLMs with CoT can be aware of high-level factors behind data (lower *OT* than *META*) but still struggle to distinguish the desired factors in Markov Blanket (higher *NMB* than COAT). COAT is more resistant to the "disturbing" factor, which is supported by the lower *NMB* column. COAT



(a) Ground truth (b) GPT-4 META

(c) GPT-3.5 COAT (d) Claude-3-Opus COAT

Figure 9: The discovered causal graphs in AppleGastronome. Compared to the ground truth results, directly adopting LLMs to reason the causal relations can easily elicit many false positive edges. In contrast, the relations recovered by COAT have a high precision and recall. The directed edge between "taste" and "juiciness" can not be recovered by COAT because of the limitations of FCI.

filters out irrelevant factors from LLMs' prior knowledge that are not reflected by the data, which is supported by the lower *OT* column. COAT robustly encourages LLM to find more expected factors through the feedback, which is supported by the higher *MB* column.

Finding 2: COAT reliably recovers the causal relationships We present quantitative and qualitative results in Table 4 and Figure 9 respectively. Compared to directly adopting LLMs to reason the causal relations, COAT significantly boosts the causal relation recovery. Meanwhile, COAT maintains high performances based on various LLMs, which further demonstrates the effectiveness of the causal feedback in COAT to improve the robustness of this system. In fact, the causal feedback focuses on making maximal use of the rich knowledge of LLMs, and reducing the reliance on the reasoning capabilities of different LLMs, to assist with causal discovery.

LLMs	Method	MB	NMB	OT	Recall	Precision	F1
GPT-4o	META	3.00±*0.82*	1.00±*0.00*	4.00±*0.82*	0.60±*0.16*	0.37±*0.09*	0.46±*0.12*
	DATA	3.00±*0.00*	0.67±*0.47*	0.00±*0.00*	0.60±*0.00*	0.83±*0.12*	0.69±*0.04*
	DATA+CoT	4.67±*0.58*	1.00±*0.00*	0.33±*0.58*	0.93±*0.12*	0.78±*0.10*	0.85±*0.10*
	COAT	4.00±*0.82*	0.00±*0.00*	0.00±*0.00*	0.80±*0.16*	1.00±*0.00*	0.88±*0.10*
GPT-4	META	2.67±*0.94*	0.67±*0.47*	2.33±*0.47*	0.53±*0.19*	0.46±*0.08*	0.49±*0.13*
	DATA	3.00±*0.00*	0.33±*0.47*	0.00±*0.00*	0.60±*0.00*	0.92±*0.12*	0.72±*0.04*
	DATA+CoT	4.33±*0.58*	0.83±*0.29*	0.17±*0.29*	0.87±*0.12*	0.81±*0.02*	0.84±*0.06*
	COAT	4.00±*0.82*	0.33±*0.47*	0.00±*0.00*	0.80±*0.16*	0.93±*0.09*	0.85±*0.11*
GPT-3.5	META	3.33±*1.25*	0.33±*0.47*	4.33±*1.25*	0.67±*0.25*	0.42±*0.12*	0.51±*0.15*
	DATA	2.67±*0.47*	0.67±*0.47*	0.00±*0.00*	0.53±*0.09*	0.81±*0.14*	0.64±*0.10*
	DATA+CoT	5.00±*0.00*	1.00±*0.00*	1.33±*0.58*	1.00±*0.00*	0.68±*0.05*	0.81±*0.04*
	COAT	3.67±*0.47*	0.00±*0.00*	0.00±*0.00*	0.73±*0.09*	1.00±*0.00*	0.84±*0.07*
Mistral-Large	META	2.00±*0.82*	0.67±*0.47*	2.67±*0.94*	0.40±*0.16*	0.37±*0.09*	0.38±*0.12*
	DATA	3.00±*0.00*	0.33±*0.47*	0.00±*0.00*	0.60±*0.00*	0.92±*0.12*	0.72±*0.04*
	DATA+CoT	4.33±*0.58*	1.00±*0.00*	0.67±*0.58*	0.87±*0.12*	0.73±*0.07*	0.79±*0.05*
	COAT	4.33±*0.47*	0.00±*0.00*	0.00±*0.00*	0.87±*0.09*	1.00±*0.00*	0.93±*0.05*
Mistral-Medium	META	3.00±*0.00*	0.67±*0.47*	1.67±*1.25*	0.60±*0.00*	0.59±*0.13*	0.59±*0.07*
	DATA	3.00±*0.00*	0.67±*0.47*	0.00±*0.00*	0.60±*0.00*	0.83±*0.12*	0.69±*0.04*
	DATA+CoT	4.33±*0.58*	1.00±*0.00*	0.67±*0.58*	0.87±*0.12*	0.73±*0.07*	0.79±*0.05*
	COAT	4.67±*0.47*	0.00±*0.00*	0.00±*0.00*	0.93±*0.09*	1.00±*0.00*	0.96±*0.05*
LLaMA3-70b	META	2.67±*0.47*	0.33±*0.47*	4.67±*0.47*	0.53±*0.09*	0.35±*0.06*	0.42±*0.07*
	DATA	2.67±*1.25*	0.33±*0.47*	0.00±*0.00*	0.53±*0.25*	0.93±*0.09*	0.63±*0.21*
	DATA+CoT	2.67±*0.58*	0.67±*0.58*	1.00±*0.00*	0.53±*0.12*	0.62±*0.13*	0.57±*0.11*
	COAT	3.67±*0.47*	0.33±*0.47*	0.00±*0.00*	0.73±*0.09*	0.93±*0.09*	0.81±*0.06*
LLaMA2-70b	META	2.33±*0.47*	0.67±*0.47*	4.67±*1.25*	0.47±*0.09*	0.32±*0.07*	0.37±*0.06*
	DATA	2.33±*0.94*	0.67±*0.47*	0.00±*0.00*	0.47±*0.19*	0.75±*0.20*	0.57±*0.20*
	DATA+CoT	3.00±*1.73*	0.67±*0.58*	0.33±*0.58*	0.60±*0.35*	0.71±*0.34*	0.65±*0.35*
	COAT	3.00±*0.00*	0.67±*0.47*	0.00±*0.00*	0.60±*0.00*	0.83±*0.12*	0.69±*0.04*
Qwen-1.5-110B	META	2.00±*0.00*	1.00±*0.00*	4.00±*0.00*	0.40±*0.00*	0.29±*0.00*	0.33±*0.00*
	DATA	3.00±*0.82*	1.00±*0.00*	0.00±*0.00*	0.60±*0.16*	0.74±*0.05*	0.66±*0.12*
	DATA+CoT	3.67±*1.53*	0.67±*0.58*	0.00±*0.00*	0.73±*0.31*	0.83±*0.17*	0.77±*0.23*
	COAT	4.00±*0.82*	0.33±*0.47*	0.00±*0.00*	0.80±*0.16*	0.93±*0.09*	0.85±*0.11*
DeepSeek-V2	META	2.33±*0.47*	1.00±*0.00*	3.00±*0.82*	0.47±*0.09*	0.37±*0.03*	0.41±*0.04*
	DATA	3.33±*0.47*	0.67±*0.47*	0.00±*0.00*	0.67±*0.09*	0.85±*0.11*	0.74±*0.05*
	DATA+CoT[1]	5.00±*0.00*	1.00±*0.00*	0.67±*0.58*	1.00±*0.00*	0.75±*0.07*	0.86±*0.04*
	COAT	3.67±*0.47*	0.67±*0.47*	0.00±*0.00*	0.73±*0.09*	0.87±*0.09*	0.78±*0.02*
Claude-3-Opus	META	2.00±*0.00*	1.00±*0.00*	3.00±*1.41*	0.40±*0.00*	0.35±*0.07*	0.37±*0.04*
	DATA	3.33±*0.47*	0.33±*0.47*	0.00±*0.00*	0.67±*0.09*	0.93±*0.09*	0.77±*0.02*
	DATA+CoT	2.67±*0.58*	0.33±*0.58*	1.00±*1.00*	0.53±*0.12*	0.67±*0.14*	0.59±*0.13*
	COAT	5.00±*0.00*	0.33±*0.47*	0.00±*0.00*	1.00±*0.00*	0.94±*0.08*	0.97±*0.04*

Table 3: Full Results on the Apple Gastronome Benchmark.

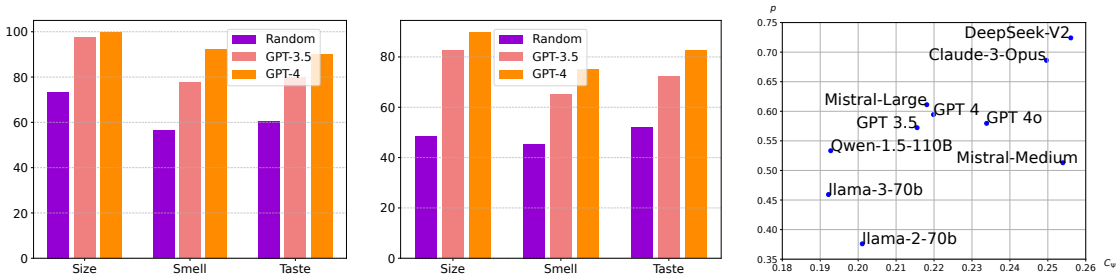
[1] *DeepSeek-V2 was no longer available when we added this baseline, so we used DeepSeek-V2.5 instead.*

LLM	Method	SHD	SID	Recall	Precision	F1
GPT-4o	pairwise	5.33±*2.62*	0.00±*0.00*	1.00±*0.00*	0.45±*0.07*	0.62±*0.07*
	COAT	1.00±*0.00*	0.00±*0.00*	1.00±*0.00*	0.79±*0.03*	0.89±*0.02*
GPT-4	pairwise	8.67±*2.05*	1.33±*1.89*	0.93±*0.09*	0.35±*0.05*	0.51±*0.06*
	COAT	2.33±*0.47*	1.00±*1.41*	0.93±*0.09*	0.70±*0.09*	0.79±*0.04*
GPT-3.5	pairwise	5.67±*4.03*	3.00±*2.94*	0.67±*0.24*	0.49±*0.36*	0.55±*0.32*
	COAT	2.00±*0.82*	2.67±*2.05*	0.75±*0.20*	0.63±*0.21*	0.68±*0.21*
Mistral-Large	pairwise	7.33±*1.89*	0.00±*0.00*	1.00±*0.00*	0.38±*0.03*	0.55±*0.03*
	COAT	2.00±*0.00*	1.33±*1.25*	0.85±*0.11*	0.74±*0.05*	0.78±*0.02*
Mistral-Medium	pairwise	6.00±*1.63*	2.00±*0.00*	0.51±*0.13*	0.37±*0.04*	0.42±*0.07*
	COAT	0.33±*0.47*	0.00±*0.00*	1.00±*0.00*	0.94±*0.08*	0.97±*0.04*
LLaMA3-70b	pairwise	4.00±*1.41*	0.33±*0.47*	0.92±*0.12*	0.51±*0.09*	0.65±*0.11*
	COAT	2.33±*0.47*	3.67±*1.25*	0.75±*0.00*	0.70±*0.07*	0.72±*0.04*
LLaMA2-70b	pairwise	3.33±*1.25*	1.00±*1.41*	0.89±*0.16*	0.48±*0.11*	0.62±*0.13*
	COAT	1.00±*0.82*	0.33±*0.47*	0.89±*0.16*	0.81±*0.14*	0.84±*0.14*
Qwen-1.5-110B	pairwise	6.67±*2.62*	2.33±*1.25*	0.62±*0.03*	0.38±*0.09*	0.47±*0.07*
	COAT	4.00±*1.63*	7.00±*4.32*	0.47±*0.34*	0.47±*0.34*	0.70±*0.10*
DeepSeek-V2	pairwise	5.00±*1.41*	1.33±*1.89*	0.93±*0.09*	0.47±*0.05*	0.62±*0.06*
	COAT	1.33±*1.25*	2.33±*1.70*	0.85±*0.11*	0.87±*0.19*	0.85±*0.14*
Claude-3-Opus	pairwise	7.33±*1.25*	0.33±*0.47*	0.93±*0.09*	0.40±*0.05*	0.56±*0.07*
	COAT	1.33±*0.47*	0.00±*0.00*	1.00±*0.00*	0.79±*0.06*	0.88±*0.04*

Table 4: Causal relation extraction results in AppleGastronome.

Finding 3: LLMs are competent in proposing potential high-level factors As discussed in Sec. 4, there are two crucial abilities for LLMs in identifying potential high-level factors. The first one is to be aware of the existence of potential factors, and the second is to synthesize and describe these factors. Inspired by this observation, we propose two novel metrics to quantify LLMs' causal ability: a) Perception that quantifies the ratio of *valid* factors (satisfying Prop. 2) proposed by LLMs in each round; b) Capacity that measures the effective mutual information drop in Assumption 3. As shown in Figure 10(c), LLMs differ largely on the perception score while comparably on the capacity score. In Table 15 (Appendix C.3), we also show the two metrics as well as the mutual information in each COAT iteration.

Finding 4: LLMs can conduct effective factor annotation As shown in Figure 10, both GPT-3.5 and GPT-4 annotate objective attributes well. Regarding subjective human preferences, the performances are still relatively high. Empirically, LLMs will not introduce new confounders. Table 5 presents the independence testing results between the annotation noises and the annotated features, and the noises themselves. It can be found that,



(a) Apple attributes Acc. (b) Preference matchness Acc. (c) LLM Ability Scatters

Figure 10: Quantitative evaluation of the causal capabilities of LLMs in COAT.

LLM	Test Object	T	P-value
GPT-4	Feature	0.2828	0.9997
	Noise	0.0327	0.9962
GPT-3.5	Feature	0.4803	0.0325
	Noise	0.0446	0.9962

Table 5: Independence tests of the annotation noises with annotated features and other noises in AppleGastronome.

the introduced noises are independent of the attributes, therefore, will not introduce much additional interference to the causal discovery procedure.

5.2 Eliciting Partial Ancestral Graph

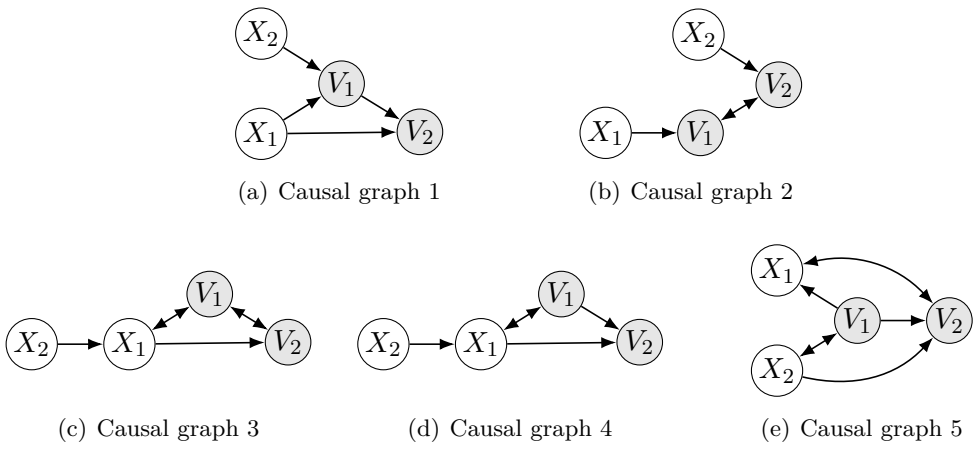
Similar to the previous setting, we consider the scenario where the rating score Y and one additional factor T (like taste) are observed. The PAG in such two-node case is always $T \circ\!\!-\!\!\circ Y$ if they are correlated, because we can only conduct one CI test given only two variables. And thus, we have little information about edge marks. The goal of this setting is to extend the node set from unstructured data, so that the skeleton and orientation in the original graph can be further refined. For example, in Figure 7(d) and Figure 7(e), by extending node E, the skeleton between node C and node D is refined, and the orientation on nodes B, Y, and A are identified.

Setup and Benchmark Construction As shown in Figure 11, we design five different causal graphs with respect to different FCI orientation rules. Each causal graph consists of 4 nodes: 2 numerical variables V_1 and V_2 that will be in the original PAG, and two factors X_1 and X_2 that are expected to be elicited from unstructured data. We consider two types of causal relationships between V_1 and V_2: (1) direct cause, i.e., $V1 \to V2$; and (2) latent confounding, i.e., $V1 \leftrightarrow V2$. Given X_1 and X_2, the edge marks can be exactly identified in all graphs except for graph 5 which is $V1 \circ\!\!\to V2$ in the extended PAG.

We choose V_2 to be score across all graphs, and randomly select 3 factors from Figure 9(a) to be the remaining nodes. Other factors will be generated as independent variables. The concrete specification for each graph is listed in Table 6. In this setting, we generate 2000

	V1	V2	X1	X2	FCI Orientation Rules
graph 1	Taste	Score	Nutrition	Market Potential	$\mathcal{R}0$, $\mathcal{R}1$
graph 2	Taste	Score	Juiciness	Size	$\mathcal{R}0$
graph 3	Smell	Score	Market Potential	Nutrition	$\mathcal{R}0$, $\mathcal{R}4$
graph 4	Juiciness	Score	Nutrition	Smell	$\mathcal{R}0$, $\mathcal{R}2$, $\mathcal{R}4$
graph 5	Market Potential	Score	Taste	Juiciness	$\mathcal{R}3$

Table 6: The variable specification for different PAGs.



(a) Causal graph 1 (b) Causal graph 2

(c) Causal graph 3 (d) Causal graph 4 (e) Causal graph 5

Figure 11: The five causal graphs used in the setting on extending PAG. The filled nodes are present in the original PAG.

samples using GLM-4-Flash, a light-weight LLM with a free API. The review generating prompt is constructed in the same way as the setting in Section 5.1.

Baselines and Evaluation We consider two baselines in this setting. The first considered one is **DATA+CoT** that prompts LLMs to decide the edge marks by analyzing a portion of samples from each value combination of V_1 and V_2. The second one is **DATA+FCI**, the ablation version of single round COAT without verification or iteration procedure.

For evaluation, we report the precision, recall, and F1 score for arrow heads identification between V_1 and V_2, and the purpose is to confirm our previous identifiability results. We also report the exact edge mark accuracy, to see whether each method can predict arrow head, arrow tail and circle mark in the edge between V_1 and V_2.

Finding 5: COAT reliably improves edge orientation by extending useful nodes We present the averaged result over all 5 graphs in Table 7(a). In the arrow-head metrics, COAT shows best performance and lowest standard deviation compared with the CoT baseline and its own ablation version. Compared with *DATA+COT*, *COAT* has significant improvement, indicating the effectiveness of COAT iteration.

Finding 6: LLMs' reasoning fails when reality deviates from its prior Instead of extending the node set in PAG, *DATA+CoT* improves orientation by LLMs' own reasoning ability with unstructured samples from each $(V1, V2)$ value combination. It can achieve competitive results when the causal graph aligns with commonsense, e.g., in graph 1, taste is the cause of the rating score. However, its performance can drop severely when the structure is different. For example, in graph 2, there is a latent confounder between taste and rating score; in graph 5, the market potential is the cause of the rating score. In the latter example, LLMs tend to believe the rating score determines the market potential. Therefore, it might be important to incorporate statistical methods to avoid subtle bias from prior.

(a) Average over graphs

		Edge Mark	**Arrow Head**		
		Accuracy	**Precision**	**Recall**	**F1**
DeepSeek-V3	DATA+CoT	48 ±*40*	52 ±*36*	54 ±*34*	53 ±*35*
	DATA+FCI	66 ±*19*	88 ±*11*	76 ±*22*	80 ±*17*
	COAT	88 ±*8*	96 ±*9*	88 ±*11*	91 ±*9*
GLM-4-Plus	DATA+CoT	60 ±*16*	74 ±*17*	70 ±*22*	69 ±*16*
	DATA+FCI	54 ±*30*	68 ±*27*	58 ±*25*	61 ±*24*
	COAT	68 ±*19*	76 ±*17*	74 ±*19*	75 ±*19*
GPT-4.1-mini	DATA+CoT	43 ±*24*	80 ±*45*	65 ±*42*	70 ±*41*
	DATA+FCI	63 ±*35*	73 ±*28*	66 ±*34*	66 ±*34*
	COAT	74 ±*24*	93 ±*11*	77 ±*26*	76 ±*25*

Table 7: Results on extending Partial Ancestral Graph.

(b) Graph 1

		Edge Mark	**Arrow Head**		
		Accuracy	**Precision**	**Recall**	**F1**
DeepSeek-V3	DATA+CoT	80 ±*45*	80 ±*45*	80 ±*45*	80 ±*45*
	DATA+FCI	80 ±*45*	80 ±*45*	80 ±*45*	80 ±*45*
	COAT	100 ±*0*	100 ±*0*	100 ±*0*	100 ±*0*
GLM-4-Plus	DATA+CoT	50 ±*0*	90 ±*22*	100 ±*0*	93 ±*15*
	DATA+FCI	80 ±*45*	80 ±*45*	80 ±*45*	80 ±*45*
	COAT	60 ±*55*	60 ±*55*	60 ±*55*	60 ±*55*
GPT-4.1-mini	DATA+CoT	50 ±*0*	100 ±*0*	100 ±*0*	100 ±*0*
	DATA+FCI	100 ±*0*	100 ±*0*	100 ±*0*	100 ±*0*
	COAT	100 ±*0*	100 ±*0*	100 ±*0*	100 ±*0*

Table 7: Results on extending Partial Ancestral Graph (continued).

(c) Graph 2

		Edge Mark	Arrow Head		
		Accuracy	**Precision**	**Recall**	**F1**
DeepSeek-V3	DATA+CoT	30 ±*45*	30 ±*45*	30 ±*45*	30 ±*45*
	DATA+FCI	80 ±*27*	100 ±*0*	80 ±*27*	87 ±*18*
	COAT	80 ±*27*	100 ±*0*	80 ±*27*	87 ±*18*
GLM-4-Plus	DATA+CoT	40 ±*42*	60 ±*55*	40 ±*42*	47 ±*45*
	DATA+FCI	90 ±*22*	100 ±*0*	90 ±*22*	93 ±*15*
	COAT	100 ±*0*	100 ±*0*	100 ±*0*	100 ±*0*
GPT-4.1-mini	DATA+CoT	50 ±*0*	100 ±*0*	50 ±*0*	67 ±*0*
	DATA+FCI	100 ±*0*	100 ±*0*	100 ±*0*	100 ±*0*
	COAT	100 ±*0*	100 ±*0*	100 ±*0*	100 ±*0*

Table 7: Results on extending Partial Ancestral Graph (continued).

(d) Graph 3

		Edge Mark	Arrow Head		
		Accuracy	**Precision**	**Recall**	**F1**
DeepSeek-V3	DATA+CoT	20 ±*45*	20 ±*45*	20 ±*45*	20 ±*45*
	DATA+FCI	40 ±*22*	80 ±*45*	40 ±*22*	53 ±*30*
	COAT	80 ±*27*	100 ±*0*	80 ±*27*	87 ±*18*
GLM-4-Plus	DATA+CoT	80 ±*27*	100 ±*0*	80 ±*27*	87 ±*18*
	DATA+FCI	40 ±*22*	80 ±*45*	40 ±*22*	53 ±*30*
	COAT	50 ±*50*	60 ±*55*	50 ±*50*	53 ±*51*
GPT-4.1-mini	DATA+CoT	63 ±*25*	100 ±*0*	75 ±*29*	83 ±*19*
	DATA+FCI	50 ±*41*	75 ±*50*	63 ±*48*	67 ±*47*
	COAT	50 ±*41*	75 ±*50*	63 ±*48*	67 ±*47*

Table 7: Results on extending Partial Ancestral Graph (continued).

Comparison between COAT-MB and COAT-PAG We empirically demonstrate the differences between the two algorithms in a variant of AppleGastronome data set which has a different causal structure as shown in Figure 12 (a). The initial variable set is {`score`, `taste`}. When applying COAT-MB on `taste`, it introduces one more factor `textual quality` that is semantically close to `juiciness`; however, the causal direction between `taste` and `score` is not identified, as the red edge shown in Figure 12 (b). On the contrary, applying COAT-PAG reveals the causal direction in Figure 12 (c), which is colored by green, by extending the uncertain factors in the graph. These results demonstrate that COAT-PAG can adaptively extend necessary factors that may be beyond the Markov Blanket to reveal causal information among the initial variables.

(e) Graph 4

		Edge Mark	Arrow Head		
		Accuracy	**Precision**	**Recall**	**F1**
DeepSeek-V3	DATA+CoT	100 ±0	100 ±0	100 ±0	100 ±0
	DATA+FCI	80 ±27	100 ±0	100 ±0	100 ±0
	COAT	90 ±22	100 ±0	100 ±0	100 ±0
GLM-4-Plus	DATA+CoT	70 ±45	80 ±45	70 ±45	73 ±43
	DATA+FCI	20 ±27	40 ±55	40 ±55	40 ±55
	COAT	60 ±42	80 ±45	80 ±45	80 ±45
GPT-4.1-mini	DATA+CoT	50 ±0	100 ±0	100 ±0	100 ±0
	DATA+FCI	40 ±55	40 ±55	40 ±55	40 ±55
	COAT	60 ±22	90 ±22	80 ±45	73 ±43

Table 7: Results on extending Partial Ancestral Graph (continued).

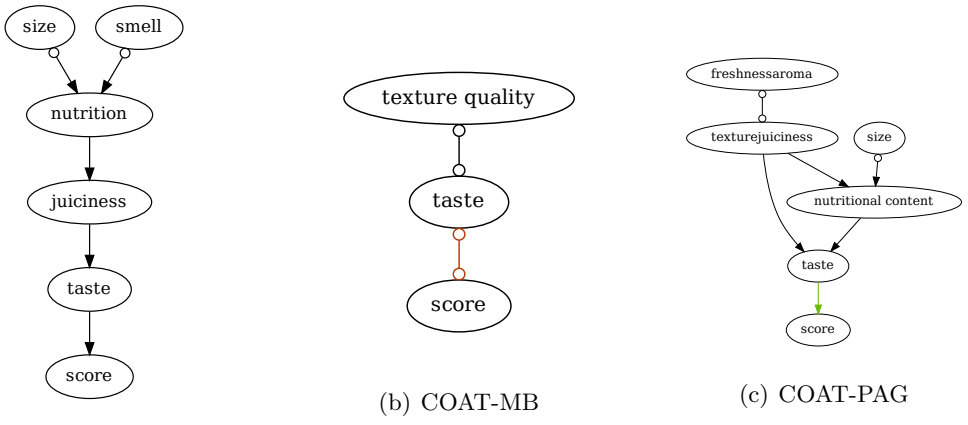
(f) Graph 5

		Edge Mark	Arrow Head		
		Accuracy	**Precision**	**Recall**	**F1**
DeepSeek-V3	DATA+CoT	10 ±22	30 ±45	40 ±55	33 ±47
	DATA+FCI	50 ±0	80 ±45	80 ±45	80 ±45
	COAT	90 ±22	80 ±45	80 ±45	80 ±45
GLM-4-Plus	DATA+CoT	60 ±55	60 ±55	60 ±55	60 ±55
	DATA+FCI	40 ±55	40 ±55	40 ±55	40 ±55
	COAT	70 ±45	80 ±45	80 ±45	80 ±45
GPT-4.1-mini	DATA+CoT	0 ±0	0 ±0	0 ±0	0 ±0
	DATA+FCI	25 ±29	50 ±58	25 ±50	25 ±50
	COAT	60 ±22	100 ±0	40 ±55	40 ±55

Table 7: Results on extending Partial Ancestral Graph (continued).

5.3 Evaluation on Neuropathic Data Set

Benchmark construction In the Neuropathic benchmark, we convert the data set into a clinical diagnosis task. In the original data set, there are three levels of causal variables, including the symptom level, radiculopathy level, and the pathophysiology level. In the experiments, we mainly consider the target variable of right shoulder impingement. When generating the clinical diagnosis notes as $\boldsymbol{x}$ using GPT-4, we avoid any mention of variables other than symptoms. We generated 100 samples for LLMs' analysis; since the number of possible factors is finite, we generate 1000 tabular data for CI tests.



(a) Ground Truth

(b) COAT-MB

(c) COAT-PAG

Figure 12: The comparison between COAT-MB and COAT-PAG in the causal discovery setting.

As we intend to leverage the Neuropathic benchmark to simulate the real-world diagnosis, after the factor proposal stage, we directly incorporate external tools to measure the values of the candidate factors. More details about the construction of the Neuropathic are given in Appendix C.5.

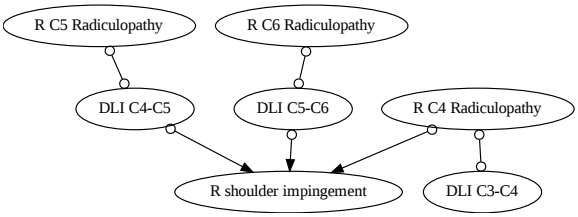
Evaluation and baselines In Neuropathic, we adopt a similar evaluation protocol and the baselines as in AppleGastronome. Nevertheless, due to the faithfulness issue of the original data set (Tu et al., 2019), for the evaluation of causal relation discovery, we mainly conduct a qualitative comparison between the ground truth that is faithful to the data, against the baselines and COAT.

Factor proposal The quantitative results on Neuropathic benchmark are given in Table 8. PA, AN, and OT refer to the parents, ancestors, and others, respectively. Accuracy and F1 measure the recovery of the causal ancestors. Similarly, we can find that COAT consistently outperforms all of the baselines regardless of which LLMs are incorporated. In particular, even with the weakest backbone model, i.e., LLaMA2-7b, COAT can still effectively leverage the intrinsic rich knowledge and beat the baselines with more powerful LLMs.

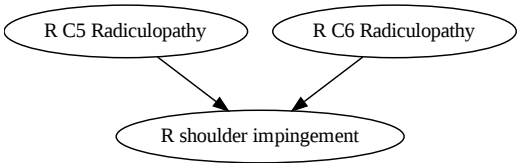
Causal relation recovery Figure 13(a) shows the causal graph obtained by FCI running on the original data, where we can find that several causal relations cannot hold on the data. As shown in Figure 13, when using LLMs to perform the reasoning, LLMs cannot identify the faithfulness issues. In contrast, COAT can imply faithful causal insights.

5.4 Eliciting Adjustment Set for Causal Inference

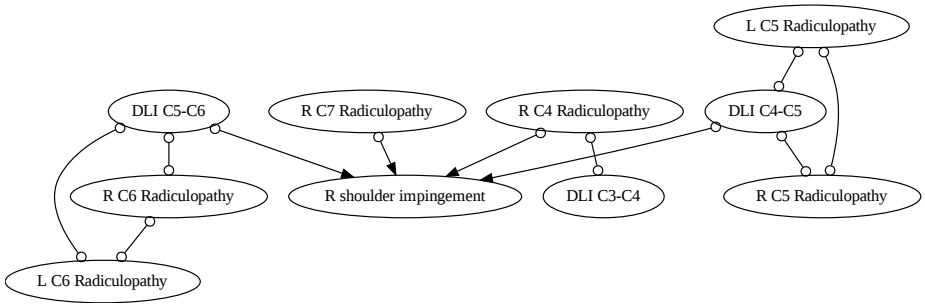
Benchmark and baselines We use two realistic data sets about marketing for estimating Average Treatment Effect. The original ones are randomized tabular data sets:



(a) Faithful ground truth



(b) GPT-4 reasoning



(c) GPT-4 COAT

Figure 13: The discovered causal graphs in Neuropathic. (b) shows the result based on directly prompting LLM to reason for the causal relations among all factors. Disconnected ones are dropped.

Hillstrom (2008): the effect of mailing on customers' visiting to a website; and Retail Hero (2019): the effect of Short Message on customers' purchasing a product. We follow the evaluation setting introduced by Dhawan et al. (2024), where $\boldsymbol{X}$ is a synthesized social-media post, and T and Y are confounded. In Hillstrom, the confounder is the `new customer indicator`, and in Retail Hero it is the `customer's age`.

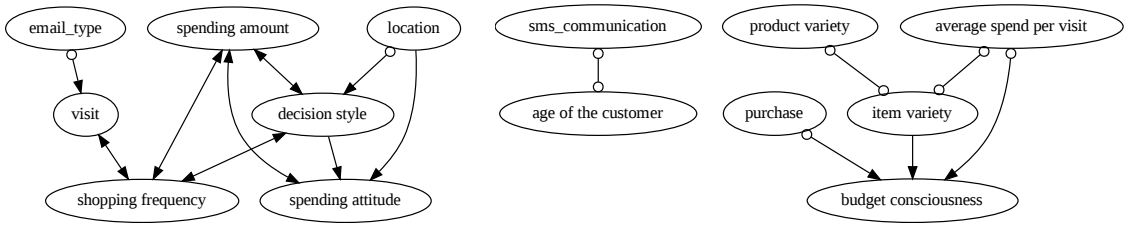
We consider two baselines: (1) **DATA+IPW**: LLMs will conduct reasoning to find the possible factors that will be utilized to estimate the ATE. (2) **DATA+FCI+IPW**: the proposed factors will be further verified by FCI algorithm based on the identified causal graphs.

In addition, we also include the human expert results as a reference. For the baseline of human expert results, a set of human-crafted factor descriptions is provided to predict the factor values by bag-of-word sentence encoders, or by LLMs (Dhawan et al., 2024). For COAT method, we do not include this prior knowledge. The input is only $(\boldsymbol{X}, T, Y)$. The factor descriptions will be identified by Algorithm 3 with access to an LLM.

Causal graphs interpretation The PAGs produced by COAT with GLM-4-Plus are displayed in Figure 14. In Figure 14(a), the direction between mailing and visit is confirmed by the elicited factors, while in Figure 14(b) the skeleton is missing, indicating a faithfulness issue in the data set. Interestingly, in Figure 14(a), COAT doesn't directly include the new customer indicator factors, but alternatively proposes some related ones like shopping frequency or spending amount, providing an unexpected point of view. In Figure 14(b),

LLM	Method	Factor Proposal				
		PA	AN	OT	Acc	F1
GPT-4	Meta	3	5	6	0.91	0.59
	DATA	2	2	0	0.95	0.50
	DATA+CoT	3	4	13	0.81	0.35
	COAT	3	6	3	0.96	0.80
GPT-3.5	Meta	3	5	6	0.91	0.59
	DATA	3	5	4	0.94	0.67
	DATA+CoT	2	2	3	0.91	0.36
	COAT	3	5	2	0.96	0.77
LLaMA2-70b	Meta	2	4	5	0.91	0.53
	DATA	3	3	1	0.95	0.60
	DATA+CoT	2	4	7	0.88	0.47
	COAT	3	6	2	0.97	0.86
LLaMA2-13b	Meta	1	3	6	0.88	0.40
	DATA	3	6	4	0.95	0.75
	DATA+CoT	0	1	10	0.81	0.12
	COAT	3	6	2	0.97	0.86
LLaMA2-7b	Meta	1	1	17	0.72	0.08
	DATA	3	6	3	0.96	0.80
	DATA+CoT	0	0	10	0.79	–
	COAT	3	6	2	0.97	0.86
Mistral-Medium	Meta	3	6	3	0.96	0.80
	DATA	3	3	2	0.94	0.66
	DATA+CoT	3	5	8	0.88	0.53
	COAT	3	6	2	0.97	0.86

Table 8: Factor proposal in Neuropathic.



(a) hillstrom (b) retailhero

Figure 14: The Causal Graph Identified by Algorithm 3 on two marketing data sets.

the customer's age factor is included in the graph, and is directly connected with the SMS variable, while its connection to the purchase variable is missing, indicating their correlation may not be very strong.

Factor Provider	Method	Hillstrom		Retail Hero	
		ATE (pp)	RMSE	ATE (pp)	RMSE
-	Ground Truth	6.09	-	3.32	-
Human Expert (Dhawan et al., 2024)	Bag-of-Words	7.57±*1.37*	2.23	2.61±*2.08*	2.42
	N-MC IPW	4.81±*0.80*	1.51	1.85±*2.01*	2.49
	N-IPW	5.23±*1.00*	1.32	3.83±*1.29*	1.39
DeepSeek-V3	DATA+IPW	4.31±*0.12*	1.78	-1.68±*1.72*	5.29
	DATA+FCI+IPW	4.73±*0.68*	1.52	2.29±*0.39*	1.10
	COAT(IPW)	5.70±*0.26*	**0.47**	3.16±*0.47*	**0.50**
GLM-4-Plus	DATA+IPW	0.18±*0.88*	5.98	73.21±*15.93*	71.68
	DATA+FCI+IPW	5.62±*0.31*	0.57	2.38±*0.51*	1.07
	COAT(IPW)	5.92±*0.17*	**0.23**	2.39±*0.25*	**0.96**
GPT-4.1-mini	DATA+IPW	3.70±*1.93*	3.07	1.04±*1.42*	2.68
	DATA+FCI+IPW	4.11±*0.37*	2.01	2.40±*0.62*	1.11
	COAT(IPW)	4.79±*1.38*	**1.90**	2.54±*0.06*	**0.78**

Table 9: ATE Estimation Results.

Note: ATE is reported in percentage points (pp), representing the absolute increase in conversion rate.

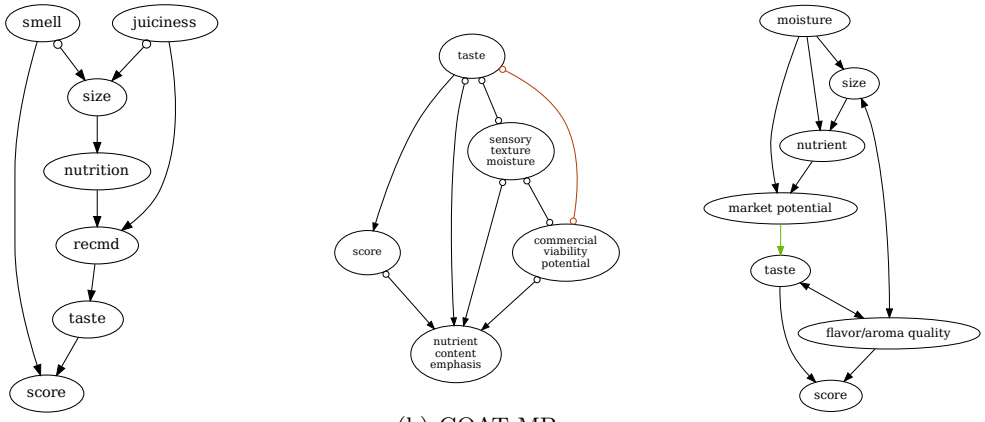
Treatment effect estimation The ATE estimation results are presented in Table 9. For each experiment, we choose the largest valid adjustment sets from the generated PAG for evaluation, if there are multiple such sets, we calculate their averaged ATE. Experiments are repeated 3 times. Compared with baselines, COAT has lowest RMSE loss. Compared with human-expert results, we find the factor quality of COAT is also competitive.

Compare COAT-MB and COAT-ADJ We empirically demonstrate the differences between the two algorithms in a variant of AppleGastronome data set which has a different causal structure as shown in Figure 15(a). The task is to estimate the causal effect from `taste` to `score` with their numerical values as well as the unstructured textual data.

After applying COAT-MB on `taste`, the causal graph on the extended node set is shown in Figure 15(b). This graph does not meet the graphical requirement on visibility because the edge taste → score is *invisible*. The red edge in the graph doesn't have an arrow head into `taste`. On the other hand, the edge is *visible* (see Definition 10) due to the green edge in Figure 15(c) found by COAT-ADJ, and one can proceed to search adjustment sets based on graphical criteria (Maathuis and Colombo, 2015; Perković et al., 2018), like {`flavor/aroma quality`, `market potential`}.

5.5 Ablation Studies

Group Size in Prompt In the COAT prompt, several samples are given and grouped by the values of the target variable. The samples in each group are randomly selected to a fixed number (like 3 samples per group). Empirically, we keep it to be 3 throughout all



(a) Ground Truth

(b) COAT-MB

(c) COAT-ADJ

Figure 15: The comparison between COAT-MB and COAT-ADJ in the causal inference setting.

Method	Cluster Size	Group Size	MB	NMB	OT	Recall	Precision	F1
META	-	-	2.67±*0.94*	0.67±*0.47*	2.33±*0.47*	0.53±*0.19*	0.46±*0.08*	0.49±*0.13*
DATA	-	3	3.00±*0.00*	0.33±*0.47*	0.00±*0.00*	0.60±*0.00*	0.92±*0.12*	0.72±*0.04*
COAT	`len(factor)+1`	3	4.00±*0.82*	0.33±*0.47*	0.00±*0.00*	0.80±*0.16*	0.93±*0.09*	0.85±*0.11*
COAT	`len(factor)+1`	1	4.67±*0.58*	0.00±*0.00*	0.00±*0.00*	0.93±*0.12*	1.00±*0.00*	0.96±*0.06*
COAT	2	3	3.67±*1.53*	0.00±*0.00*	0.00±*0.00*	0.73±*0.31*	1.00±*0.00*	0.82±*0.22*

Table 10: Results about Ablation Study on Hyperparameters.

experiments (sometimes smaller than 3 if samples are not enough). In practice, it is mainly constrained by the LLM's context length.

The Number of Clusters When constructing feedback, we first use clustering to separate the data set and then find the cluster where the target variable is not explained well by current factors. Empirically, we set the number of clusters to be one plus the number of current factors.

We conduct ablation studies of COAT with GPT-4 using different hyperparameters. As shown in Table 10, one can observe that COAT is not sensitive to these hyperparameters and performs more robustly than the baselines under different hyperparameter setups.

Cross-Model Validation of Synthetic Data. The AppleGastronome data set is constructed through a two-phase process: (1) the tabular data (i.e., apple attributes and scores) are generated via Bernoulli distributions according to a human-specified causal structure; (2) GPT-4 is prompted to generate textual reviews based on each sample's attributes and score. In Table 3, to mitigate potential correlated errors between the LLM used for text generation and those used for factor proposal, we report results across 10 distinct LLMs for the proposal stage.

Review Generation	Factor Proposal	MB	NMB	OT	Recall	Precision	F1
DeepSeek-V3-0324	DeepSeek-V3-0324	3.75±*0.66*	0.00±*0.00*	0.44±*0.51*	0.75±*0.13*	0.91±*0.10*	0.82±*0.12*
	GPT-OSS-120B	5.00±*0.00*	0.00±*0.00*	0.00±*0.00*	1.00±*0.00*	1.00±*0.00*	1.00±*0.00*
	Qwen3-Next-80B-A3B	2.94±*0.10*	0.33±*0.58*	0.00±*0.00*	0.59±*0.02*	0.92±*0.14*	0.71±*0.04*
GPT-OSS-120B	DeepSeek-V3-0324	4.08±*1.18*	0.33±*0.58*	0.00±*0.00*	0.82±*0.24*	0.91±*0.15*	0.86±*0.20*
	GPT-OSS-120B	5.00±*0.00*	0.00±*0.00*	0.00±*0.00*	1.00±*0.00*	1.00±*0.00*	1.00±*0.00*
	Qwen3-Next-80B-A3B	4.33±*1.15*	0.00±*0.00*	0.00±*0.00*	0.87±*0.23*	1.00±*0.00*	0.92±*0.14*
Qwen3-Next-80B-A3B	DeepSeek-V3-0324	4.33±*0.47*	0.00±*0.00*	0.00±*0.00*	0.87±*0.09*	1.00±*0.00*	0.93±*0.05*
	GPT-OSS-120B	4.72±*0.25*	0.00±*0.00*	0.00±*0.00*	0.94±*0.05*	1.00±*0.00*	0.97±*0.03*
	Qwen3-Next-80B-A3B	3.25±*0.43*	0.00±*0.00*	0.00±*0.00*	0.65±*0.09*	1.00±*0.00*	0.79±*0.06*

Table 11: Cross-Model Validation of AppleGastronome Construction.

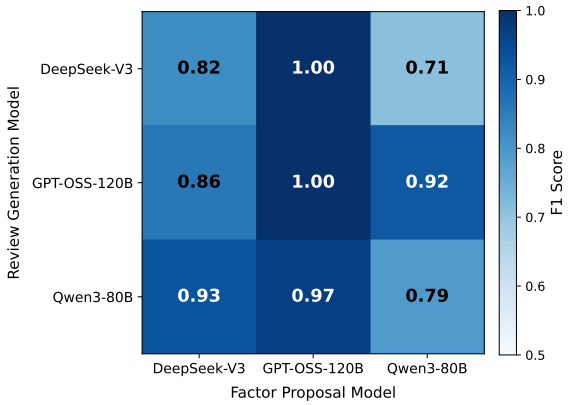


Figure 16: Heatmap for the cross-model validation of synthetic data.

To further validate the robustness of our evidence, we conduct a cross-model validation study: we generate three additional variants of the textual reviews using different LLMs (`DeepSeek-V3-0324`, `GPT-OSS-120B`, and `Qwen3-Next-80B-A3B`) based on the same underlying tabular data, and then evaluate each LLM as a factor proposer on all three review data sets.

As shown in Table 11, the factor proposal performance remains robust across different combinations of generation and proposal models. Notably, the precision scores are consistently high across nearly all configurations, often reaching 1.00. In terms of recall and F1, `GPT-OSS-120B` consistently achieves the highest recall and near-perfect F1 scores (0.97–1.00) as a proposer across all three review variants, and achieves a perfect F1 score on self-generated reviews. Meanwhile, `DeepSeek-V3-0324` and `Qwen3-Next-80B-A3B` achieve second and third F1 scores on their respective self-generated reviews. This indicates that

Factor	Type	Factor Definition
Juiciness	Correlated	**1**: Abundant or moist texture described in text **0**: Not mentioned or ambiguous description **-1**: Dry or lacking moisture
Color	Irrelevant	**1**: Vibrant or appealing color mentioned **0**: Not mentioned or neutral description **-1**: Uneven, dull, or unappealing color

Table 12: Definitions of injected factors for robustness stress testing. *Juiciness* serves as a disturbing candidate (correlated with outcome but not in the true Markov Blanket), while *Color* represents a completely irrelevant factor.

Injection	Iteration	MB	NMB/OT	Injected Factors	F1
juiciness	1	0.00±*0.00*	1.00±*0.00*	1.00±*0.00*	0.00±*0.00*
	2	3.20±*0.98*	0.20±*0.40*	0.00±*0.00*	0.75±*0.16*
	3	3.60±*0.80*	0.00±*0.00*	0.00±*0.00*	0.83±*0.10*
	4	4.00±*0.89*	0.00±*0.00*	0.00±*0.00*	0.88±*0.11*
color	1	0.00±*0.00*	0.00±*0.00*	0.00±*0.00*	0.00±*0.00*
	2	2.60±*0.49*	0.60±*0.80*	0.00±*0.00*	0.64±*0.14*
	3	3.40±*0.80*	0.40±*0.49*	0.00±*0.00*	0.77±*0.12*
	4	4.00±*0.63*	0.40±*0.49*	0.00±*0.00*	0.85±*0.09*

Table 13: Stress test result with non-MB factor injected in the initial set. For the *juiciness* factor, it was preserved in the first round, and then was removed after 1 iteration. For the *color* factor, since it is not relevant to the textual data, it was rejected in the first round by statistical testing.

the performance is largely driven by the factor-proposer LLM. These results suggest that the AppleGastronome data set is robust against model-specific biases.

COAT-MB with Unexpected Factor Injection To assess whether COAT-MB can recover from wrong factor proposals, we conduct a stress test by manually injecting "unexpected factors" into the initial round: (1) *juiciness*—a factor correlated with the target but not part of the true Markov Blanket, and (2) *color*—a completely irrelevant factor with no statistical relationship to the data. To conduct the injection, we put their factor definitions in Table 12 as the factor proposal of the first round in COAT-MB.

As shown in Table 13, when *juiciness* is injected, the algorithm initially preserves it in round 1 (Injected Factors = 1.00). Critically, as the COAT-MB continues to propose and verify new factors, it starts to identify the true MB factors (MB size increases to 3.20), and this wrong factor is quickly discarded by round 2 (Injected Factors drops to 0.00). For the completely irrelevant *color* factor, it is dropped immediately in round 1 (Injected Factors = 0.00) due to the independence between its factor values and the target's ones. These



(a) Two noises (b) Three noises

Figure 17: Stress test result with a variable set of pure noises. The initial variables U_1, U_2, U_3 are not relevant to the textual data, and therefore COAT-PAG returns no additional factors.

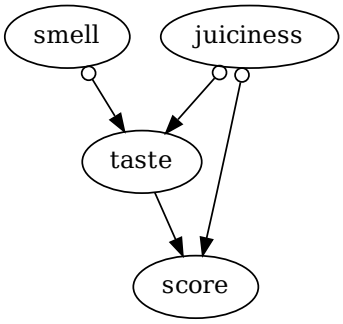
results demonstrate that COAT-MB is robust to possibly wrong factor proposals with the help of the iterative feedback loops.

COAT-PAG with Pure Noise Variables. To evaluate whether COAT-PAG can recover when the initial variable set $\mathcal{U}$ contains only noise variables with no relationship to the textual data, we conduct a stress test with two synthetic noise structures: (1) *two correlated noises* ($U_1 \to U_2$), where U_2 is constructed as a linear function of U_1 with Gaussian noise; and (2) *three noises with collider structure* ($U_1 \to U_2 \leftarrow U_3$), where U_2 is a common effect of two independent noise sources U_1 and U_3. For each setting, we apply COAT-PAG on $\mathcal{U}$ and the five textual data sets in Figure 11 respectively.

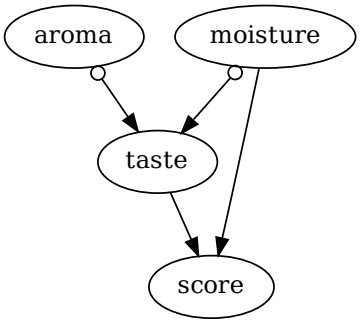
As illustrated in Figure 17, when COAT-PAG is initialized with these pure noise variables over the textual data set, the algorithm correctly rejects all the factors proposed from the text $\boldsymbol{X}$ based on independence testing. The resulting graphs, which are consistent across five data sets, preserve only the initial variable set $\mathcal{U}$ and the causal structures among the noise variables themselves. This demonstrates that COAT-PAG can abstain from hallucinating factors in the extreme case.

COAT-ADJ with Unobserved Confounders To evaluate how COAT-ADJ behaves when some confounding factors cannot be expressed as a function of unstructured data $\boldsymbol{X}$, we conduct a stress test comparing two scenarios: (1) *fully observable case* where all confounders are extractable from $\boldsymbol{X}$ (Figure 18 (a–b)), and (2) *partially observable case* where a latent confounder (*juiciness*, shown as dashed oval) affects both T and Y but is not captured in $\boldsymbol{X}$ (Figure 18 (c–d)). We apply COAT-ADJ to both data sets with initial graph $T \to Y$ (*taste* $\to$ *score*).

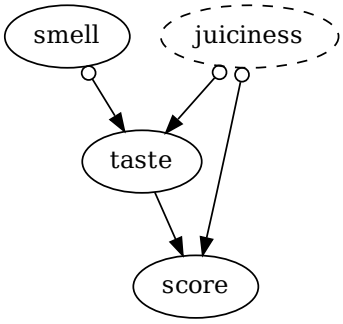
In the fully observable case, the algorithm correctly identifies the edge $T \to Y$ as *visible* (see Definition 10) because there exists a node (*aroma*) that is not adjacent to Y yet has a directed edge into T. Consequently, COAT-ADJ would proceed to find valid adjustment sets (in this case, it can be {moisture}). In the partially observable case, the edge $T \to Y$ is rendered *invisible* due to the unobserved *juiciness*. In such situation, COAT-ADJ would



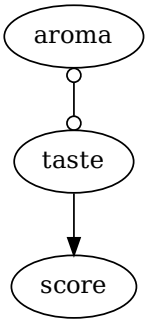
(a) fully observable case - ground truth



(b) fully observable case - result



(c) partially observable case - ground truth



(d) partially observable case - result

Figure 18: Stress test result with latent confounder for the treatment effect variables. For (a) and (b), there is no latent variable, and the result from COAT-ADJ indicates the edge $T \to Y$ is *visible* since there exists another node (aroma) that is not adjacent to Y and has an edge into T. For (c) and (d), the edge $T \to Y$ is invisible due to the latent confounder (juiciness).

report the invisibility and abstain from causal estimation. These results demonstrate that COAT-ADJ behaves conservatively when faced with possible confounders that cannot be elicited from the $\boldsymbol{X}$.

6. Real-World Case Studies

In this section, we use COAT framework (COAT-MB Algorithm) to analyze unstructured data from three real-world scenarios: (1) brain tumor detection with MRI images (Bhuvaji et al., 2024), where COAT-MB elicits visual factors; (2) climatic reanalysis data with fine-grained time and space coverage (Compo et al., 2011), where LLM can write code to interact with a database. (3) three-year news summary about one stock from the New York Times (Innovations, 2023), where COAT is applied to analyze the sequential data.

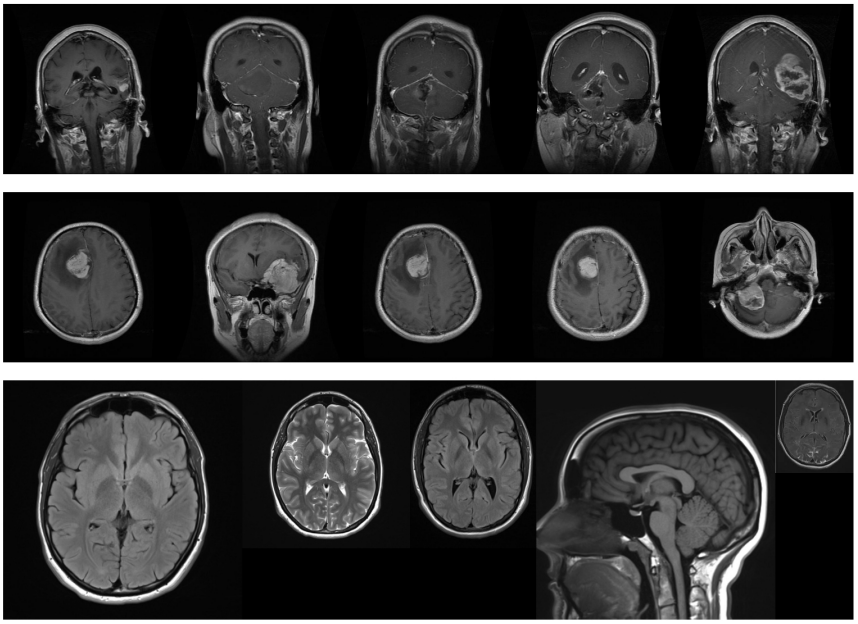


Figure 19: An example input from Brain Tumor data. Each row contains random samples from the same category (top-down: glioma, meningioma, and no tumor).

6.1 Eliciting Tumor-Related Factors from Brain MRI Images

In this section, we utilize the COAT to explore the image data set. Magnetic Resonance Imaging (MRI) is an important technique for detecting tumors in the human brain. The images are from an open Kaggle data set (kaggle/brain-tumor-classification-mri) with an open-sourced project (Bhuvaji et al., 2024). Each sample is a scanning MRI of a human brain. In this case study, the interesting variable is the tumor type. We consider three types of MRI images: glioma, meningioma, and no tumor. We include 20 images for each category and the total sample size is 60.

Data Processing We use gpt-4-vision-preview to handle image samples. As the current GPT-4 cannot process multiple images simultaneously, we concatenate samples from different categories into one picture, as shown in Figure 19, and provide additional instructions in the prompt to explain the format. For each proposed factor, the LLMs will go through all 60 sample images individually to evaluate the factor value.

Result As shown in Figure 21, there are two visual factors that appeared in the final causal graph: *contrast enhancement* and *mass effect*. One may refer to Figure 20 to see detailed descriptions of these two factors. We verify the proposed factors by searching medical literature according to keywords in factor descriptions. We found that both factors are aligned with existing research (Lyndon et al., 2019; Haydar et al., 2022; Tunç et al., 2021; Watts et al., 2014). Therefore, we believe COAT framework can provide a reasonable starting point for domain experts.

Contrast Enhancement

- 1: Strong, uniform enhancement typical of a meningioma.
- 0: No abnormal enhancement or no mass present.
- -1: Heterogeneous enhancement with or without areas of necrosis typical of a glioma.

Mass Effect

- 1: Mass causing significant displacement of normal brain structures (can occur with both meningioma and glioma but is more pronounced in glioma due to intra-axial location).
- 0: No mass is present, or the specific criteria for 1 or -1 are not met.
- -1: Mass causing minimal or no displacement of brain structures (more common in meningioma due to extra-axial location).

Figure 20: Detailed descriptions of these two final factors in the Brain Tumor case study.

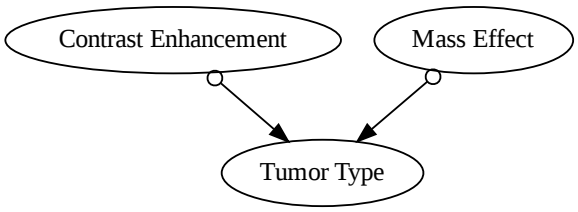


Figure 21: Final causal graph by COAT-MB in the Brain Tumor case study.

6.2 Eliciting Climate Factors with LLMs Using Tools

The NOAA 20th Century Reanalysis V3 data set (Compo et al., 2011; Slivinski et al., 2019) contains high-dimensional information about Earth's atmosphere with a fine-grained time-and-spatial coverage. We use COAT to explore El Niño-Southern Oscillation (ENSO) phenomenon: irregular fluctuations in sea surface temperatures (McPhaden, 1999) in the Pacific Ocean. The focus of this study is the future change in monthly SST in the Nino3 region, which could be an important indicator of ENSO events.

LLMs using tools The data set is in the NetCDF format (network Common Data Format), so it is not convenient to be directly fed to LLMs. Therefore, the LLM is required to write code to define specific factors. An example is provided in Lst. 1. The *measurement* specifies the climate variables like precipitation rate or temperature; the *level* specifies the vertical location, like surface or specific pressure level; the region is a rectangle about the area concerned. This function will output a single time series about the measurement on the specified level averaged within the specified region.

Feedback construction As shown in Figure 22, two types of information are included in the feedback: *intermediate causal graph* and *OLS regression result* of factors in the current estimation of a Markov blanket on the target variable. These two types of information can be drawn naturally from the COAT's intermediate results.

```
    def get_air_temperature_at_cloud_layer_top_nino3():
        '''
        Air Temperature at Cloud Layer Top (Nino3 Region)
        '''
        # Declare observation
        observation_name = Observation(
            measurement="Air Temperature",
            level="Isentropic Levels",
            region={'latitude_min': -5, 'latitude_max': 5, '
    longitude_min': 210, 'longitude_max': 270},
            detailed_level=300
        )

        # Return the observation values directly
        return observation_name

    factor_dict['Air Temperature at Cloud Layer Top (Nino3 Region)']
    = get_air_temperature_at_cloud_layer_top_nino3()
```

Listing 1: LLM can be prompted to propose factors using complex tools.

Result interpretation In this case study, we are treating non-stationary climate data, while assuming the causal structure is invariant. Under this assumption, the non-stationarity is actually helpful for causal structure learning (Huang et al., 2020; Mooij et al., 2020). To this end, we utilize the CD-NOD algorithm (Huang et al., 2020) to fully utilize the changing causal modules for better identifiability. CD-NOD will first identify the non-stationary nodes, whose conditional distribution given the causal parents is changing with time, and then use them for better causal structure recovery. Four factors are identified to be non-stationary, as shown in Figure 23.

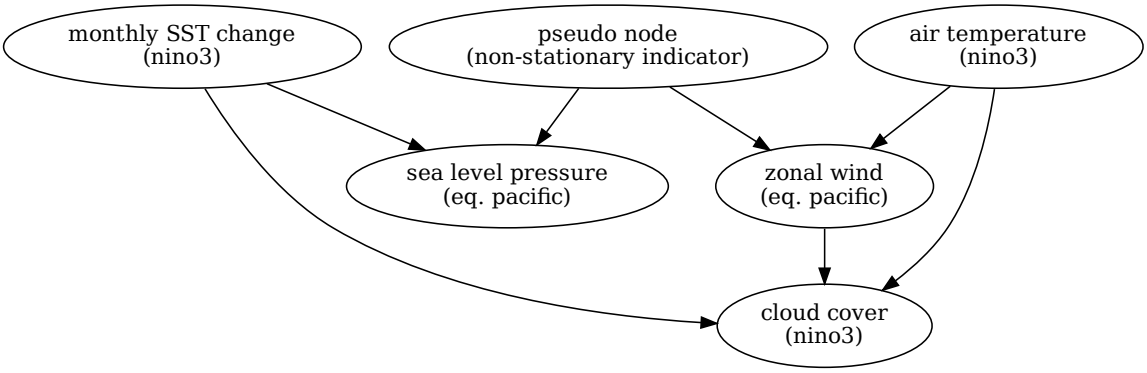
We have adjusted the nodes' names, shapes, and colors for better visualization. There are 14 nodes in total, with 13 factors identified by COAT. Each factor is a time series about a certain climate measurement above a specific level averaged over a specific region. For simplicity, we only considered instantaneous causal relations among those time series.

There are three regions identified to be relevant to the ENSO phenomenon:

- **Equatorial Pacific Region (Orange Nodes).** This region (5N-5S, 120W-280W) is one of the most active places about ENSO. It becomes significantly warm during the El Niño phase and becomes significantly cold during the La Niña phase (Zhang et al., 2017).

- **Nino3 Region (Blue Nodes).** This region (5N-5S, 150W-90W) is one of the most classical regions to monitor the El Niño events by scientists. It is also used by humans to construct the target variables (McPhaden, 1999).

- **South American Coastal Region (Green Nodes).** This region (0N-20S, 80W-60W) includes the Peruvian coastal up-welling system, and is noticed to be relevant to the ENSO cycle (Tarazona and Arntz, 2001).



Follow the instructions to complete the task.
 # Result
Causal Graph Results *[Causal graph is visualized in the figure aside]*

Regression Results on a Markov Blanket *[The table is simplified for brevity]*

Variable	**coef**	**std err**	**t**	**P>|t|**
sea surface air temp gradient (eq. pacific)	-0.0003	0.001	-0.60	0.548
ocean-atm heat flux (cen. eq. pac.)	0.0023	0.001	2.81	0.005
lag ocean-atm heat flux (cen. eq. pac.)	0.0094	0.001	12.16	0.000
lag albedo variability (e. eq. pac.)	-0.9733	0.034	-28.41	0.000
lag next-month shift nino3 SST	0.6513	0.014	45.32	0.000

R^2(uncentered) = 0.695, F = 951.9, N = 2513

 # Task
What are the new factors that might influence 'next month's shift for nino3 SST (Sea Surface Temperature)'?
1. Analyze the results (Causal Graph result & Regression Results) step by step.
2. Guess potential new factors based on what you have known.

Figure 22: The prompt with feedback based on previous loop in the ENSO case study.

Some insights are delivered by paths in the output causal graph:

- **Sea level Pressure.** This factor is about the sea-level pressure on the equatorial Pacific region. The pressure gradient will influence the movement of warm water, and thus influence the sea surface temperature (SST) change (Bjerknes, 1969). In addition, pressure can influence the water evaporation and thus regulate through the water circulation. This also matches another indirect path Sea level Pressure → Sensible Heat Net Flux → Volumetric Soil Moisture → Cloud Cover → SST Change.

- **Momentum Flux, V-component.** This factor is about the vertical movement of air. It is crucial in driving atmospheric convection (Bjerknes, 1969), and it is related to the Walker Circulation, which is an important component in the ENSO system(Wang, 2004). Also, it could influence the change in sea level pressure and indirectly influence the SST change.

- **Cloud Cover.** The factor is the fraction of the sky covered by clouds in the NINO3 region. It could influence the SST Change through solar radiation as well as water circulation. It is confirmed to have a significant correlation (Liu et al., 2016) with ENSO events and plays an important role in the atmospheric circulation and hydrological cycle (Mishra, 2019).

- **Soil Temperature.** This might be a novel hypothesis proposed by COAT, since we found no sufficient research to confirm this point to the best of our knowledge. This causal graph has also suggested two possible indirect mechanisms: (1) through *Volumetric Soil Moisture* and *Cloud Cover*; and (2) through *Convective Precipitation Rate* and *Sea Level Pressure*. Therefore, this finding encourages more serious investigations of these hypotheses.

6.3 Eliciting Financial Factors with Sequential News Data

In this section, we utilize COAT to explore time series data with text. The data set consists of the stock value of Microsoft (MSFT) from 2006 to 2009 and its news summary (only include news in *the New York Times*). This is a subset of an open Kaggle data set (kaggle/stock-price-and-news-realted-to-it). Each sample is one trading day with the company's close stock price and news. The target value is the future return rate, and we are curious about factors in the related news. We fed data during the first 200 trading days to COAT, and we used the following 400 trading days for evaluation.

The target value is a binary variable: 1 for positive return rate in next week, 0 otherwise. News fed to COAT are grouped according to the value of Y_t. To keep the whole prompt within the context limit, only 3 samples will be randomly included in each group.

Factor processing Given one proposed factor, denote its annotated value at the t-th trading day as $a_t \in \{-1, 0, 1\}$. Each factor is averaged over a rolling window of the past M days:

$$S_t = \frac{1}{M} \sum_{t-M+1 \le i \le t} a_i. \tag{15}$$

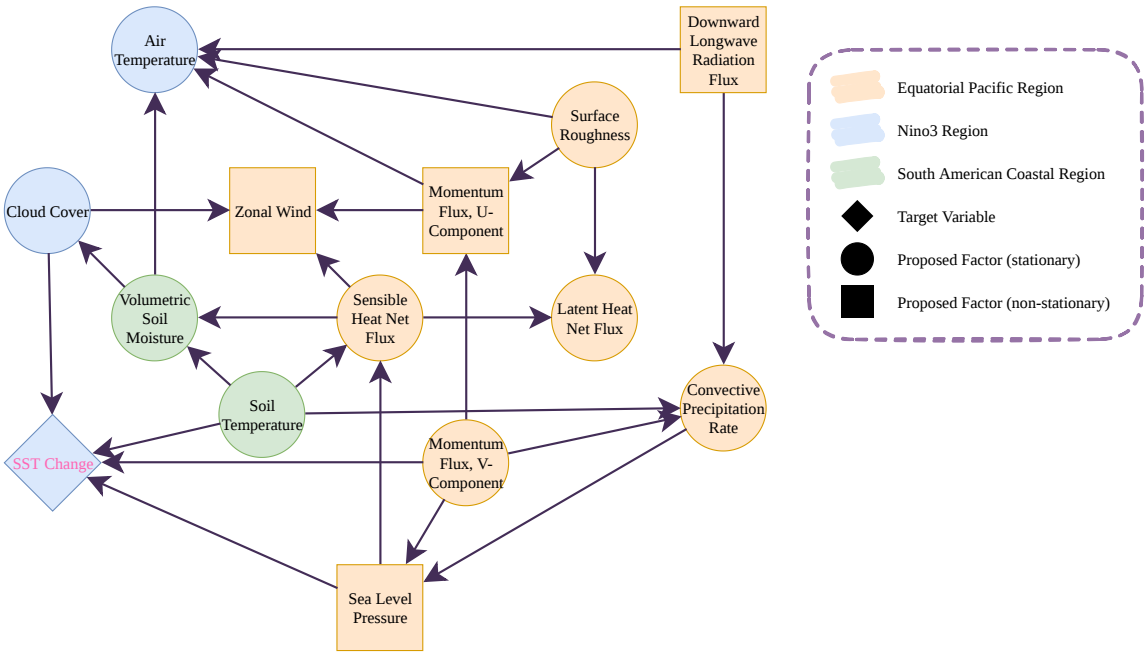


Figure 23: The final causal graph found by COAT-MB in the ENSO case study.

Each variable, including return rate, is normalized by the rolling standard deviation over the past M days:

$$F_t = \frac{1}{\mathrm{Sd}\left(\{S_i\}_{t-M+1\le i\le t}\right)+1} S_t. \tag{16}$$

An example of a processed factor can be seen in Figure 24.

Resulting causal graph As shown in Figure 25, four factors are identified in the final causal graph to form a Markov blanket of return rate: *Product Focus*, *Legal and Regulatory Issues*, *Market Strategy*, and *Innovation and Technology Focus*. One can refer to Figure 26 for more detailed descriptions of these factors. One factor (*Innovation and technology focus*) is identified as a possible cause of the return rate; this matches the nature of the company's type and reflects people's expectation of the company to keep creating innovative computer software. It is also interesting to see that COAT captures the structure between factors and *market strategy*, where *product focus* and *legal and regulatory issues* are identified as potential causes. It also implies the existence of a latent confounder between *market strategy* and *return rate* that may not be significantly reflected in news text.

Evaluation by trading strategies For each factor $\{F_t\}$ processed in Equation 16, we establish a trading strategy on it to see the performance in the out-of-sample trading days. At each trading day t, we fit the model $r_t = \alpha + \beta F_t$, using samples $\{r_i, F_i\}_{t-200\le i<t}$. After the estimation, we make decision based on $\hat{r}_t := \hat{\alpha} + \hat{\beta} F_t$. If $\hat{r}_t > 0$, we *go long* with 1 unit of capital, i.e., purchase that amount of stock, the gain will be $r \times 1$ units of capital. If $\hat{r}_t < 0$, we *go short* with 1 unit of capital, i.e., borrow and sell that amount of stocks

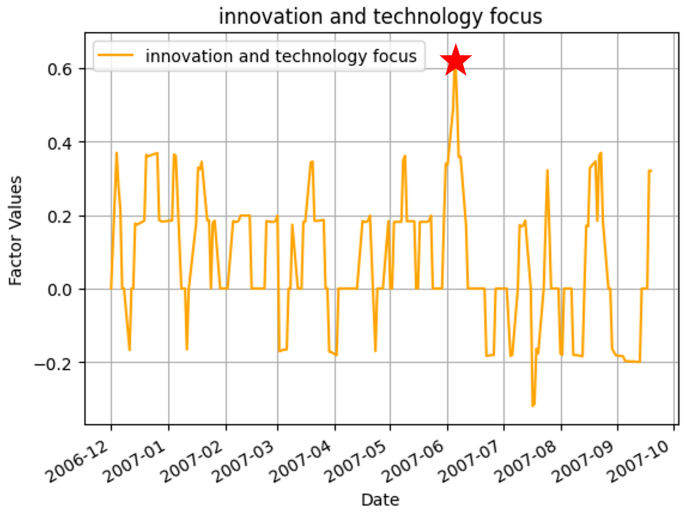


Figure 24: Example of one processed factor *Innovation and Technology Focus* during the first 200 trading days. The red star marks the highest value. In June 2007, there was a discussion about future competition in desktop operating systems and the trend towards web-based services Markoff (2007).

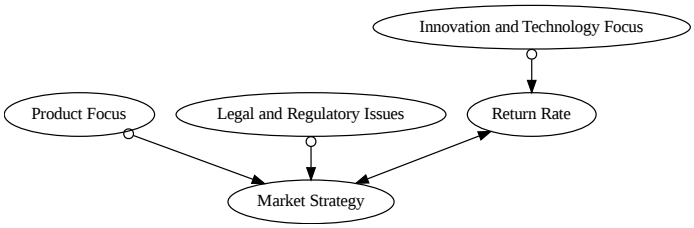


Figure 25: The final causal graph by COAT-MB in the Stock News case study.

Product Focus

- - 1: Text primarily discusses the launch, performance, or features of a company's product.
- - 0: Otherwise; or not mentioned.
- - -1: Text does not focus on products but rather on other aspects such as company strategy, business practices, or financial results.

Legal and Regulatory Issues

- - 1: Text is centered around legal proceedings, regulatory actions, or oversight related to a company's business practices.
- - 0: Otherwise; or not mentioned.
- - -1: Text does not discuss legal or regulatory issues but focuses on other topics like product information or market performance.

Market Strategy

- - 1: Text discusses company strategies related to market expansion or competitive actions.
- - 0: Otherwise; or not mentioned.
- - -1: Text focuses on internal company issues, product details, or legal/regulatory matters without a direct link to market strategy.

Innovation and Technology Focus

- - 1: Articles highlighting new technologies, innovations, or significant advancements in existing technology.
- - 0: Otherwise; or not mentioned.
- - -1: Articles discussing established technologies, standard updates, or non-innovative topics.

Figure 26: The descriptions of the proposed factors in the Stock News data.

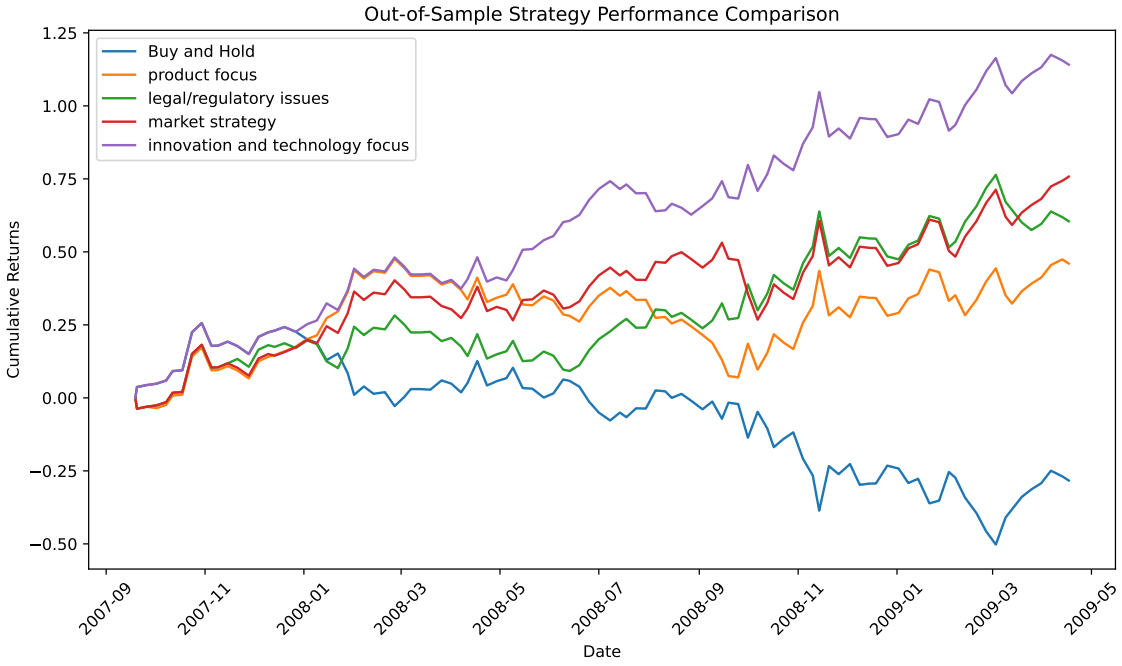


Figure 27: The out-of-sample cumulative return of trading strategies based on different factors in the Stock News case study.

immediately and buy them back next time, the gain will be $-r \times 1$ units of capital. We introduce an additional *Buy and Hold* baseline to always go long one unit of capital. The trading evaluation is after the 200-th day and thus is not accessible by COAT.

The cumulative return plot is shown in Figure 27, and the metrics commonly used in economic literature (Fama, 1970; Wachter, 2013; Bybee et al., 2024) are shown in Table 14. One important metric to see a trading strategy's effectiveness is the Sharpe ratio: a measure of risk-adjusted return, showing the excess return (over the risk-free rate, we set it to be 2%) per unit of standard deviation. A higher Sharpe ratio indicates better performance per unit of risk. We see that the *Innovation and Technology Focus* factor yields the highest Sharpe ratio and outperforms other non-causal factors.

7. Discussion and Conclusions

In this work, we reformulate and extend our previous work (Liu et al., 2024) that is only on the Markov blanket setting, and extend the COAT framework for *partial ancestral graphs* (COAT-PAG) and *adjustment sets* (COAT-ADJ). In general, COAT aims to leverage the rich knowledge learned by LLMs to elicit causal representation from unstructured data, like images or text. The key module of COAT is a propose-and-verify procedure where the advantages of LLMs are employed on understanding unstructured data and actively proposing possible candidate high-level factors, and these factors will be further verified by causal discovery methods. First, we propose an iterative method that can elicit Markov Blanket for a given target variable from its paired unstructured data. Based on this method,

	Buy and Hold	product focus	legal/regulatory issues	market strategy	innovation and technology focus
Expected Return	-3.50	0.29	0.38	0.48	0.72
Sharpe Ratio	-2.38	0.70	0.94	1.20	1.86
t-Stat	-3.00	0.89	1.19	1.51	2.35
Information Ratio	-	2.45	2.50	2.56	2.73
α	-	3.79	3.88	3.98	4.22
α t-Stat	-	3.09	3.15	3.23	3.44
Max Loss	-0.50	-0.15	-0.15	-0.15	-0.15
Skew	-0.43	-0.14	-0.15	-0.41	-0.38

Table 14: Performance of trading strategy according to each factor.

we further present an approach to extending a given Partial Ancestral Graph to maximally refine its skeleton and arrowhead orientation. Furthermore, we extend our framework to find adjustment sets for cause-effect estimation.

For evaluation, we construct AppleGastronome benchmark in the two critical settings: (1) Eliciting Markov Blanket; and (2) Extending Partial Ancestral Graph. By comparing with LLMs' reasoning approach, COAT presents robust performance when the reality deviates from LLMs' priors, and is a reliable framework to integrate the advantages of LLMs into causality tasks. We define two metrics: perception score and capacity score, to measure the LLMs' causal ability under the COAT framework; the empirical results over ten predominant LLMs support their competence. We show COAT is useful and reliable in semi-realistic benchmarks and real-world case studies. In Neuropathic benchmark, COAT can elicit factors in radiculopathy and pathophysiology with only symptom descriptions, where LLMs utilize their intrinsic domain knowledge with verification from causal discovery methods. In ATE estimation data sets, COAT presents lowest RMSE loss compared with baselines, and is competitive with results from human experts. In real-world case studies, COAT elicits two visual factors related to tumor type by analyzing tumor images data; COAT can also propose factors by writing codes to interact with climate database. These results are consistent with domain literature. In the financial case study, the proposed factors present high Sharpe ratios in out-of-distribution trading simulations.

The intensive empirical results, as well as the identifiability results of our proposed methods, demonstrate the insight that LLMs and causal discovery methods are mutually beneficial. On the one hand, LLMs that learn rich world knowledge about the world can assist with eliciting high-level hidden variables from low-level observational data. On the other hand, causal discovery methods can provide verification and feedback to LLMs, improving the robustness when LLMs' priors don't fit the specific data.

There are several open questions that can be further investigated in the future. First, in this paper we consider the observational data, which can be extended to the setting where LLMs are able to choose and conduct interventions on certain target. This will bring more information on the causal mechanism. Second, in this paper we mainly focus on the constraint-based methods in causal discovery; one can explore how to fully utilize other methods like ICA-based approach, and also causal representation learning methods. Third,

the COAT framework can also integrate experimental design methods in new perspectives: (1) by expanding the subgraph around an intervention target, COAT can resolve undetermined edges using unstructured data, reducing the need for redundant experiments; and (2) with the expanded node set and the refined causal structure, COAT delivers richer input to methods for experimental design, and yields more valuable experimental data sets.

Acknowledgments

CXL and BH were supported by NSFC Major Research Plan No. 92570109, RGC Young Collaborative Research Grant No. C2005-24Y, RGC Research Fellow Scheme No. RFS2627-2S02, RGC General Research Fund No. 12202026, RGC General Research Fund No. 12200725, NSFC General Program No. 62376235, Guangdong and Hong Kong Universities "1+1+1" Joint Research Collaboration Scheme, HKBU Faculty Niche Research Areas No. RC-FNRA-IG/25-26/SCI/05, and HKBU CSD Departmental Incentive Scheme. TLL was partially supported by the following Australian Research Council projects: DP260102466, FT220100318, and DP220102121. MG was supported by ARC DP240102088 and WIS–MBZUAI 14257. JC was supported by RGC Young Collaborative Research Grant No. C2005-24Y. KZ would like to acknowledge the support from NSF Award No. 2229881, AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from MBZUAI-WIS Joint Program, and the Al Deira Causal Education project. Support for the NOAA-CIRES-DOE Twentieth Century Reanalysis Project version 3 dataset is provided by the U.S. Department of Energy, Office of Science Biological and Environmental Research (BER), by the National Oceanic and Atmospheric Administration Climate Program Office, and by the NOAA Physical Sciences Laboratory. The authors declare no competing interests.

Appendix of COAT

Contents

A Table of Notations **56**

B Proofs for Theoretical Results **57**

C More Details about Experiments **61**
C.1 More Details on Constructing AppleGastronome 61
C.2 More Details on Prompts for AppleGastronome 61
C.3 More Details of Results on AppleGastronome 62
C.4 Implementation of the FCI Algorithm 62
C.5 More Details on Constructing Neuropathic 73
C.6 More Details of Results on Neuropathic 74
C.7 Discussion on the Time Complexity 74
C.8 Resources 75

D COAT with Different Causal Discovery Algorithms **80**

E Summary of Benchmark Data **82**

F COAT Output for Users in Practice **83**

Appendix A. Table of Notations

Notation	Description
Ψ	The LLM for factor proposal.
$\boldsymbol{p}$ ($\boldsymbol{p}^t$)	Prompt used (at the t-th COAT round) for factor proposal.
Ψ_s	The LLM for factor parsing.
$\boldsymbol{p}_p$	Prompt used for factor parsing.
$\mathcal{A}$	Algorithm for causal discovery. In this paper we mainly use FCI.
$\boldsymbol{X}$	Unstructured data; treated as a high-dimensional random vector.
$\mathcal{U}$	The set of given structured random variables; and $\mathcal{U} = \{U_1, U_2, \cdots, U_m\}$
$\boldsymbol{U}$	The random vector consists of variables in $\mathcal{U}$, i.e., $\boldsymbol{U} = [U_1, U_2, \cdots, U_m]^\top$
Y	The target variable from $\mathcal{U}$; or the effect variable in the causal inference setting.
$\mathcal{D}$	Full sample set $\mathcal{D} = \{\boldsymbol{x}^{(j)}, \boldsymbol{u}^j\}_{j=1}^n$. $(\boldsymbol{x}^{(j)}, \boldsymbol{u}^{(j)})$ is a realization of $(\boldsymbol{X}, \boldsymbol{U})$.
$\boldsymbol{w}_i$	The i-th factor proposed by LLMs, like *sweetness*, *size*, or *scent*
$\mathcal{W}$	The set of proposed factors from $\boldsymbol{X}$; and $\mathcal{W} = \{W_1, W_2, \cdots, W_k\}$, where $W_i \triangleq \boldsymbol{w}_i(\boldsymbol{X})$
$\boldsymbol{W}$	The random vector consists of variables in $\mathcal{W}$, i.e., $\boldsymbol{W} = [W_1, W_2, \cdots, W_k]^\top$
$h_{\mathcal{S}}(\cdot)$	LLM-induced representation over index $\mathcal{S}$, e.g., $h_{\{1,3\}}(\cdot) = (\boldsymbol{w}_1(\cdot), \boldsymbol{w}_3(\cdot))$. $h(\cdot) \triangleq h_{[k]}(\cdot)$
$\mathcal{C}$	The predefined value space for LLM-proposed factors, e.g., set $\mathcal{C}$ as $\{-1, 0, 1\}$
$\mathcal{V}$	The augmented structured variable set by COAT, i.e., $\mathcal{V} \triangleq \mathcal{W} \cup \mathcal{U}$.
$\boldsymbol{V}$	The random vector consists of variables in $\mathcal{V}$, i.e., $\boldsymbol{V} = [U_1, U_2, \cdots, U_m, W_1, W_2, \cdots, W_k]^\top$
$\boldsymbol{v}^{(j)}$	The realization of random vector $\mathcal{V}$ in the j-th sample.
$\mathcal{O}$	The set of all possible factors that can be defined based on $\boldsymbol{X}$.
$\mathcal{E}$	The maximal observable variable set $\mathcal{E} \triangleq \mathcal{O} \cup \mathcal{U}$.
T	The treatment variable from $\mathcal{U}$ in the causal inference setting.
$\mathcal{P}_{\mathcal{V}}$	The partial ancestral graph over node set $\mathcal{V}$; similar for $\mathcal{P}_{\mathcal{E}}$
$\widehat{\mathcal{D}}^t$	The small subset of sample used at the t-th COAT round.
$\mathcal{W}^t$	The set of factors proposed by LLMs at the t-th round. $\mathcal{W}^{\leq t} \triangleq \mathcal{W}^1 \cup \cdots \cup \mathcal{W}^t$
C_Ψ	Capacity Score of LLM Ψ
p_Ψ or p	Perception Score of LLM Ψ

Appendix B. Proofs for Theoretical Results

Proof [Proof for Proposition 2] From Eq. 8, we have

$$H(Y \mid h_{[k]}(\boldsymbol{X}); \boldsymbol{w}_{k+1}(\boldsymbol{X})) < H(Y \mid h_{[k]}(\boldsymbol{X})); \tag{17}$$

Since $Y \perp\!\!\!\perp h_{[k+1]\setminus\mathcal{S}}(X) \mid h_{\mathcal{S}}(X)$, we have

$$H(Y \mid h_{\mathcal{S}}(\boldsymbol{X}); h_{[k+1]\setminus\mathcal{S}}(\boldsymbol{X})) = H(Y \mid h_{\mathcal{S}}(\boldsymbol{X})). \tag{18}$$

Therefore:

$$\begin{aligned} H(Y \mid h_{\mathcal{S}}(\boldsymbol{X})) &= H(Y \mid h_{\mathcal{S}}(\boldsymbol{X}), h_{[k+1]\setminus\mathcal{S}}(\boldsymbol{X})) \\ &= H(Y \mid h_{[k+1]}(\boldsymbol{X})) \\ &= H(Y \mid h_{[k]}(\boldsymbol{X}), \boldsymbol{w}_{k+1}(\boldsymbol{X})) \\ &< H(Y \mid h_{[k]}(\boldsymbol{X})) \end{aligned} \tag{19}$$

Also note that

$$H(Y \mid h_{\mathcal{S}}(\boldsymbol{X}), \boldsymbol{X}) = H(Y \mid \boldsymbol{X}) = H(Y \mid h_{[k]}(\boldsymbol{X}), \boldsymbol{X}), \tag{20}$$

therefore:

$$H(Y \mid h_{\mathcal{S}}(\boldsymbol{X})) - H(Y \mid h_{\mathcal{S}}(\boldsymbol{X}), \boldsymbol{X}) < H(Y \mid h_{[k]}(\boldsymbol{X})) - H(Y \mid h_{[k]}(\boldsymbol{X}), \boldsymbol{X}), \tag{21}$$

which is Eq. 9. ■

Proof [Proof for Proposition 4] Let n_s be the number of rounds that LLM proposed at least one usable factor satisfying Assumption 3. Since t is large, its Binomial distribution $\mathrm{Binom}(t, p)$ can be approximated by Gaussian distribution $\mathcal{N}\left(tp, tp(1-p)\right)$. To enforce

$$\begin{aligned} \Pr\left(\frac{I\left(Y; \boldsymbol{X} \mid h_{\leq t}(\boldsymbol{X})\right)}{I\left(Y; \boldsymbol{X}\right)} < \epsilon\right) &\geq \Pr\Big(\left(1 - C_\Psi\right)^{n_s} < \epsilon\Big) \\ &= \Pr\Big(n_s > \frac{\log \epsilon}{\log\left(1 - C_\Psi\right)}\Big) \\ &= \Pr\Big(\frac{n_s - tp}{\sqrt{tp(1-p)}} > \frac{1}{\sqrt{tp(1-p)}}\left(\frac{\log \epsilon}{\log\left(1 - C_\Psi\right)} - tp\right)\Big) \\ &\geq 1 - \delta, \end{aligned} \tag{22}$$

we have

$$\frac{1}{\sqrt{tp(1-p)}}\left(\frac{\log \epsilon}{\log\left(1 - C_\Psi\right)} - tp\right) < z_\delta = -|z_\delta| \tag{23}$$

with Gaussian distribution approximation.

Isolating $\sqrt{t} > 0$ from the above inequality, we would have the desired result. ■

Proof [Proof for Proposition 6] Under the Gaussian assumption, the sample partial correlation r between Y and W given $\mathcal{M}$ can be transformed via Fisher's z-transformation:

$$z = \frac{1}{2} \ln \frac{1 + r}{1 - r}.$$

When the true partial correlation is ρ, the statistic z is approximately normally distributed with mean $\zeta = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ and variance $1/(n-|\mathcal{M}|-3)$ (Fisher, 1921). The null hypothesis $H_0 : \rho = 0$ is rejected at level α if

$$|z| > \frac{z_{1-\alpha/2}}{\sqrt{n-|\mathcal{M}|-3}}.$$

Under the alternative hypothesis $H_1 : \rho \neq 0$, the power of this two-sided test is

$$1-\beta \;=\; \Pr\left(|z| > \frac{z_{1-\alpha/2}}{\sqrt{n-|\mathcal{M}|-3}} \;\middle|\; \rho\right) \;\approx\; \Phi\Big(\sqrt{n-|\mathcal{M}|-3}\,|\zeta| - z_{1-\alpha/2}\Big),$$

where Φ is the cumulative distribution function of the standard normal distribution. Setting the non-centrality parameter to guarantee power at least $1-\beta$ yields

$$\sqrt{n-|\mathcal{M}|-3}\,|\zeta| - z_{1-\alpha/2} \;\geq\; z_{1-\beta}.$$

Solving for n and substituting $\zeta = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ gives the claimed bound. ■

Proof [Proof for Proposition 8] First we would show that the adjacency within $\mathcal{U}'$ are preserved given $\mathrm{MB}(\mathcal{U}')$. Then, we show the arrow heads would be preserved by analyzing the FCI orientation rules $\mathcal{R}0$–$\mathcal{R}4$ of Zhang (2008a).

Adjacency. Suppose $\alpha, \beta \in \mathcal{U}'$ are adjacent in $\mathcal{P}_\mathcal{E}$, then they cannot be m-separated by any subset of $\mathcal{E} \setminus \{\alpha, \beta\}$. Since $\mathcal{V} \setminus \{\alpha, \beta\} \subseteq \mathcal{E} \setminus \{\alpha, \beta\}$, they are adjacent in $\mathcal{P}_\mathcal{V}$.

Suppose $\alpha, \beta \in \mathcal{U}'$ are not adjacent in $\mathcal{P}_\mathcal{E}$. We can construct a set $\mathcal{C} \subseteq \mathcal{E}$ as follows: Begin with $\mathcal{C} = \emptyset$, if $\alpha \perp\!\!\!\perp \beta \mid \mathcal{C}$, stop the construction. For each path m-connecting α and β conditional on $\mathcal{C}$, if it consists of colliders, then removing any of these nodes from $\mathcal{C}$ leads to m-connection by previous processed path, thus α and β must be adjacent, this is a contradiction. Let θ be the non-collider closest to α, and put it into $\mathcal{C}$. We show $\theta \in \mathrm{MB}(\alpha)$ by induction: (1) if it is adjacent to α, then we are done; (2) if not, it would be connected with colliders from $\mathrm{MB}(\alpha)$, thus it has to be included in $\mathrm{MB}(\alpha)$. By this construction, we have $\alpha \perp\!\!\!\perp \beta \mid \mathcal{C}$ and $\mathcal{C} \subseteq \mathrm{MB}(\alpha) \subseteq \mathcal{V}$, which means $\alpha, \beta \in \mathcal{U}'$ are not adjacent in $\mathcal{P}_\mathcal{V}$.

Arrow heads. We construct $\widehat{\mathcal{P}}_\mathcal{E}$ from $\mathcal{P}_\mathcal{V}$ as follows: (1) we extend node set from $\mathcal{V}$ to $\mathcal{E}$; (2) Let the edges between $\mathcal{U}'$ and $\mathcal{E} \setminus \mathcal{U}'$ in $\widehat{\mathcal{P}}_\mathcal{E}$ be the same as $\mathcal{P}_\mathcal{E}$. By this construction, the skeleton in $\widehat{\mathcal{P}}_\mathcal{E}$ is same as $\mathcal{P}_\mathcal{E}$. If for some nodes $\alpha, \beta \in \mathcal{U}'$, the edge $\alpha \leftarrow\!* \beta$ occurs in $\mathcal{P}_\mathcal{E}$ but not in $\mathcal{P}_\mathcal{V}$, some orientation rules in $\mathcal{R}0 - \mathcal{R}4$ should be applicable in $\widehat{\mathcal{P}}_\mathcal{E}$ for some edges among $\mathcal{U}'$, and we will show this is impossible. If for some nodes $\alpha, \beta \in \mathcal{U}'$, the edge $\alpha \leftarrow\!* \beta$ occurs in $\mathcal{P}_\mathcal{V}$, some rules in $\mathcal{R}0 - \mathcal{R}4$ are applied.

If $\mathcal{R}0$ is applicable to $\widehat{\mathcal{P}}_\mathcal{E}$, then there exists an unshielded collider $< \alpha, \gamma, \beta >$ with $\gamma, \beta \in \mathcal{U}'$, and γ should have been processed by the loop in Algorithm 2. Therefore, since $\alpha \in \mathrm{MB}(\gamma) \subseteq \mathrm{MB}(\mathcal{U}')$, the triple is also unshielded in $\mathcal{P}_\mathcal{V}$, which means $\mathcal{R}0$ is applicable in $\mathcal{P}_\mathcal{V}$, contradicting with that $\mathcal{P}_\mathcal{V}$ is maximally informative. The similar argument can be applied on $\mathcal{R}1 - \mathcal{R}4$.

If $\mathcal{R}1$ is applicable to $\widehat{\mathcal{P}}_\mathcal{E}$, the antecedent implies an unshielded triple $< \alpha, \beta, \gamma >$, and $\beta, \gamma \in \mathcal{U}'$. Since there is an uncertain mark $\circ$ on β (i.e., $\beta \circ\!\!-\!\!* \gamma$), β should have been processed by the loop in Algorithm 2. Therefore, we have $\alpha \in \mathrm{MB}(\beta) \subseteq \mathrm{MB}(\mathcal{U}')$, α should

be included in $\mathcal{V}$ and the triple should also be unshielded in $\mathcal{P}_{\mathcal{V}}$. Thus, $\mathcal{R}1$ is also applicable to $\mathcal{P}_{\mathcal{V}}$, which is a contradiction.

If $\mathcal{R}2$ is applicable to $\widehat{\mathcal{P}}_{\mathcal{E}}$, then there exist $\alpha \to \beta \ast\!\!\to \gamma$ or $\alpha \ast\!\!\to \beta \to \gamma$, and $\alpha \ast\!\!-\!\!\circ \gamma$, and $\alpha, \gamma \in \mathcal{U}'$. Then γ should have been processed by the loop in Algorithm 2 due to its uncertain mark $\circ$. Therefore, $\alpha, \beta \in \mathrm{MB}(\gamma) \subseteq \mathrm{MB}(\mathcal{U}')$. Thus, $\mathcal{R}2$ is also applicable to $\mathcal{P}_{\mathcal{V}}$, which is a contradiction.

If $\mathcal{R}3$ is applicable to $\widehat{\mathcal{P}}_{\mathcal{E}}$, then there exist $\alpha \ast\!\!\to \beta \leftarrow\!\!\ast \gamma$, $\alpha \ast\!\!-\!\!\circ \theta \circ\!\!-\!\!\ast \gamma$, $\theta \ast\!\!-\!\!\circ \beta$ and $\theta, \beta \in \mathcal{U}'$. Therefore, β should have been processed by the loop in Algorithm 2, which means $\alpha, \gamma, \theta \in \mathrm{MB}(\beta) \subseteq \mathrm{MB}(\mathcal{U}')$ and thus $\mathcal{R}3$ is also applicable to $\mathcal{P}_{\mathcal{V}}$, i.e., contradiction.

If $\mathcal{R}4$ is applicable to $\widehat{\mathcal{P}}_{\mathcal{E}}$, its antecedent implies a discriminating path $p = < \theta, \cdots, \alpha, \beta, \gamma >$ where $\beta, \gamma \in \mathcal{U}'$ and $\beta \circ\!\!-\!\!\ast \gamma$ would be oriented as $\beta \to \gamma$ or $\beta \leftrightarrow \gamma$, θ is not adjacent to γ, and each nodes between θ and β is a collider and a parent of γ. Note that $\gamma \in \mathrm{MB}(\beta)$, thus all nodes between θ and β are included in $\mathrm{MB}(\beta)$ due to a path like $\alpha \to \gamma \leftarrow\!\!\ast \beta$. Then the path $\theta \ast\!\!\to \cdots \leftrightarrow \gamma \leftarrow\!\!\ast \beta$, we have $\theta \in \mathrm{MB}(\beta)$. Therefore $p \subseteq \mathrm{MB}(\beta) \subseteq \mathrm{MB}(\mathcal{U}')$, and hence $\mathcal{P}_{\mathcal{V}}$ is not maximally informative, contradiction.

■

Definition 17 (Higher order of Markov Blanket) *For a integer $p \in \mathbb{N}_+$ and a set of random variables $\mathcal{Z}$, $MB(X)$ is the Markov Blanket of X relative to $\mathcal{Z}$. If $p = 1$, then $MB^{(p)}(X) := MB(X)$ as usual; if $p > 1$, then $MB^{(p)}(X) := \bigcup_{X' \in MB^{(p-1)}(X)} MB(X')$.*

Proof [Proof for Proposition 13] Let $\boldsymbol{u}$ be a path from T to Y in $\mathcal{P}_{\mathcal{E}}$. We have $\boldsymbol{u} \subseteq \mathrm{MB}^{(2)}(\boldsymbol{u}) \subseteq \mathcal{V}$ by the Loop at step 3 in Algorithm 3. Therefore, its adjacency and arrow heads are preserved in $\mathcal{P}_{\mathcal{V}}$ according to Proposition 8. Thus $\boldsymbol{u}$ is a possibly directed path from T to Y in $\mathcal{P}_{\mathcal{V}}$ if and only if it is possibly directed in $\mathcal{P}_{\mathcal{E}}$.

Since $\mathcal{P}_{\mathcal{E}}$ is amenable relative to (T, Y), each possibly directed path $\boldsymbol{u}$ from T to Y starts with a visible directed edge $T \to V_1$. By the definition of visible edge, we need to consider two cases. In the first case, there is an unshielded triple $V_0 \ast\!\!\to T \to V_1$. We have $V_0 \in \mathrm{MB}(T)$, thus its adjacency and arrow heads would preserved by Proposition 8, and the arrow tail would be preserved by applying the FCI orientation rule $\mathcal{R}1$ of Zhang (2008a). In the second case, there is a node V_0 which is not adjacent to V_1, and a path $V_0 \ast\!\!\to \cdots \leftrightarrow \alpha_k \leftrightarrow T \to V_1$ that each non-endpoint node $(\alpha_1 \cdots \alpha_k)$ is a collider and a parent of V_1. This is a discriminating path, and we have shown this path is inside $\mathrm{MB}(T)$ in the proof for Proposition 8, and its arrow tails are preserved in $\mathcal{P}_{\mathcal{V}}$ according to Corollary 9.
■

Proof [Proof for Proposition 14] Since α and γ are two adjacent nodes in a possibly directed path from T to Y, we have $\{\alpha, \gamma\} \subseteq \mathrm{MB}^{(2)}(\{\alpha, \gamma\}) \subseteq \mathcal{V}$ by the Loop 1 in Algorithm 3.

We construct $\widehat{\mathcal{P}}_{\mathcal{E}}$ from $\mathcal{P}_{\mathcal{V}}$ as follows: (1) we extend node set from $\mathcal{V}$ to $\mathcal{E}$; (2) Let the edges between $\mathcal{U}'$ and $\mathcal{E} \setminus \mathcal{U}'$ in $\widehat{\mathcal{P}}_{\mathcal{E}}$ be the same as $\mathcal{P}_{\mathcal{E}}$. By this construction, the skeleton in $\widehat{\mathcal{P}}_{\mathcal{E}}$ is same as $\mathcal{P}_{\mathcal{E}}$. By Proposition 8, the arrow heads are also preserved.

If the edge $\alpha \to \gamma$ occurs in $\mathcal{P}_{\mathcal{E}}$ but not in $\mathcal{P}_{\mathcal{V}}$, i.e., $\alpha \circ\!\!\to \gamma$ in $\mathcal{P}_{\mathcal{V}}$, then some of the FCI orientation rules $\mathcal{R}8$–$\mathcal{R}10$ from Zhang (2008a) should be applicable in $\widehat{\mathcal{P}}_{\mathcal{E}}$ on this edge, and we will show that this is impossible.

If $\mathcal{R}8$ is applicable to $\alpha \circ\!\!\rightarrow \gamma$ in $\mathcal{P}_{\mathcal{E}}$, then we have $\alpha \rightarrow \beta \rightarrow \gamma$, since there is an uncertain mark $\circ$ on α, it has been processed by the Loop 2 in Algorithm 3. Therefore, $\gamma \in \mathrm{MB}(\alpha) \subseteq \mathrm{MB}^{(2)}(\{\alpha, \gamma\}) \subseteq \mathcal{V}$, thus $\mathcal{R}8$ could have been applied to $\mathcal{P}_{\mathcal{V}}$, and violates the fact that $\mathcal{P}_{\mathcal{V}}$ is maximally informative.

If $\mathcal{R}9$ is applicable to $\alpha \circ\!\!\rightarrow \gamma$ in $\mathcal{P}_{\mathcal{E}}$, then there is a uncovered possibly directed path $\boldsymbol{p}$ from α to γ. Since $\boldsymbol{p}$ is a part of possibly directed path from T to Y, it holds that $\boldsymbol{p} \subseteq \mathrm{MB}^{(2)}(\boldsymbol{p}) \subseteq \mathcal{V}$ by the Loop 1 in Algorithm 3, thus $\mathcal{R}9$ could have been applied to $\mathcal{P}_{\mathcal{V}}$, and thus leads to a contradiction.

If $\mathcal{R}10$ is applicable to $\alpha \circ\!\!\rightarrow \gamma$ in $\mathcal{P}_{\mathcal{E}}$, then we have $\beta \rightarrow \gamma \leftarrow \theta$, $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are uncovered possibly directed paths from α to β, and from α to θ respectively. Note that $\beta, \theta \in \mathrm{MB}(\alpha)$, and $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are in part of possibly directed path from T to Y. Therefore, with same argument, this would lead to a contradiction. ■

Proof [Proof for Proposition 16] By the construction of Algorithm 3, we have a set $\mathcal{U}'$ that $\mathrm{MB}(\mathcal{U}') \subseteq \mathcal{V}$ and $\mathrm{MB}(\{T, Y\}) \subseteq \mathcal{V}$, and for any node V on a path between T and Y, $V \in \mathcal{U}'$. To show $\mathcal{Z}$ is also an adjustment set relative to (T, Y) in $\mathcal{P}_{\mathcal{V}}$, we need to check whether it satisfies the three conditions in Definition 12.

(1) Amenability. Since $\mathcal{P}_{\mathcal{E}}$ has an adjustment set, then it has to be amenable, therefore, according to Proposition 13, $\mathcal{P}_{\mathcal{V}}$ is amenable.

(2) Forbidden set. We need to show that: if a node $\alpha \in \mathcal{U}'$ satisfies $\alpha \notin \mathrm{Forb}(T, Y, \mathcal{P}_{\mathcal{V}})$, then $\alpha \notin \mathrm{Forb}(T, Y, \mathcal{P}_{\mathcal{E}})$. By definition, if $\alpha \in \mathrm{Forb}(T, Y, \mathcal{P}_{\mathcal{E}})$, then there is a possible directed path from T to α. Thus, by Proposition 8, there is no such a path in $\mathcal{P}_{\mathcal{V}}$ containing an arrowhead towards T, thus we have $\alpha \in \mathrm{Forb}(T, Y, \mathcal{P}_{\mathcal{V}})$. On the other hand, if $\alpha \in \mathrm{Forb}(T, Y, \mathcal{P}_{\mathcal{V}})$ there is a possibly directed path $\boldsymbol{u}$ from $\beta \in \mathrm{S}(T, Y, \mathcal{P}_{\mathcal{V}})$ to α. By the construction of Algorithm 3, $\beta \in \mathrm{S}(T, Y, \mathcal{P}_{\mathcal{E}})$ and $\boldsymbol{u}$ is also a possibly directed path in $\mathcal{P}_{\mathcal{E}}$, thus $\alpha \in \mathrm{Forb}(T, Y, \mathcal{P}_{\mathcal{E}})$.

(3) Blocking. By Proposition 14 and Corollary 15, each definite status paths between T and Y are preserved. Thus if a node can m-separate some such paths in $\mathcal{P}_{\mathcal{E}}$, it m-separates them in $\mathcal{P}_{\mathcal{V}}$. ■

Appendix C. More Details about Experiments

C.1 More Details on Constructing AppleGastronome

In the AppleGastronome benchmark, we consider the target variable as a rating score of the apple by several gastronomes. Each apple has its own attributes, including size, smell, and taste (or sweetness). Each gastronome has a unique preference for some attributes of the apple. They will give a rating as well as write a review according to the matchness of the apple with respect to their preference. We generate the review using GPT-4 by fetching GPT-4 the preferences and the apple attributes.

The prompts for generating the unstructured inputs are given in Figure 28.

```
You are a picky gastronome on apples. You
are ready to evaluate apples and write
reviews. Your writing should be clear,
solid and convincing to suppliers and
customers.

Task
Please write a short review about the
evaluation results for a given apple.

Evaluation Results:
- a high content of essential nutrients,
including dietary fiber, vitamin C, etc.
- musty or rotten
- more sour than sweet
- dry and lacking Moisture
- might not bring the expected returns and
could even lead to losses

Additional Requirement:
- Combine those evaluation results into
more detailed reasoning.
- Single paragraph; No quotation marks.
- Modern English
- Begin this part by a mark [Review Begin].
- no more than 60 words.
```

Figure 28: Illustration of prompts for generating AppleGastronome.

Examples of AppleGastronome are given in Figure 29.

C.2 More Details on Prompts for AppleGastronome

The prompts for factor proposal are given in Figure 30.

The prompt for factor annotation is given in Figure 31.

The prompts for constructing feedback are given in Figure 32.

```
## Score: -3
- The assessed apple falls short in offering the essential nutrients that consumers demand,
presenting a concerning deficiency. Its diminutive stature coupled with a distinct tartness that
overshadows any sweetness detracts from its palate appeal. Moreover, the dry texture fails to
satisfy those seeking a juicy bite. Given these drawbacks, this apple variety may disappoint
buyers, posing a financial risk for suppliers.

- Despite its diminutive stature, this apple fails to meet even basic quality expectations. The
  presence of a musty or rotten scent signals underlying issues that can neither be overlooked
  by discerning consumers nor excused by vendors seeking profit. Ultimately, stocking this fruit
  is a risky venture that's far more likely to incur losses than yield any anticipated returns.

… …

## Score: 3
- This apple variety, rich in essential nutrients like dietary fiber and vitamin C, impresses
with its large size and delightful strong aroma. Its balanced flavor, leaning more towards sweet
than sour, and abundant refreshing moisture guarantee a delectable eating experience. This
apple's significant market potential is evident, and it certainly merits broader acclaim for its
high-quality attributes.
```

Figure 29: Illustration of examples in AppleGastronome.

C.3 More Details of Results on AppleGastronome

The detailed causal graph results are given from Figure 33 to Figure 37. Independent tests about annotation on the Apple Gastronome benchmark are shown in Table 5. Full Results of Causal Metrics each Round of each LLM on Apple Gastronome Benchmark are shown in Table 15.

C.4 Implementation of the FCI Algorithm

We use a third-party open-sourced Python library to perform the FCI algorithm: `https://causal-learn.readthedocs.io/en/latest/`.

We set $\alpha = 0.05$, and `independence_test_method="fisherz"` throughout all experiments. Other parameters are kept as the default.

You are an excellently helpful AI assistant for analyzing, abstracting, and processing data. Now try your best.

Rating=0
…
Rating=1
…
Rating=2
…

Tasks: Factor Abstraction.
What are the high-level factors behind the text that contribute to the allocation of groups?
- Propose as many as possible factors based on your knowledge.
- Adjust your candidate factors by your observation on given data.
- The proposed factors should be diverse and different semantically. Their overlapping should be minimized.

Figure 30: Illustration of the prompt for factor proposal.

```
You are an excellently helpful AI assistant
for analyzing, abstracting, and processing
data.
Now try your best.

# Data

The reviewer comments on an apple that are
randomly picked:
{text}


# Tasks

For the given sample, what is the most
appreciate factor value based on the
criterion?

{factor_state}


**Your final output should follow this
template**:
**"The value is: ___."**
```

Figure 31: Illustration of the prompt for factor annotation.

```
# Data
… <groups of samples> …

# Tasks: Factor Abstraction.

**What are the high-level factors behind the text that contribute to the
allocation of groups?**

- Propose your candidate factors based on your observation on given data.
- The proposed factors should be diverse and different semantically. Their
  overlapping should be minimized.

# About Output

Your output should contain parts described as follows.

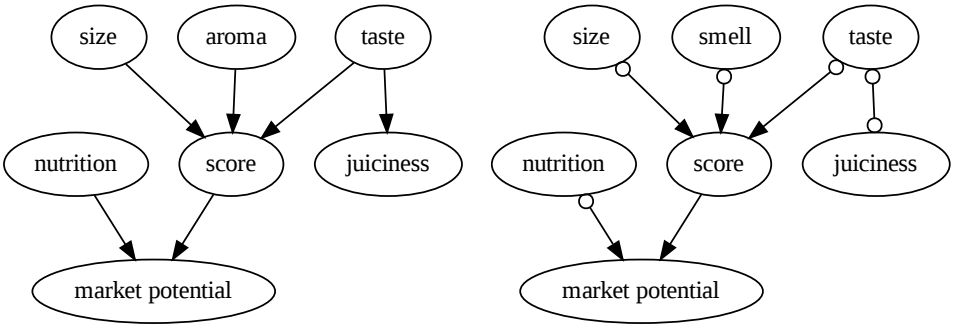
**Part 1**: Consideration.

In this part, feel free to write down the process of your considerations.
Hint: You need to abstract, identify, and design suitable factors with
corresponding criteria, and each factor is only allowed to take value from
[-1, 0, 1] .

**Part 2**: Factor filtration. You should decide whether to use each of the
proposed factors by following criteria:

- Each new factor should be helpful to distinguish the groups.
- Each Factor should focus on one concrete aspect and try to avoid
  overlapping.
- Avoid overlapping with those existing factors:
  - size
  - aroma
```

Figure 32: Illustration of the prompt for feedback.

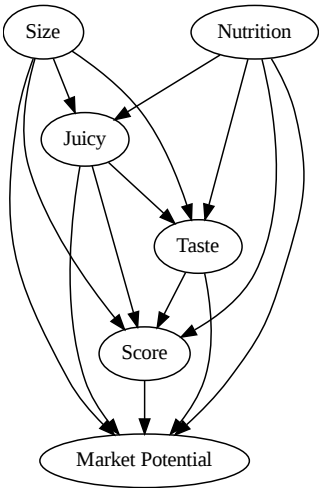


(a) Ground truth (b) FCI results

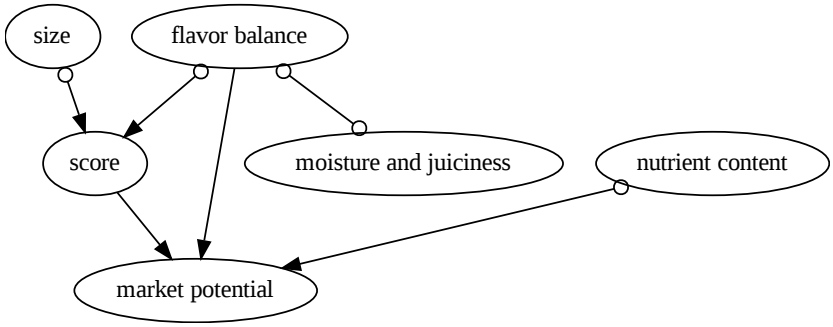
Figure 33: Ground truth and faithful (via FCI algorithm) causal graphs in AppleGastronome.

LLMs	COAT Round	Perception Score	Capacity Score	$I(y; x\|h_S)$
GPT-4o	iter 1	0.82±*0.02*	0.24±*0.01*	0.28±*0.02*
	iter 2	0.44±*0.08*	0.26±*0.02*	0.21±*0.03*
	iter 3	0.38±*0.12*	0.19±*0.09*	0.24±*0.05*
GPT-4	iter 1	0.78±*0.02*	0.18±*0.02*	0.45±*0.04*
	iter 2	0.61±*0.08*	0.26±*0.05*	0.28±*0.05*
	iter 3	0.39±*0.28*	0.22±*0.00*	0.29±*0.02*
GPT-3.5	iter 1	0.81±*0.02*	0.21±*0.01*	0.42±*0.05*
	iter 2	0.52±*0.11*	0.23±*0.02*	0.31±*0.04*
	iter 3	0.38±*0.10*	0.21±*0.03*	0.33±*0.11*
Mistral-Large	iter 1	0.78±*0.02*	0.20±*0.01*	0.41±*0.04*
	iter 2	0.61±*0.08*	0.20±*0.03*	0.29±*0.02*
	iter 3	0.33±*0.33*	0.32±*0.00*	0.25±*0.05*
Mistral-Medium	iter 1	0.78±*0.02*	0.20±*0.02*	0.39±*0.05*
	iter 2	0.39±*0.15*	0.19±*0.10*	0.32±*0.04*
	iter 3	0.36±*0.10*	0.37±*0.14*	0.25±*0.08*
LLaMA2-3-70b	iter 1	0.75±*0.00*	0.19±*0.02*	0.47±*0.08*
	iter 2	0.36±*0.31*	0.19±*0.00*	0.38±*0.02*
	iter 3	0.27±*0.38*	0.18±*0.00*	0.36±*0.03*
LLaMA2-2-70b	iter 1	0.77±*0.02*	0.21±*0.05*	0.43±*0.13*
	iter 2	0.08±*0.12*	0.35±*0.00*	0.35±*0.03*
	iter 3	0.28±*0.04*	0.14±*0.07*	0.33±*0.03*
Qwen-1.5-110B	iter 1	0.77±*0.02*	0.17±*0.01*	0.45±*0.09*
	iter 2	0.56±*0.16*	0.21±*0.12*	0.32±*0.08*
	iter 3	0.28±*0.21*	0.18±*0.02*	0.33±*0.11*
DeepSeek-V2	iter 1	0.89±*0.08*	0.17±*0.02*	0.42±*0.05*
	iter 2	0.62±*0.10*	0.20±*0.02*	0.34±*0.02*
	iter 3	0.67±*0.24*	0.40±*0.12*	0.29±*0.06*
Claude-3-Opus	iter 1	0.77±*0.02*	0.18±*0.00*	0.43±*0.06*
	iter 2	0.56±*0.08*	0.39±*0.11*	0.24±*0.06*
	iter 3	0.75±*0.25*	0.15±*0.00*	0.22±*0.05*

Table 15: Full Results of Causal Metrics in Each Round for Each LLM on the Apple Gastronome Benchmark.

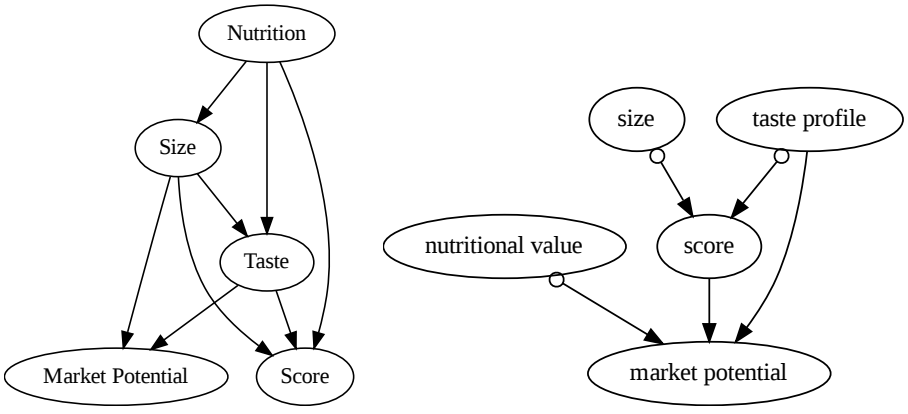


(a) GPT-4 reasoning



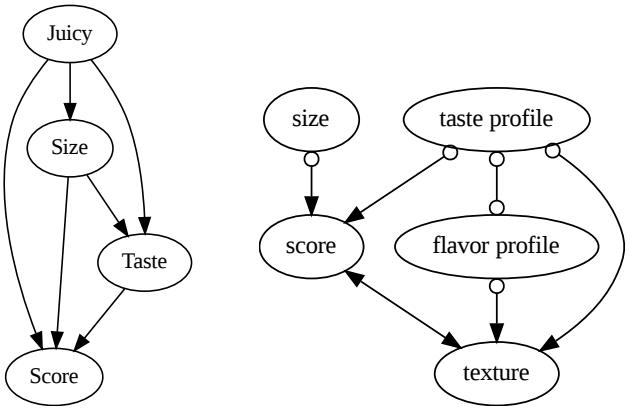
(b) GPT-4 COAT

Figure 34: Causal graphs with GPT-4 in AppleGastronome.



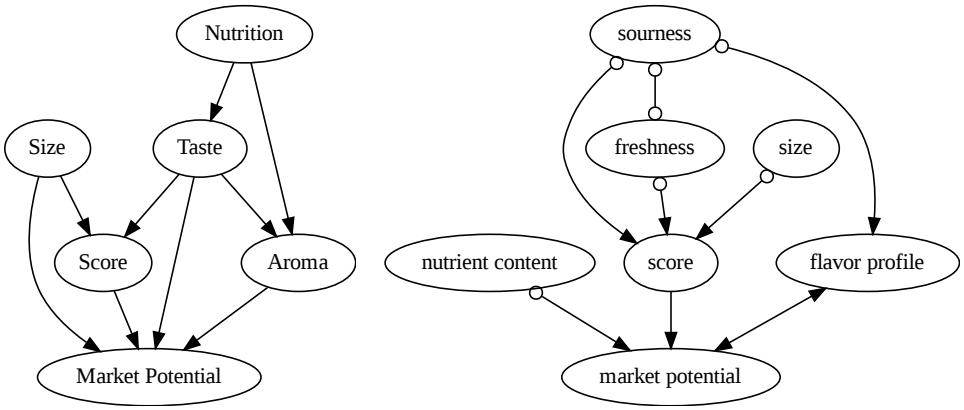
(a) GPT-3.5 reasoning (b) GPT-3.5 COAT

Figure 35: Causal graphs with GPT-3.5 in AppleGastronome.



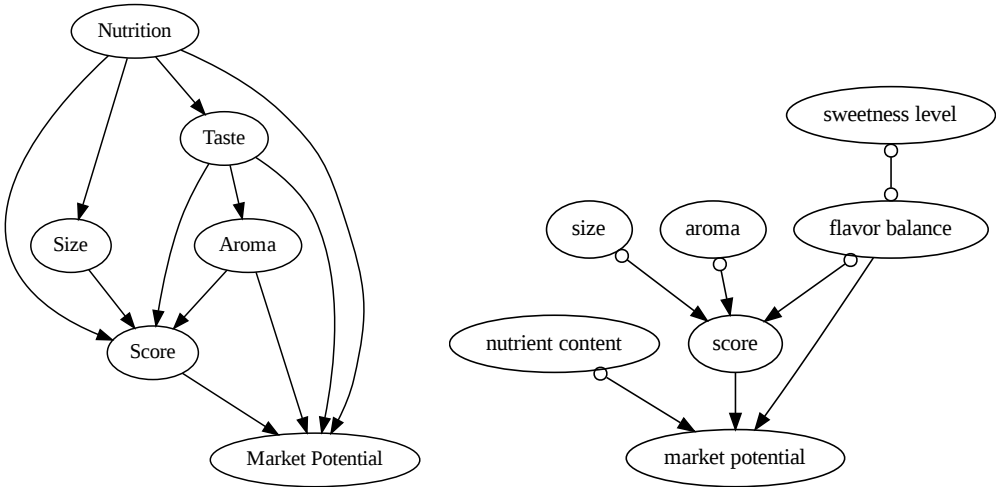
(a) LLaMA2 reasoning (b) LLaMA2 COAT

Figure 36: Causal graphs with Llama-2 in AppleGastronome.



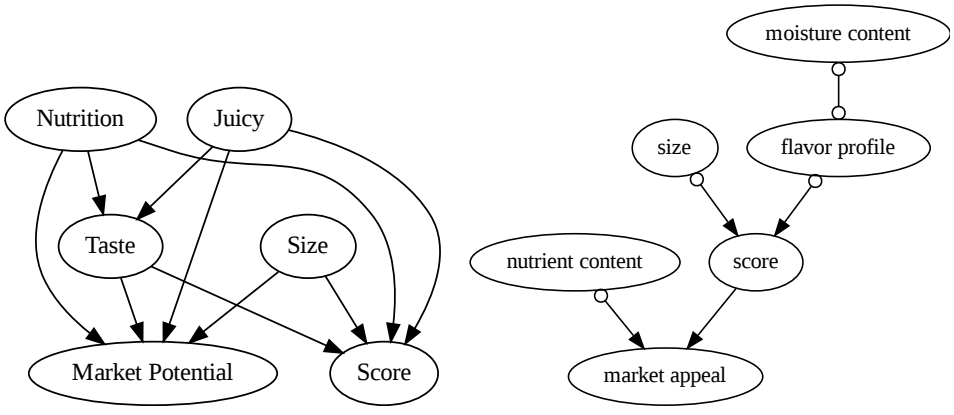
(a) Mistral reasoning (b) Mistral COAT

Figure 37: Causal graphs with mistral Medium in AppleGastronome.



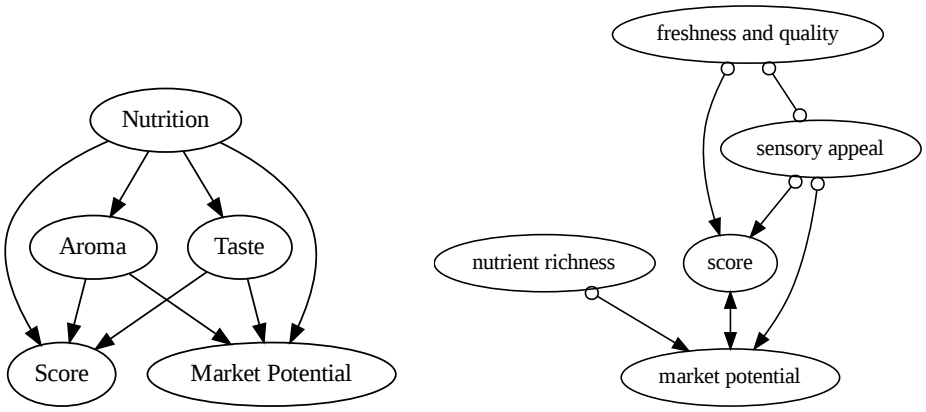
(a) Claude-3-Opus reasoning (b) Claude-3-Opus COAT

Figure 38: Causal graphs with Claude-3-Opus in AppleGastronome.



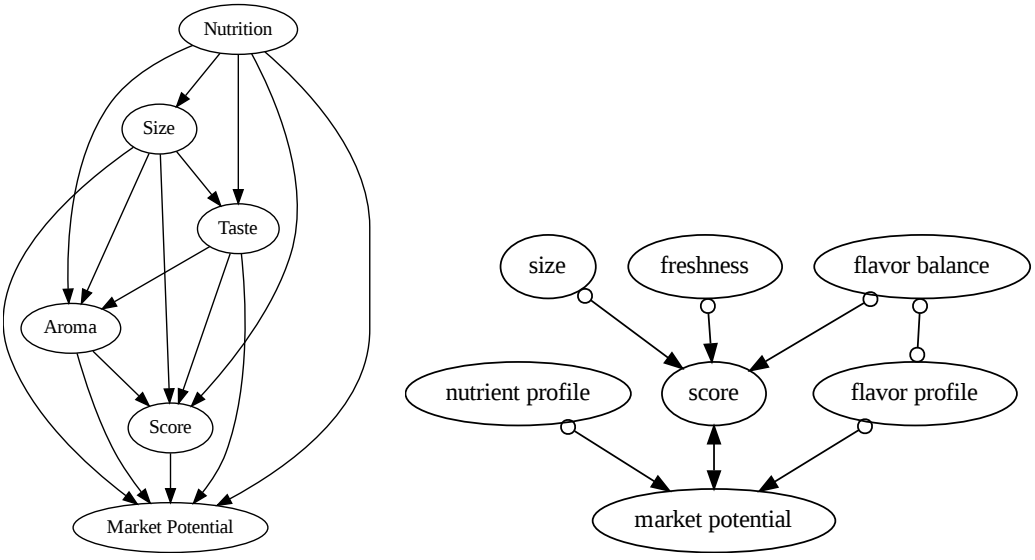
(a) DeepSeek-V2 reasoning (b) DeepSeek-V2 COAT

Figure 39: Causal graphs with DeepSeek-V2 in AppleGastronome.



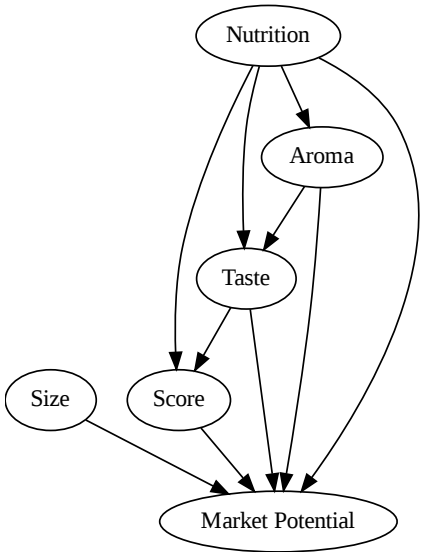
(a) Llama-3-70b reasoning (b) Llama-3-70b COAT

Figure 40: Causal graphs with Llama-3-70b in AppleGastronome.

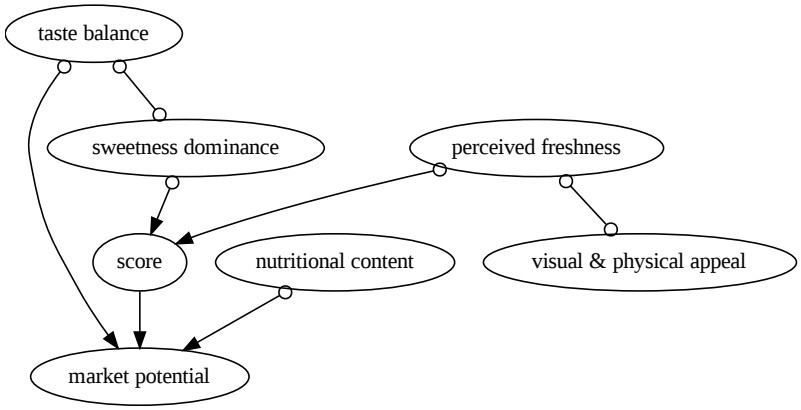


(a) mistral-Large reasoning (b) mistral-Large COAT

Figure 41: Causal graphs with mistral-Large in AppleGastronome.

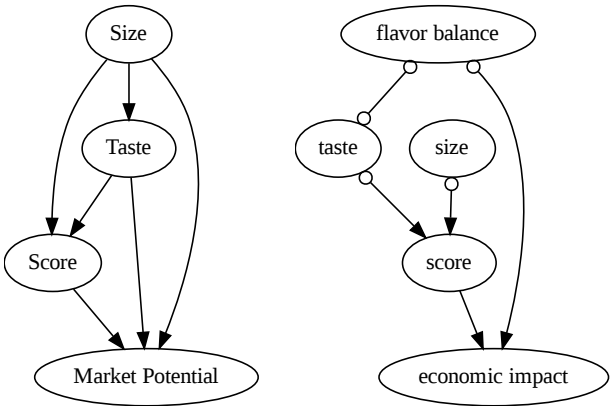


(a) qwen-1.5-110B reasoning



(b) qwen-1.5-110B COAT

Figure 42: Causal graphs with qwen-1.5-110B in AppleGastronome.



(a) GPT-4o reasoning (b) GPT-4o COAT

Figure 43: Causal graphs with GPT-4o in AppleGastronome.

C.5 More Details on Constructing Neuropathic

```
You are a helpful and professional
assistant to Neurologists. Try your
best!

Write a clinical notes based on
following information:

Patient id: Neuro_{idx:05}
Age: {10+np.random.choice(50)}
Gender: {get_gender()}

---

Write one summary (one paragraph, do not
do inference) about **symptom
diagnosis**:
  {symptom_notes}
---
L: Left
R: Right
DLI: discoligamentous injury
Radi: radiculopathy

---

Your clinical notes should surrounded by
<note> ... </note>
```

Figure 44: Illustration of prompts for generating Neuropathic.

In the Neuropathic benchmark, we convert the data set into a clinical diagnosis task. In the original data set, there are three levels of causal variables, including the symptom-level, radiculopathy-level and the pathophysiology-level. In experiments, we mainly consider the target variable of right shoulder impingement. When generating the clinical diagnosis notes as $\boldsymbol{x}$ using GPT-4, we will avoid any mentioning of variables other than symptoms.

As we intend to leverage the Neuropathic benchmark to simulate the real-life diagnosis, after the factor proposal stage, we directly incorporate external experts that measure the

values of the candidate factors. The prompts to generate the diagnosis records are given in Figure 44.

Examples of Neuropathic are given in Figure 45.

```
## Group with 'R shoulder impingement'=0
----
Patient ID: Neuro_00000
Age: 57
Gender: Male

Symptom Diagnosis Summary:
The patient presents with a left shoulder impingement and associated intracapsular problems. The
symptoms include limited and painful range of motion with a positive outcome when performing a
"thumbs up" test on the left side. Right upper limb assessment also reveals issues; the patient
wears an armband suggesting ongoing management for right lower arm disorders. The patient
reports lumbago and pain localized to the left thigh. Physical therapy assessment (L pta) is
noted, indicating involvement in a therapeutic intervention. Joint assessments reveal a left hip
joint condition and right hip arthritis, accompanied by right hamstring discomfort.
```

Figure 45: Illustration of examples in Neuropathic.

C.6 More Details of Results on Neuropathic

The detailed causal graph results are given from Figure 46 to Figure 50.

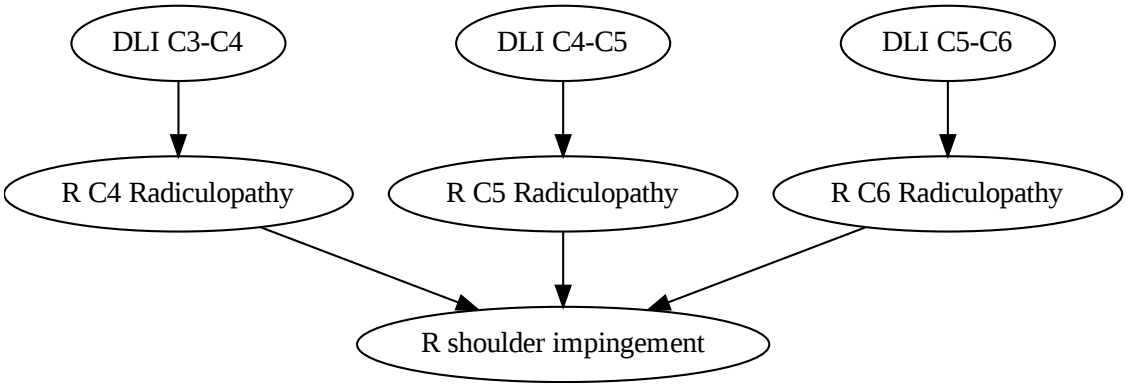
C.7 Discussion on the Time Complexity

Assume there are m samples with n possible factors.

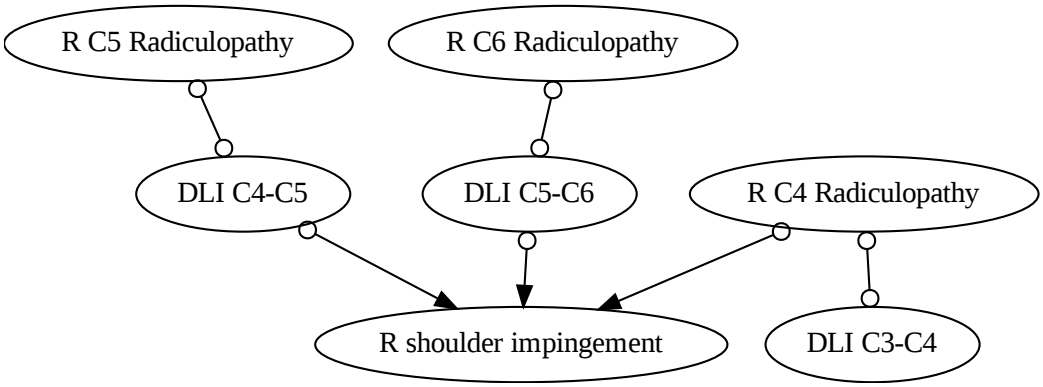
For Factor Proposal: META only needs to interact once with LLM, so it is $O(1)$. DATA runs one single COAT round, so it is $O(1)$. COAT interacts with LLMs multiple rounds. In each round, at least one new factor should be proposed; otherwise, the loop will stop. So it is $O(n)$.

For Factor Annotation: META: Not applicable. DATA: At most n factors would be proposed in a single round. And each of them needs m times annotations by LLMs for all samples. So it is $O(nm)$. COAT: At most n factors would be proposed during all rounds. And each of them needs m times annotations by LLMs for all samples. So it is $O(nm)$.

For Causal Discovery: Pair-wise reasoning by LLMs: $O(n^2)$. COAT: LLM is not involved. The computational cost of it depends on the numeric methods. The FCI algorithm used by COAT has a time complexity that grows exponentially with n. Learning the causal graph over a large number of nodes effectively is still an open problem in causal discovery literature.



(a) Ground truth

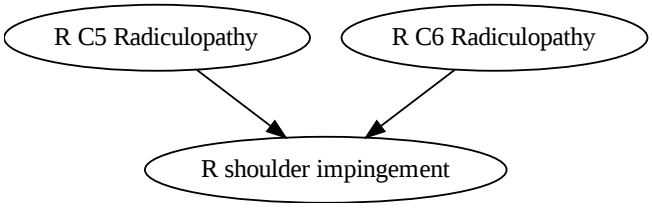


(b) FCI results

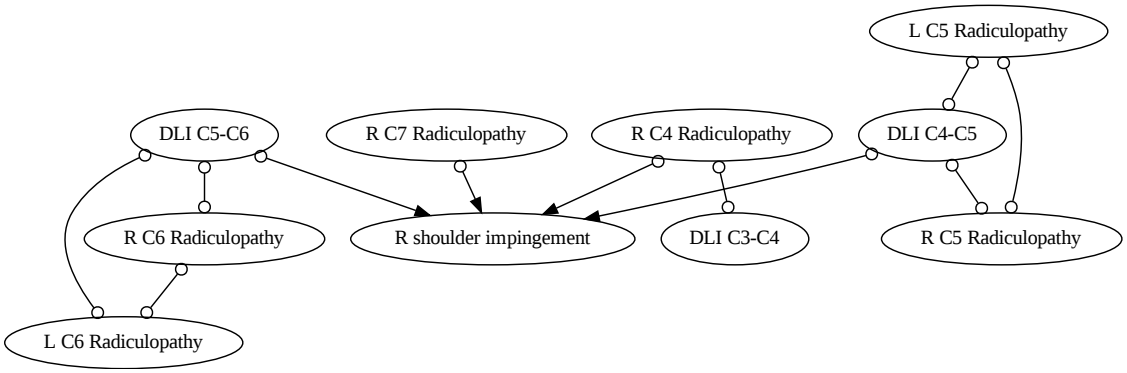
Figure 46: Ground truth and faithful (via FCI algorithm) causal graphs in Neuropathic.

C.8 Resources

We utilized a system comprising two Intel Xeon E5-2630v4 processors with 2.2GHz, two NVIDIA Tesla P40 GPUs, and 256 GB of memory. For conversations with large language models (LLMs), we leveraged the poe.com platform, while annotations were facilitated using the OpenAI API and the Mistral API.



(a) GPT-4 reasoning

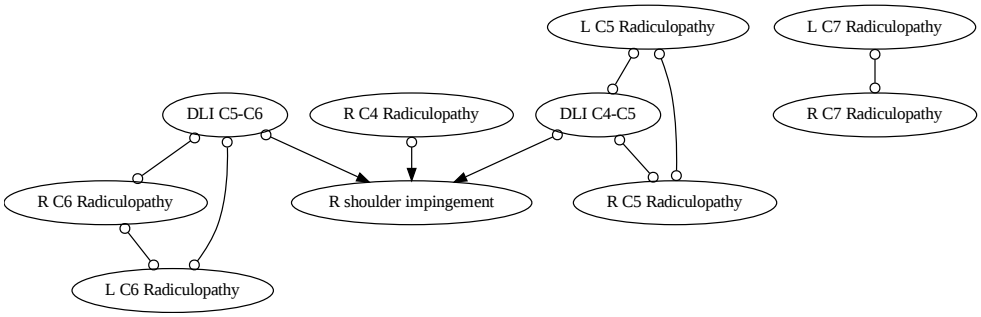


(b) GPT-4 COAT

Figure 47: Causal graphs with GPT-4 in Neuropathic.

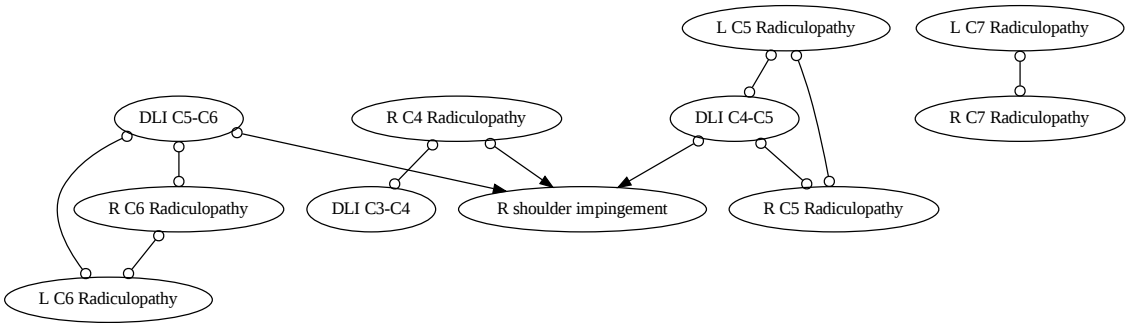


(a) GPT-3.5 reasoning

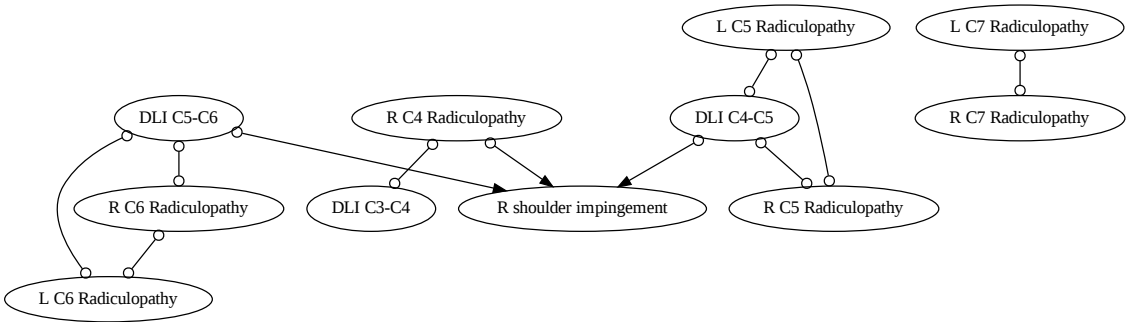


(b) GPT-3.5 COAT

Figure 48: Causal graphs with GPT-3.5 in Neuropathic.

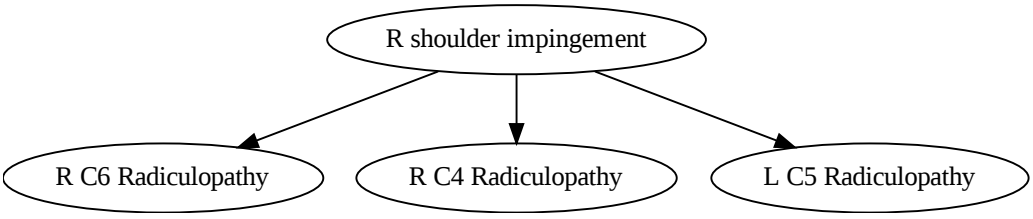


(a) LLaMA-2-70b reasoning

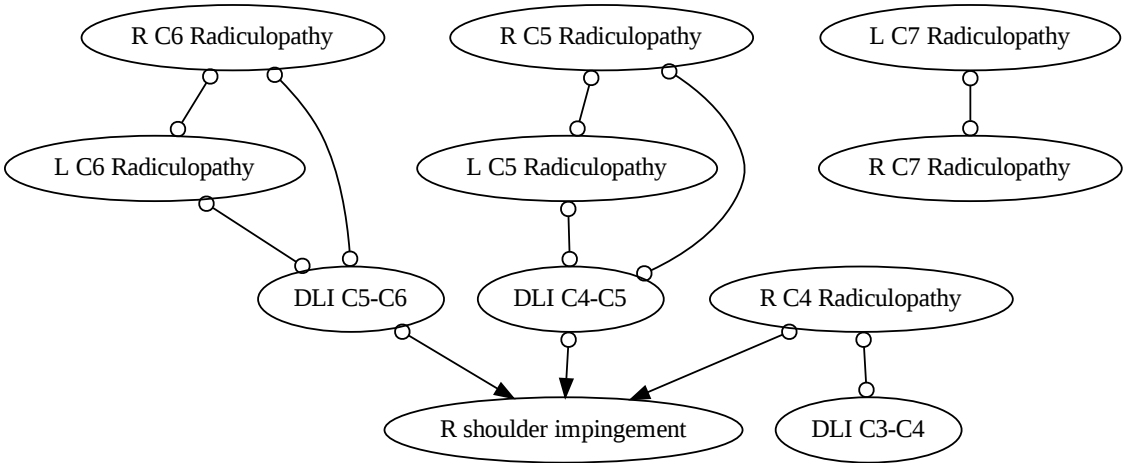


(b) LLaMA-2-70b COAT

Figure 49: Causal graphs with LLaMA-2-70b in Neuropathic.



(a) Mistral-Medium reasoning



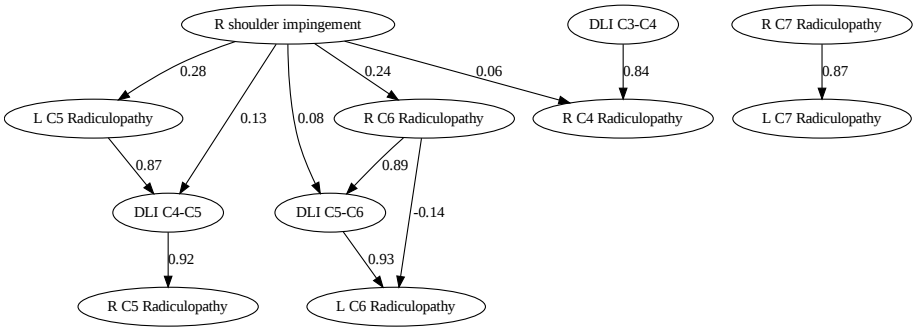
(b) Mistral-Medium COAT

Figure 50: Causal graphs with Mistral-Medium in Neuropathic.

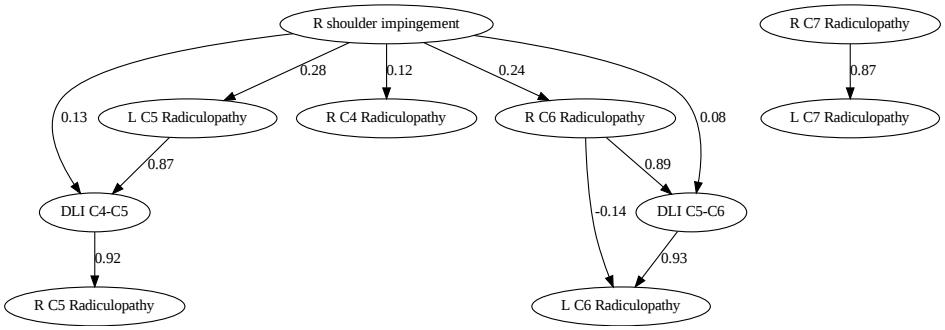
LLM	Method	Factor Proposal				
		PA	AN	OT	Acc	F1
GPT-4	Meta	3	5	6	0.91	0.59
	Data	2	2	0	0.95	0.50
	COAT	3	6	3	0.96	0.80
	Data (LiNGAM)	3	3	0	0.96	0.67
	COAT (LiNGAM)	3	6	4	0.95	0.75
GPT-3.5	Meta	3	5	6	0.91	0.59
	Data	3	5	4	0.94	0.67
	COAT	3	5	2	0.96	0.77
	Data (LiNGAM)	2	3	4	0.91	0.46
	COAT (LiNGAM)	3	5	4	0.94	0.67
LLaMA2-70b	Meta	2	4	5	0.91	0.53
	Data	3	3	1	0.95	0.60
	COAT	3	6	2	0.97	0.86
	Data (LiNGAM)	3	4	4	0.92	0.57
	COAT (LiNGAM)	3	6	4	0.95	0.75
Mistral-Med	Meta	3	6	3	0.96	0.80
	Data	3	3	2	0.94	0.66
	COAT	3	6	2	0.97	0.86
	Data (LiNGAM)	3	3	3	0.92	0.50
	COAT (LiNGAM)	3	6	3	0.96	0.80

Table 16: Causal discovery results in Neuropathic. PA, AN, and OT refer to the parents, ancestors, and others, respectively. Accuracy and F1 measure the recovery of the causal ancestors.

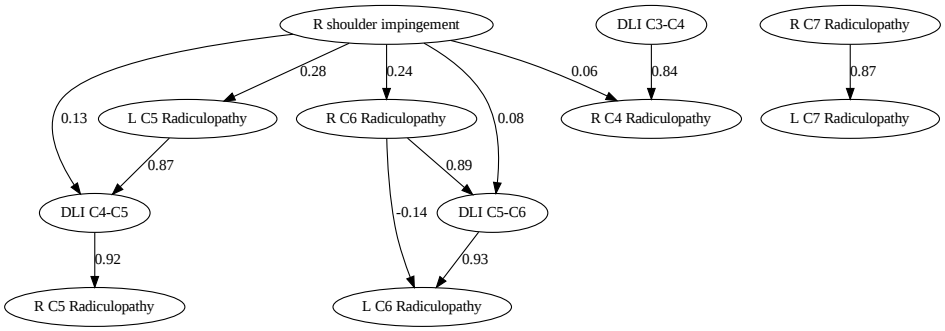
Appendix D. COAT with Different Causal Discovery Algorithms



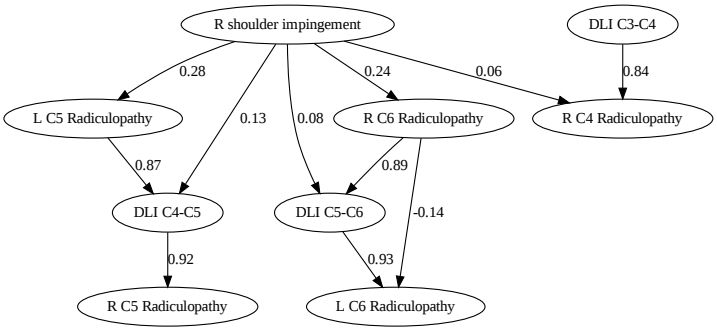
(a) GPT-4 COAT with LiNGAM



(b) GPT-3.5 COAT with LiNGAM



(c) llama-2-70b COAT with LiNGAM



(d) Mistral-med COAT with LiNGAM

Figure 51: The discovered causal graphs in Neuropathic with LiNGAM.

Appendix E. Summary of Benchmark Data

Benchmark Name	Type	Source	Sample	Sample Size	Ground Truth
Apple Gastronome	Synthetic Data	This paper. It will be open-sourced under CC-BY 4.0.	Textual	200 textual reviews.	Yes
Neuropathic	Semi Real-world Data	Tabular samples from Tu et al. (2019) on their GitHub repo under CC-BY 4.0. Textual samples are generated by this paper.	Textual	100 synthetic textual data for prompt construction; 1000 tabular data for CI tests.	Yes
Brain Tumor	Real-world Data	An open Kaggle data set (kaggle/brain-tumor-classification-mri) with an open-sourced project. (Bhuvaji et al., 2024). MIT License.	Image	60 MRI samples for COAT and CI tests. The sample size is small due to the expensive API cost. One can easily construct a much larger data set by following our instructions.	No
Stock News	Real-world Data	An open Kaggle data set (kaggle/stock-price-and-news-realted-to-it) (Innovations, 2023). MIT License.	Textual with time index	200 samples for COAT experiments. 400 samples for empirical validation.	No
ENSO	Real-world Data	The NOAA 20th Century Reanalysis V3 data set (Compo et al., 2011) from their website at https://psl.noaa.gov (with CC0 1.0 License). We developed a set of simple tools for LLMs to utilize the official database.	NetCDF	Monthly observation for 1836/01 to 2015/12; 1.0 degree latitude x 1.0 degree longitude global grid (360x181); 90+ climate variables.	No

Table 17: Summary of Benchmark Data

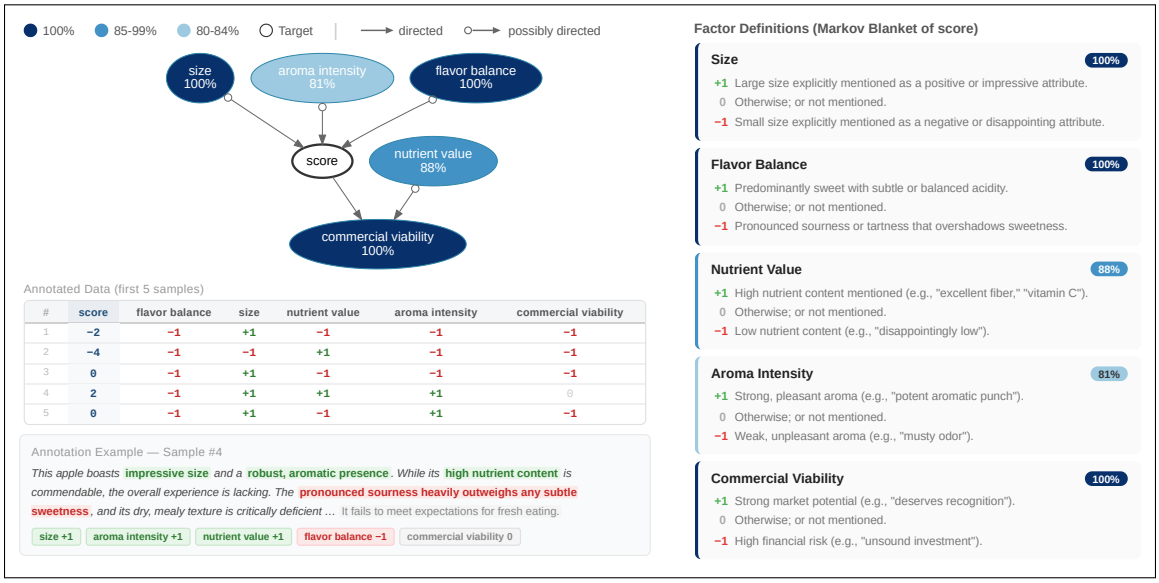


Figure 52: Schematic overview of COAT-MB output components for end users. It has four integrated modules (conceptual layout): factor definitions with uncertainty indicators (right), the inferred causal structure (upper-left), structured annotation results (middle-left), and annotation details (bottom-left).

Appendix F. COAT Output for Users in Practice

Figure 52 presents a conceptual layout of COAT-MB's practical output. The schematic illustrates four core components presented to users: *(i) Factor definitions* (right): the final factors identified by the system with uncertainty quantification; *(ii) Causal graph* (upper-left): the learned causal structure among the factors and target variable; *(iii) Annotation tables* (middle-left): structured data with factor values after the annotation process on raw samples; and *(iv) Annotations details* (bottom-left): the annotation details for raw samples, possibly with the rationale of the decisions about each factor's value.

References

Ahmed Abdulaal, Adamos Hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2024.

Raj Agrawal, Chandler Squires, Karren Yang, Karthikeyan Shanmugam, and Caroline Uhler. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3400–3409. PMLR, 2019.

Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. In *Advances in Neural Information Processing Systems*, pages 15516–15528, 2022.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pages 372–407, 2023.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2009.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

Anthropic. Claude 3 family. `https://www.anthropic.com/news/claude-3-family`, 2024. Accessed: 2024-05-20.

Taiyu Ban, Lyuzhou Chen, Xiangyu Wang, and Huanhuan Chen. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint*, arXiv:2306.16902, 2023.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 507–556. Association for Computing Machinery, 2022. doi: 10.1145/3501714.3501743.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". *arXiv preprint*, arXiv:2309.12288, 2023.

Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, and Sameer Dedge. Brain tumor classification using deep learning algorithms. `https://github.com/SartajBhuvaji/Brain-Tumor-Classification-Using-Deep-Learning-Algorithms`, 2024. Accessed: 2024-05-19.

Jakob Bjerknes. Atmospheric teleconnections from the equatorial Pacific. *Monthly Weather Review*, 97(3):163–172, 1969.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, pages 21865–21877, 2020.

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. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint*, arXiv:2303.12712, 2023.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 2023.

Stephen Burgess and Simon G Thompson. *Mendelian randomization: methods for using genetic variants in causal estimation*. CRC Press, 2015.

Leland Bybee, Bryan Kelly, Asaf Manela, and Dacheng Xiu. Business news and business cycles. *Journal of Finance*, 79(5):3105–3147, 2024. doi: 10.1111/jofi.13377.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, Kaili Ma, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. In *Advances in Neural Information Processing Systems*, 2022a.

Yongqiang Chen, Yatao Bian, Kaiwen Zhou, Binghui Xie, Bo Han, and James Cheng. Does invariant graph learning via environment augmentation learn invariance? In *Advances in Neural Information Processing Systems*, 2023a.

Yongqiang Chen, Wei Huang, Kaiwen Zhou, Yatao Bian, Bo Han, and James Cheng. Understanding and improving feature learning for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, 2023b.

Yongqiang Chen, Kaiwen Zhou, Yatao Bian, Binghui Xie, Bingzhe Wu, Yonggang Zhang, Kaili Ma, Han Yang, Peilin Zhao, Bo Han, and James Cheng. Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization. In *International Conference on Learning Representations*, 2023c.

Yongqiang Chen, Yatao Bian, Bo Han, and James Cheng. Interpretable and generalizable graph learning via subgraph multilinear extension. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

Zhengming Chen, Feng Xie, Jie Qiao, Zhifeng Hao, Kun Zhang, and Ruichu Cai. Identification of linear latent variable model with arbitrary distribution. In *AAAI Conference on Artificial Intelligence*, pages 6350–6357, 2022b.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. Unveiling causal reasoning in large language models: Reality or mirage? *Advances in Neural Information Processing Systems*, 37:96640–96670, 2024.

Kristy Choi, Chris Cundy, Sanjari Srivastava, and Stefano Ermon. LMPriors: Pre-trained language models as task-specific priors. *arXiv preprint*, arXiv:2210.12530, 2022.

Gilbert P. Compo, Jeffrey S. Whitaker, Prashant D. Sardeshmukh, Nobuki Matsui, Robert J. Allan, Xungang Yin, Byron E. Gleason, Russell S. Vose, Glenn Rutledge, Pierre Bessemoulin, S. Brönnimann, M. Brunet, R. I. Crouthamel, A. N. Grant, P. Y. Groisman, P. D. Jones, M. C. Kruk, A. C. Kruger, G. J. Marshall, M. Maugeri, H. Y. Mok, Ø. Nordli, T. F. Ross, R. M. Trigo, X. L. Wang, S. D. Woodruff, and S. J. Worley. The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, 137(654):1–28, 2011.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in GPT-4V(ision): Bias and interference challenges. *arXiv preprint*, arXiv:2311.03287, 2023.

Shaobo Cui, Zhijing Jin, Bernhard Schölkopf, and Boi Faltings. The odyssey of commonsense causality: From foundational benchmarks to cutting-edge reasoning. *arXiv preprint arXiv:2406.19307*, 2024.

Nikita Dhawan, Leonardo Cotta, Karen Ullrich, Rahul G Krishnan, and Chris J Maddison. End-to-end causal effect estimation from unstructured natural language data. *arXiv preprint arXiv:2407.07018*, 2024.

Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. *arXiv preprint*, arXiv:2312.11001, 2023.

Brett Drury, Hugo Gonçalo Oliveira, and Alneu de Andrade Lopes. A survey of the extraction and applications of causal relations. *Natural Language Engineering*, 28:361 – 400, 2022.

Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25:383–417, 1970.

R. A. Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:1–32, 1921.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. Modeling document-level causal structures for event causal relation identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1808–1817, 2019.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

Moisés Goldszmidt and Judea Pearl. Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions. *KR*, 92:661–672, 1992.

Sander Greenland, Judea Pearl, and James M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999.

Bo Han, Jiangchao Yao, Tongliang Liu, Bo Li, Oluwasanmi Koyejo, and Feng Liu. Trustworthy machine learning: From data to models. *Found. Trends Priv. Secur.*, 7:74–246, 2025. URL `https://api.semanticscholar.org/CorpusID:278211257`.

Norwood Russell Hanson. *Patterns of discovery : an inquiry into the conceptual foundations of science*. Cambridge University Press, 1958.

Nisreen Haydar, Khatoun Alyousef, Usama Alanan, Rana Issa, Fawaz Baddour, Zuheir Al-Shehabi, and Moatasem Hussein Al-Janabi. Role of magnetic resonance imaging (MRI) in grading gliomas comparable with pathology: A cross-sectional study from Syria. *Annals of Medicine and Surgery*, 82:104679, 2022.

Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. doi: 10.1080/01621459.1986.10478354.

Pedram Hosseini, David A Broniatowski, and Mona Diab. Predicting directionality in causal relations in text. *arXiv preprint arXiv:2103.13606*, 2021.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 2008.

Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In *International Conference on Machine Learning*, pages 2901–2910, 2019.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. *Advances in Neural Information Processing Systems*, pages 5549–5561, 2022.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In *Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, 2023.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, pages 859–868, 2019.

Bidec Innovations. Stock price and news related to it, 2023. URL `https://www.kaggle.com/datasets/BidecInnovations/stock-price-and-news-realted-to-it`. Accessed: 2023-10-01.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv preprint*, arXiv:2401.04088, 2024.

Yibo Jiang and Bryon Aragam. Learning nonparametric latent causal graphs with unknown interventions. In *Advances in Neural Information Processing Systems*, volume 36, pages 60468–60513, 2023.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? *arXiv preprint*, arXiv:2306.05836, 2023b.

Elahe Khatibi, Ziyu Wang, and Amir M Rahmani. CDF-RAG: Causal dynamic feedback for adaptive retrieval-augmented generation. *arXiv preprint arXiv:2504.12560*, 2025.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217, 2020.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint*, arXiv:2305.00050, 2023.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding, 2021.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11455–11472. PMLR, 17–23 Jul 2022.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

Sebastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, pages 428–484, 2022.

Andrew Kyle Lampinen, Stephanie C.Y. Chan, Ishita Dasgupta, Andrew Joo Hun Nam, and Jane X Wang. Passive learning of active causal strategies in agents and language models. In *Advances in Neural Information Processing Systems*, 2023.

Junyi Li, Yongqiang Chen, Chenxi Liu, Qianyi Cai, Tongliang Liu, Bo Han, Kun Zhang, and Hui Xiong. Can large language models help experimental design for causal discovery? *arXiv preprint arXiv:2503.01139*, 2025a.

Xin Li, Zhuo Cai, Shoujin Wang, Kun Yu, and Fang Chen. A survey on enhancing causal reasoning ability of large language models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 399–416. Springer, 2025b.

Xiu-Chuan Li, Kun Zhang, and Tongliang Liu. Causal structure recovery with latent variables under milder distributional and graphical assumptions. In *International Conference on Learning Representations*, 2024.

Chenxi Liu and Kun Kuang. Causal structure learning for latent intervened non-stationary data. In *International Conference on Machine Learning*, pages 21756–21777, 2023.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. Discovery of the hidden world with large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 102307–102365. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/b99a07486702417d3b1bd64ec2cf74ad-Paper-Conference.pdf`.

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. Everything has a cause: Leveraging causal inference in legal text analysis. *arXiv preprint arXiv:2104.09420*, 2021.

Yinge Liu, Ninglian Wang, Lingang Wang, Zhongming Guo, and Xiaobo Wu. Variation of cloud amount over China and the relationship with ENSO from 1951 to 2014. *International Journal of Climatology*, 36(8):2931–2941, 2016.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359, 2020.

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. *arXiv preprint*, arXiv:2307.02390, 2023.

Daniel Lyndon, Joseph A Lansley, Jane Evanson, and Anant S Krishnan. Dural masses: meningiomas and their mimics. *Insights into imaging*, 10(1):11, 2019.

Zhiheng Lyu, Zhijing Jin, Rada Mihalcea, Mrinmaya Sachan, and Bernhard Schölkopf. Can large language models distinguish cause from effect? In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

Marloes H Maathuis and Diego Colombo. A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060–1088, 2015.

Daniel Malinsky and Peter Spirtes. Learning the structure of a nonstationary vector autoregression. In *International Conference on Artificial Intelligence and Statistics*, pages 2986–2994, 2019.

John Markoff. Competing as software goes to web. *The New York Times*, 2007. URL `https://www.nytimes.com/2007/06/05/technology/05compute.html`. Accessed: 2024-03-14.

Michael J McPhaden. Genesis and evolution of the 1997-98 El Niño. *Science*, 283(5404): 950–954, 1999.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010. doi: 10.1111/j.1467-9868.2010.00740.x.

MetaAI. Meta Llama 3. `https://ai.meta.com/blog/meta-llama-3/`, 2024. Accessed: 2024-05-20.

Anoop Kumar Mishra. Investigating changes in cloud cover using the long-term record of precipitation extremes. *Meteorological Applications*, 26(1):108–116, 2019.

Francesco Montagna, Atalanti Mastakouri, Elias Eulig, Nicoletta Noceti, Lorenzo Rosasco, Dominik Janzing, Bryon Aragam, and Francesco Locatello. Assumption violations in causal discovery and the robustness of score matching. *Advances in Neural Information Processing Systems*, 36:47339–47378, 2023.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

OpenAI. ChatGPT. `https://chat.openai.com/chat/`, 2022.

OpenAI. GPT-4 technical report, 2023.

OpenAI. Hello GPT-4o! `https://openai.com/index/hello-gpt-4o/`, 2024. Accessed: 2024-05-20.

OpenDataScience. X5 RetailHero. `http://ods.ai/competitions/x5-retailhero-uplift-modeling`, 2019.

Judea Pearl. Graphical models, causality, and intervention. *Statistical Science*, 8(3):266–269, 1993. doi: 10.1214/ss/1177010894.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.

Judea Pearl and Thomas S Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H Maathuis. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018.

Ronan Perry, Julius von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *Advances in Neural Information Processing Systems*, pages 10904–10917, 2022. URL `https://arxiv.org/abs/2206.02013`.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2023.

Nicholas J. Radcliffe. Hillstrom's MineThatData email analytics challenge: An approach using uplift modelling. `https://stochasticsolutions.com/pdf/HillstromChallenge.pdf`, 2008.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58, 1978.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint*, arXiv:2302.04761, 2023.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *arXiv preprint*, arXiv:2102.11107, 2021.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Ilya Shpitser, Tyler VanderWeele, and James M Robins. On the validity of covariate adjustment for estimating causal effects. *arXiv preprint arXiv:1203.3515*, 2012.

Laura C. Slivinski, Gilbert P. Compo, Jeffrey S. Whitaker, Prashant D. Sardeshmukh, Benjamin S. Giese, Chesley McColl, Rob Allan, Xungang Yin, Russell Vose, Holly Titchner, John Kennedy, Lawrence J. Spencer, Linden Ashcroft, Stefan Brönnimann, Manola Brunet, Dario Camuffo, Richard Cornes, Thomas A. Cram, Richard Crouthamel, Fernando Domínguez-Castro, J. Eric Freeman, Joëlle Gergis, Ed Hawkins, Philip D. Jones, Sylvie Jourdain, Alexey Kaplan, Hisayuki Kubota, Frank Le Blancq, Tsz-Cheung Lee, Andrew Lorrey, Jürg Luterbacher, Maurizio Maugeri, Cary J. Mock, G. W. Kent Moore, Rajmund Przybylak, Christa Pudmenzky, Chris Reason, Victoria C. Slonosky, Catherine A. Smith, Birger Tinz, Blair Trewin, Maria Antónia Valente, Xiaolan L. Wang, Clive Wilkinson, Kevin Wood, and Przemysław Wyszyński. Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, 145(724):2876–2908, 2019. doi: 10.1002/qj.3598. URL `https://doi.org/10.1002/qj.3598`.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*, volume 81 of *Lecture Notes in Statistics*. Springer New York, 1993. ISBN 978-1-4612-7650-0. doi: 10.1007/978-1-4612-2748-9.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Uncertainty in Artificial Intelligence*, pages 499–506, 1995.

Peter Spirtes, Clark Glymour, Richard Scheines, and Robert Tillman. *Automated Search for Causal Relations: Theory and Practice*, 2018.

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, pages 32540–32560, 2023.

J. Tarazona and W. Arntz. *The Peruvian Coastal Upwelling System*, pages 229–244. Springer Berlin Heidelberg, 2001.

Panagiotis Tigas, Yashas Annadani, Andrew Jesson, Bernhard Schölkopf, Yarin Gal, and Stefan Bauer. Interventions, where and how? experimental design for causal models at scale. *Advances in Neural Information Processing Systems*, 35:24130–24143, 2022.

Christian Toth, Lars Lorch, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peharz, and Julius von Kügelgen. Active bayesian causal inference. *Advances in Neural Information Processing Systems*, 35:16261–16275, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint*, arXiv:2302.13971, 2023.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *The Journal of Machine Learning Research*, 16(1):2147–2205, 2015.

Ruibo Tu, Kun Zhang, Bo C. Bertilson, Hedvig Kjellström, and Cheng Zhang. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. In *Advances in Neural Information Processing Systems*, pages 12773–12784, 2019.

Birkan Tunç, David A Hormuth II, George Biros, and Thomas E Yankeelov. Modeling of glioma growth with mass effect by longitudinal magnetic resonance imaging. *IEEE Transactions on Biomedical Engineering*, 68(12):3713–3724, 2021. doi: 10.1109/TBME.2021.3085523.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. Causal order: The key to leveraging imperfect experts in causal inference. *arXiv preprint arXiv:2310.15117*, 2023.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, pages 48603–48638, 2023.

Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like DAGs? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4), 2022. ISSN 0360-0300.

Jessica A Wachter. Can time-varying risk of rare disasters explain aggregate stock market volatility? *The Journal of Finance*, 68(3):987–1035, 2013.

Chunzai Wang. ENSO, Atlantic climate variability, and the Walker and Hadley circulations. In *The Hadley circulation: Present, past and future*, pages 173–202. Springer, 2004.

Xinyue Wang, Kun Zhou, Wenyi Wu, Har Simrat Singh, Fang Nan, Songyao Jin, Aryan Philip, Saloni Patnaik, Hou Zhu, Shivam Singh, Parjanya Prashant, Qian Shen, and Biwei Huang. Causal-Copilot: An autonomous causal analysis agent. *arXiv preprint arXiv:2504.13263*, 2025.

J Watts, G Box, A Galvin, P Brotchie, N Trost, and T Sutherland. Magnetic resonance imaging of meningiomas: a pictorial review. *Insights into imaging*, 5:113–122, 2014.

Noah Weber, Rachel Rudinger, and Benjamin Van Durme. Causal inference of script knowledge. *arXiv preprint arXiv:2004.01174*, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality? In *UAI 2022 Workshop on Causal Representation Learning*, 2022. URL https://openreview.net/forum?id=DbJXEqU0kaM.

Anpeng Wu, Haoxuan Li, Kun Kuang, Keli Zhang, and Fei Wu. Learning causal relations from subsampled time series with two time-slices. In *International Conference on Machine Learning*, 2024.

Xingyu Wu, Kui Yu, Jibin Wu, and Kay Chen Tan. LLM cannot discover causality, and should be restricted to non-decisional support in causal discovery. *arXiv preprint arXiv:2506.00844*, 2025.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. The rise and potential of large language model based agents: A survey. *arXiv preprint*, arXiv:2309.07864, 2023.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. In *International Conference on Machine Learning*, pages 5541–5550, 2018.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint*, arXiv:2304.05524, 2023a.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 2024a.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896, 2008a. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2008.08.001. URL https://www.sciencedirect.com/science/article/pii/S0004370208001008.

Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9(7), 2008b.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: a general setting. In *Proceedings of the 41st International Conference on Machine Learning*, pages 60057–60075, 2024b.

Yi Ge Zhang, Mark Pagani, Jorijntje Henderiks, and Haojia Ren. A long history of equatorial deep-water upwelling in the Pacific Ocean. *Earth and Planetary Science Letters*, 467:1–9, 2017. ISSN 0012-821X.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint*, arXiv:2309.01219, 2023b.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 16411–16422. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/6801fa3fd290229efc490ee0cf1c5687-Paper-Conference.pdf`.

Shuang Zhou, Mingquan Lin, Sirui Ding, Jiashuo Wang, Canyu Chen, Genevieve B Melton, James Zou, and Rui Zhang. Explainable differential diagnosis with dual-inference large language models. *npj Health Systems*, 2(1):12, 2025.