Discovery of the Hidden World with Large Language Models

Anonymous Authors[1]

Abstract

Despite progress in the past decades, it has been a major impediment to broader real-world applications of causal methods for relying on high-quality measured variables given by human experts. In this position paper, we argue that the rise of large language models (LLMs) trained on massive observations of the world has great potential to bridge the gap. In particular, we envision a system consisting of two *mutually beneficial* components: LLMs and causal discovery methods (CDs). On the one hand, LLMs that learn rich world knowledge about the world can assist with discovering high-level hidden variables from low-level observational data. CDs, on the other hand, can uncover the underlying causal relations between the discovered high-level variables with guarantees. More importantly, CDs can also provide feedback to improve the identification of the variables. The combination of LLMs and CDs iteratively improves the discovery of the hidden causal world, and hence opens up a broader adoption of various causal methods. To provide justifications for our envision, we also introduce a preliminary implementation of the system **C**ausal representati**O**n **A**ssistan**T** (COAT). We present a comprehensive analysis of COAT with four case studies ranging from analysis of human reviews to diagnosis of neuropathic and brain tumors. The effectiveness of COAT demonstrates a promising direction toward foundation models that empower various downstream applications of causal methods and ultimately the general artificial intelligence.

1. Introduction

Despite the progress in the past decades, existing causal discovery algorithms mainly rely on high-quality measured variables given by human experts (Spirtes et al., 2000; 2010; Vowels et al., 2022). However, the causal variables and their measurements are usually available in a wide range of real-world applications. For example, Amazon sellers who want to analyze the factors related to user ratings only have user reviews, generated by the underlying user preferences for certain product characteristics. Therefore, the lack of measured high-quality causal variables has been a major impediment to broader real-world applications of causal or causality-inspired methods (Schölkopf et al., 2021).

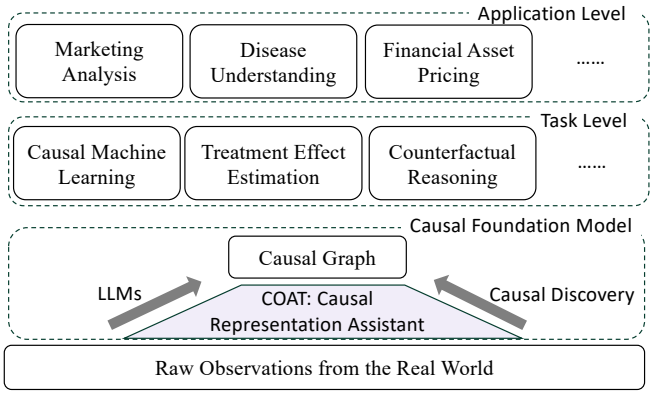


Figure 1. COAT combines both strengths of LLMs that learn the rich knowledge of the world, and causal discovery methods to uncover the hidden causal world. With the uncovered causal knowledge, COAT can empower broader applications of causal methods.

In this position paper, we argue that the recent emergence of Large Language Models (LLMs) (Brown et al., 2020; OpenAI, 2022; Touvron et al., 2023a; OpenAI, 2023) has a great potential to mitigate the gap. In particular, as shown in Fig. 1, we envision an entangled system towards causal foundation models, which consists of two *mutually beneficial* components: LLMs and causal discovery methods (CDs). On the one hand, LLMs that learn rich world knowledge about the world can assist with discovering high-level hidden variables from low-level observational data (Bubeck et al., 2023), or even certain commonsense causal knowledge (Zhang et al., 2023a; Kiciman et al., 2023; Anonymous, 2024). on the other hand, it has also been found that the reliability of LLMs in reasoning of causality 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.,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

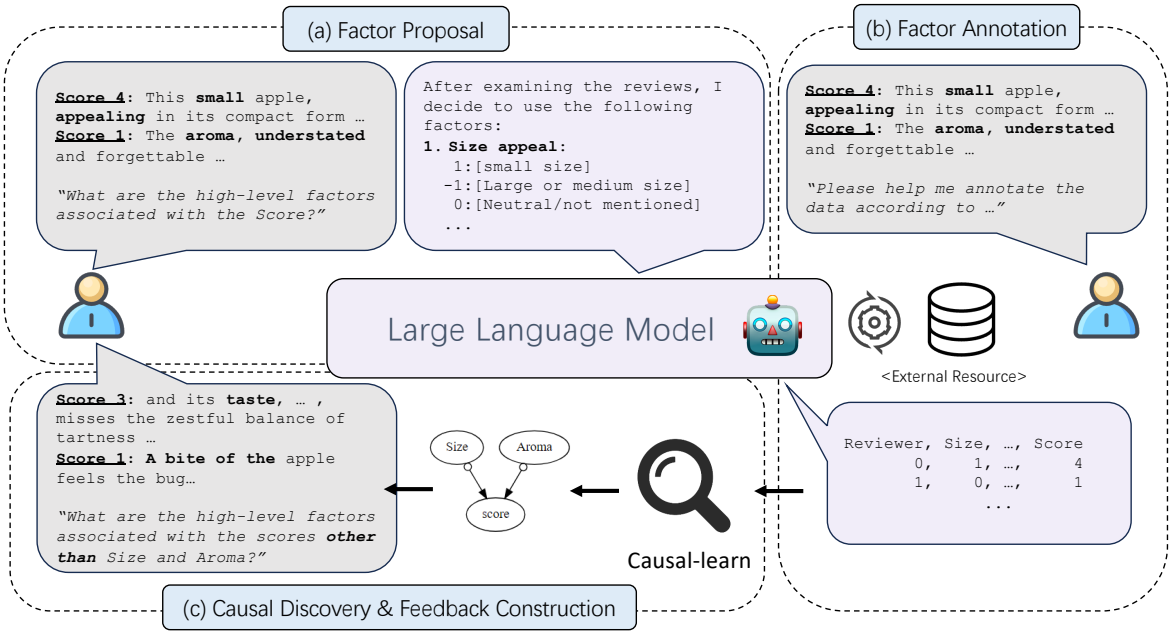


Figure 2. Illustration of COAT framework to analyze the rating scores of AppleGastronome. COAT aims to uncover the underlying Markov Blanket with respect to the given ratings of the apples (i.e., factors that fit the preferences of gastronomes). COAT first (a) adopts an LLM to read, comprehend, and relate the rich knowledge about tasting the 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 associate more related knowledge to uncover more desired causal factors.

2023c; Cui et al., 2023; Berglund et al., 2023), as well as the ethical considerations (Lee et al., 2023; Tu et al., 2023). Nevertheless, CDs can uncover the underlying causal relations between the discovered high-level variables with guarantees. More importantly, CDs can also provide feedback to improve the identification of the variables. The combination of LLMs and CDs iteratively improves the discovery of the hidden causal world, and hence opens up a broader adoption of various causal methods.

To provide justifications for our envision, we also introduce a preliminary implementation of the system **C**ausal representati**O**n **A**ssistan**T** (COAT). Specifically, we will be focusing on discovering the Markov Blanket with respect to a high-value target variable from the unstructured observational data. When given a new task, such as identifying causally related high-level variables about the rating score of an apple from multiple gastronomes, shown as in Fig. 2, COAT first leverages an LLM to relate its acquired knowledge about the task, and comprehend the sample inputs to propose candidate causal factors such as the apple size and smells associated with certain meta-information (e.g., annotation guidelines). Then, COAT can either fetch the data of the candidate factors from external tools or resources when available, or employ another LLM to annotate the raw observational data according to the guidelines. Afterward, the annotated data is readily input to the suitable CD algorithm to obtain the intermediate causal hypothesis. Throughout the paper, we mainly adopt the FCI algorithm (Spirtes et al., 2000) as it allows the existence of latent confounders and is compatible with our objective to find the Markov Blanket.

Nevertheless, LLMs often require proper prompts to unlock their full capability and elicit high-quality outputs (Wei et al., 2022). Meanwhile, LLMs may also give certain unreliable factors or miss necessary factors (Bubeck et al., 2023; Zhang et al., 2023c). Therefore, COAT also leverages the CD algorithm to audit the candidate causal factors and filter out inappropriate factors. For example, one could filter out factors that are independent of the target variable. Meanwhile, COAT can also construct feedback to further improve the factor discovery by LLMs. For example, one could leverage samples where the current factors can not explain the target variable very well, to prompt LLMs to look into the samples and find better factors.

We present a comprehensive analysis of COAT with four case studies ranging from analysis of human reviews to diagnosis of neuropathic and brain tumors (see more details in Appendix B.7). The experiments demonstrate that with COAT, LLMs can effectively identify the underlying causal variables, and provide reliable annotations for CDs

to uncover the causal relations. The strong performances of COAT not only provide preliminary verification of our envision, but more importantly, demonstrate a promising direction toward foundation models that empower various downstream applications of causal methods and ultimately the general artificial intelligence.

2. Related Work

We discuss closely related work in this section and leave more details in Appendix A.

Causal discovery aims to discover the *unknown* causal relations from the observational data (Spirtes et al., 2010; 2000), which is critical to both real-world applications and scientific discoveries (Pearl & Robins, 1999; Pearl & Mackenzie, 2018). 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, and can easily suffer from the low quality of annotated data (e.g., latent confounders) (Dong et al., 2023). In this work, we argue that incorporating LLMs that learn world knowledge from massive training data could effectively relieve the need for artificial causal factor annotation. Meanwhile, COAT also opens up a new line to learn causal representations with rich pre-trained knowledge as well as feedback from causal discovery algorithms (Schölkopf et al., 2021).

Reasoning with LLMs has achieved remarkable performance across a variety of tasks with few demonstrations of the samples (Brown et al., 2020; OpenAI, 2022; Ouyang et al., 2022; OpenAI, 2023). The strong capabilities of LLMs show that it is evident that LLMs could acquire and understand commonsense knowledge about the world (Bubeck et al., 2023). The power of LLMs can be further unlocked with suitable context as inputs (Wei et al., 2022). Nevertheless, LLMs have also been found to make mistakes in basic algorithmic reasoning (Cobbe et al., 2021; Dziri et al., 2023), easy to hallucinate nonfactual results (Zhang et al., 2023c), tend to learn shortcuts or dataset biases (Liu et al., 2023a; Berglund et al., 2023; Cui et al., 2023), and fall short in complex planning and reasoning tasks (Bubeck et al., 2023). The drawbacks of LLMs render it *risky* to rely on the reasoning results to derive any rigorous results. Therefore, we do not directly derive the results from LLMs. Rather, we merely leverage the learned world knowledge in LLMs to find useful causal factors by constructing proper instructions based on causal feedback.

Causal learning with LLMs has received lots of attention results from the community (Kiciman et al., 2023; Zhang et al., 2023a). Kiciman et al. (2023) find that LLMs can recover the pairwise causal relations very well. Lampinen et al. (2023) show that transformer-based agents can learn causal strategies passively if allowed intervention during tests. Choi et al. (2022); Long et al. (2023a); Ban et al. (2023); Anonymous (2024) propose to incorporate the causal discovery results by LLMs as a prior or constraint to improve the performance of data-driven causal discovery algorithms. However, Willig et al. (2022); Zečević et al. (2023) find that LLMs can not understand causality while simply retelling the causal knowledge contained in the training data. Zhang et al. (2023a); Jin et al. (2023b;a) find that LLMs can hardly provide satisfactory answers for discovering new knowledge or decision-making tasks. Although Long et al. (2023b) find that LLMs can build causal graphs with 3-4 nodes, Tu et al. (2023) find that the performance of LLMs in more complex causal discovery remains limited as LLMs can hardly understand new concepts and knowledge. The aforementioned debate implies the limitations in directly adopting the causal discovery results by LLMs, which motivates us to incorporate the existing causal discovery algorithms with rigorous guarantees instead of LLMs to learn the causal relations.

Despite certain successes of indirectly or directly leveraging the LLMs to reason for the causal relations, the unreliability of LLM reasoning poses risks to trusting the causal discovery results. Different from existing approaches, we advocate for another paradigm that merely leverages LLMs to find high-level causal variables. The CD algorithm in COAT can provide necessary guarantees of the discovered causal relations, as well as feedback for maximally eliciting the rich world knowledge of LLMs. We envision that the system consists of two mutually beneficial components that provide a promising direction to develop foundation models empowering broader applications of various causal or causality-based methods.

3. Causal Discovery with LLMs

In this work, we focus on discovering causation from text data while our framework could also be extended to other modalities such as images (Radford et al., 2021; Betker et al., 2023; OpenAI, 2023). Different from typical causal discovery settings with well-structured data (Spirtes et al., 2010), we start from the unstructured empirical observations where the causal variables are not available.

3.1. Problem Definition

We consider uncovering the unobserved latent variables $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_l)$ with respect to an observed high-value target variable y, as well as their causal relations $\boldsymbol{e}$ from the *observed unstructured* data $\boldsymbol{x}$ (e.g., natural languages, images, etc.). The target variable y can be considered as the trigger to query the underlying causation (e.g., an indicator for getting on a disease, etc.). Nevertheless, our framework can also be generalized to discover the whole causal graph.

```
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 3. Illustration of the prompt for factor proposal.

Factors behind unstructured data. We consider a broad spectrum of unstructured data, predominantly text, symbolically represented as $\boldsymbol{x} \in \mathcal{X}$. This data is assumed to be generated by an unknown mapping $g : \cup\{\mathcal{Z}_i\}_{i=1}^{l} \times \mathcal{N} \to \mathcal{X}$, such that $\boldsymbol{x} = g(\boldsymbol{z}, \boldsymbol{\epsilon})$. The latent factors $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_l)$ with $\boldsymbol{z}_i \in \mathcal{Z}_i$ are the underlying causal variables that control the generation of $\boldsymbol{x}$. The randomness is captured by $\boldsymbol{\epsilon} \in \mathcal{N}$, where $\mathcal{N}$ is the noise space.

Observable target variable. The observable target variable is denoted as $y \in \mathbb{R}$, and is generated as $y = f_y(\text{PA}(y), \epsilon_y)$ where $\text{PA}(y) \subseteq \boldsymbol{z}$ refer to the direct parents of y in the causal graph and ϵ_y captures the noise when generating y via f_y. The observable variable y serves as a crucial element in our analysis, which triggers the causal discovery, as well as provides a measurable outcome associated with the data.

Learning latent causal structures. There exists an underlying causal structure $\mathcal{G} = (\boldsymbol{z}, \boldsymbol{e})$ where $\boldsymbol{z}$ is the set of nodes representing the latent factors and $\boldsymbol{e}$ is the set that contains edges linking the nodes. The objective of COAT is, given n i.i.d. observations of $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$,

a) to identify a set of latent factors $\hat{\boldsymbol{z}} \subseteq \boldsymbol{z}$ such that all of the underlying useful factors to predict y are identified $\text{MB}(y) \subseteq \hat{\boldsymbol{z}}$, where $\text{MB}(y)$ refers to the Markov Blanket (Pearl, 1991) with respect to y such that $y \perp\!\!\!\perp \boldsymbol{z}_i \setminus \text{MB}(y) | \text{MB}(y)$;

b) to uncover the causal relations among $\text{MB}(y)$;

The discovery process resembles local causal discovery, which offers meaningful insights about the target variable y (Aliferis et al., 2010; Gupta et al., 2023). Nevertheless, by shifting the target variable to other identified causal factors, we could also recover the whole causal graph.

3.2. The COAT Framework

Under the aforementioned setup, we propose a new framework called COAT that leverages LLMs to propose and annotate factors given the unstructured observation. By additionally incorporating a causal discovery algorithm to audit the proposed factors and provide feedback, COAT could secure reliability when leveraging the learned knowledge by LLMs. Therefore, COAT is proceeded in an iterative manner among the two modules. The algorithm of COAT is given in Algorithm 1, which consists of four major steps.

Factor proposal. Denote an LLM as Ψ that takes k observations $\widehat{\mathcal{D}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{k}$ as inputs. Usually, k is a relatively small number (i.e., $k \leq n$) due to the limited context understanding capability of LLMs (Liu et al., 2023b). Ψ will be instructed by $\boldsymbol{s}^t$ to find potential causal factors using the pre-trained knowledge from large observations of the world:

$$(\hat{\boldsymbol{z}}^t, \hat{\boldsymbol{s}}^t) = \Psi(\widehat{\mathcal{D}}^t, \boldsymbol{s}^t), \tag{1}$$

where $\hat{\boldsymbol{z}} = \{\hat{\boldsymbol{z}}_1, \hat{\boldsymbol{z}}_2, ..., \hat{\boldsymbol{z}}_m\}$ is the set of proposed factors, and $\hat{\boldsymbol{s}} = \{\hat{\boldsymbol{s}}_1, \hat{\boldsymbol{s}}_2, ..., \hat{\boldsymbol{s}}_m\}$ is the set of corresponding semantic descriptions of the factors, such as the guidelines for curating the specific values of the factors. We will use the superscript t to denote the inputs and outputs at the i-th iteration: $\widehat{\mathcal{D}}^t, \boldsymbol{s}^t, \hat{\boldsymbol{z}}^t, \hat{\boldsymbol{s}}^t$. The factor proposal stage aims to imitate the process of human experts in selecting and defining causal variables (Pearl & Mackenzie, 2018). $\widehat{\mathcal{D}}$ will be grouped according to values of y, and then Ψ will be instructed to explain the potential factors associated with the changes in y, shown as in Fig. 3. In addition, the metadata about the task such as the task description and context can also be incorporated into $\boldsymbol{s}$ if available.

Factor parsing. Once with the candidate factors, we will then collect the values of the factors from the unstructured observations. In previous works, they are usually collected from human experts according to the given factors (Spirtes et al., 2000). When without any additional information, in the factor proposal stage, Ψ will also need to give the corresponding value curation guidelines in $\hat{\boldsymbol{s}}$. Then, another LLM Ψ_s could be leveraged to read the guidelines and parse the unstructured observations into structured data:

$$\hat{\boldsymbol{z}}^t = \Psi_s(\mathcal{D}, \hat{\boldsymbol{v}}^t, \hat{\boldsymbol{s}}^t, \boldsymbol{s}_p), \tag{2}$$

where $\boldsymbol{s}_p$ refers to the additional instruction to prompt Ψ_s to parse the observed data, and $\hat{\boldsymbol{v}} = \{\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2, ..., \hat{\boldsymbol{v}}_n\}$ refers to the parsed values for the corresponding factors in $\hat{\boldsymbol{z}}$.

Algorithm 1 The COAT Algorithm

Required: Dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$; LLM for factor proposal Ψ; Model for factor parsing Ψ_s; causal discovery algorithm $\mathcal{A}$; Feedback constructor $\mathcal{F}$; Maximal rounds T;
Random sampling $\widehat{\mathcal{D}}^0$;
Constructing $\boldsymbol{s}^0$;
while not converge and current round $t < T$ **do**
  $(\hat{\boldsymbol{z}}^t, \hat{\boldsymbol{s}}^t) \leftarrow \Psi(\widehat{\mathcal{D}}^t, \boldsymbol{s}^t)$; `//factor proposal`
  $\hat{\boldsymbol{v}}^t \leftarrow \Psi_s(\mathcal{D}, \hat{\boldsymbol{z}}^t, \hat{\boldsymbol{s}}^t, \boldsymbol{s}_p^t)$; `//factor parsing`
  $\widehat{\mathcal{G}}^t \leftarrow \mathcal{A}(\hat{\boldsymbol{v}}^t \cup \{y\})$; `//causal discovery`
  $(\widehat{\mathcal{D}}^{t+1}, \boldsymbol{s}^{t+1}) \leftarrow \mathcal{F}(\widehat{\mathcal{G}}^t, \mathcal{D}, \boldsymbol{s}^t)$; `//feedback`
end while
return $\mathcal{G}^T$

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 obtain $\hat{\boldsymbol{z}}$ 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). In experiments, we examine both strategies and show that COAT can effectively uncover the hidden factors under both factor parsing schemes.

Causal discovery. With the given values $\hat{\boldsymbol{v}} \cup \{y\}$ associated with the candidate factors $\hat{\boldsymbol{v}} \cup \{y\}$, a causal discovery algorithm $\mathcal{A}$ can be allocated to reason about the causal hypothesis from the parsed data:

$$\widehat{\mathcal{G}}^t = \mathcal{A}(\hat{\boldsymbol{z}}^t \cup \{y\}), \tag{3}$$

where $\widehat{\mathcal{G}}^t = (\hat{\boldsymbol{z}}^t, \hat{\boldsymbol{e}}^t)$ is the discovered causal hypothesis. In general, the inputs in each round may contain noises as well as the latent confounders, any causal discovery algorithms with suitable theoretical assumptions could be used for $\mathcal{A}$. The noises injected through LLM-based parsing may be of independent interest to the literature of causal discovery (Glymour et al., 2019; Vowels et al., 2022).

In this work, we demonstrate the idea of COAT via the FCI algorithm (Spirtes et al., 2010) as it is flexible to the functional classes of the underlying generation process, allows for the existence of latent confounders, and aligns well with our objective of discovering $\mathrm{MB}(y)$.

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; Bubeck et al., 2023). In complex tasks, it usually requires a decomposition into intermediate steps (Wei et al., 2022). 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:

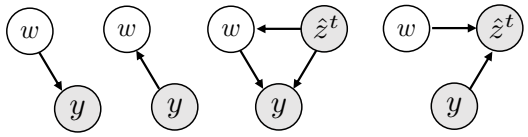


Figure 4. Illustration of variables that could be discovered with COAT. When w is the direct parent or child of y, finding hard-to-explain samples can help uncover it, with a sufficiently powerful LLM. When w is the direct parent and also a child of the discovered variable $\hat{\boldsymbol{z}}^t$, or the spouse of y with $\hat{\boldsymbol{z}}^t$ as the common child as y, conditioning on $\hat{\boldsymbol{z}}^t$ facilitates the discovery of w.

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

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

3.3. Causal Feedback

After setting up COAT framework, let us elaborate more on the causal feedback used in our demonstration. Since we are interested in discovering all of the relevant factors of the target variable y, let us assume there exists some underlying latent variable $w \in \mathrm{MB}(y)$, but not discovered (i.e., $w \notin \hat{\boldsymbol{z}}^t$) at round t. Then, we know w is not conditional independent with y given $\hat{\boldsymbol{z}}^t$ by the property of $\mathrm{MB}(y)$ (Pearl, 1991) (Note that $H(\cdot)$ is the notation for entropy):

$$H(y|\hat{\boldsymbol{z}}^t, w) < H(y|\hat{\boldsymbol{z}}^t), \tag{5}$$

which also means incorporating w into the causal factors could facilitate the explanation and predictions of y. If the observational data contains sufficiently diverse examples, and the LLM is sufficiently powerful, Eq. 5 can help with progressively expanding the discovered factors and uncovering all the useful factors in $\mathrm{MB}(y)$, as shown in Fig. 4.

In general, LLMs have been demonstrated to be able to resolve a variety of tasks when given a few samples (Brown et al., 2020; Bubeck et al., 2023). When given the current observations $\widehat{\mathcal{D}}^t$, Ψ is expected to find useful $\hat{w}$ such that

$$H_{\widehat{\mathcal{D}}^t}(y|\hat{\boldsymbol{z}}^t) - H_{\widehat{\mathcal{D}}^t}(y|\hat{\boldsymbol{z}}^t, \hat{w}) \geq \Delta H_\Psi, \tag{6}$$

where $H_{\widehat{\mathcal{D}}^t}(y|\hat{\boldsymbol{z}}^t)$ refers to the conditional entropy measured on $\widehat{\mathcal{D}}^t$, and ΔH_Ψ represents the sensitivity of the LLM Ψ in identifying the associated factors. Intuitively, more powerful LLMs tend to have a lower ΔH_Ψ. For the samples where Ψ can not find the desired w, it means $\hat{\boldsymbol{z}}^t$ is sufficient to capture the variances in y. To this end, we need to select suitable $\widehat{\mathcal{D}}^*$ such that

$$\widehat{\mathcal{D}}^* = \arg\max_{\widehat{\mathcal{D}} \subseteq \mathcal{D}} H_{\widehat{\mathcal{D}}}(y|\hat{\boldsymbol{z}}^t), \tag{7}$$

which means the currently discovered causal variables $\hat{\boldsymbol{z}}^t$ can not well explain the target variable y. Recall that one of the key metric for measuring the quality of discovered Markov Blanket variables is the predictivity (Pearl, 1991). Therefore, Eq. 7 can be solved by converting it into a classification problem, where $\widehat{\mathcal{D}}^*$ are the samples that the fitted classification model yields a large prediction error. In our experiments, we implement the classification as a clustering with respect to $\hat{\boldsymbol{z}}^t$ to perform sample selection of $\widehat{\mathcal{D}}^*$. The clustering elicits c groups $\{\widehat{\mathcal{D}}_i\}_{i=1}^c$ associated with $\{\hat{\boldsymbol{z}}_i^t\}_{i=1}^c$ in the i-th group:[1]

$$\{\widehat{\mathcal{D}}_i\}_{i=1}^c, \{\hat{\boldsymbol{z}}_i^t\}_{i=1}^c = \text{K-Means}(\mathcal{D}, \hat{\boldsymbol{z}}^t). \tag{8}$$

With the groups obtained, we take the group with the largest conditional entropy to construct the feedback. The entropy of i-th group is estimated by $\hat{\boldsymbol{h}}_i = -\sum \hat{p}_i(Y = y) \log \hat{p}_i(Y = y)$, where the summation is on the set of Y's values occurred in the i-th cluster, and $\hat{p}_i(Y = y)$ is the estimated probability mass function evaluated in the $\widehat{\mathcal{D}}_i$.

3.4. Practical Discussions

After establishing the main framework of COAT, we then discuss some cases where we need to handle them properly in practice, due to the instability and variety of LLM outputs.

Factor filtering. In some cases, LLMs may output several factors that have similar semantics, or even 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 get rid of these variables.

Factor pool. In some cases, LLMs may discover useful factors in early rounds while may be discarded. For example, the underlying spouse variables of the target label y may be independent with 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.

4. Analysis And Discussions of Feedback Driven Causal Discovery in COAT

The previous discussion leaves many aspects mysterious that the success of COAT in finding useful causal information still largely depends on the capability of LLMs. In other words, it remains unknown to what extent LLMs can find useful factors and parse the data properly, in order to find the underlying Markov Blanket of the target variable.

[1]For the sake of clarity, here we denote the empirical values of $\hat{\boldsymbol{z}}^t$ within each group as $\hat{\boldsymbol{z}}_i{}^t$, with a little abuse of notation.

Table 1. Causal discovery results in AppleGastronome. MB, NMB and OT refer to the number of causal factors discovered in the underlying Markov Blanket, in the causal graph but not the Markov Blanket, and the other variables. Recall, precision, and F1 for factor proposal evaluate the discovered causal ancestors.

LLM	METHOD	FACTOR PROPOSAL					
		MB	NMB	OT	RECALL	PRECISION	F1
GPT-4	META	4	1	2	0.67	0.29	0.40
	DATA	4	1	0	1.00	0.60	0.75
	COAT	3	0	0	1.00	1.00	1.00
GPT-3.5	META	5	1	4	1.00	0.30	0.46
	DATA	5	1	0	1.00	0.50	0.67
	COAT	5	0	0	1.00	0.75	0.86
LLAMA2	META	3	0	5	0.67	0.40	0.50
	DATA	2	1	0	0.67	0.50	0.57
	COAT	2	1	0	0.67	0.80	0.73
MISTRAL	META	3	1	0	0.67	0.50	0.57
	DATA	3	1	0	1.00	0.75	0.86
	COAT	3	1	0	1.00	0.75	0.86

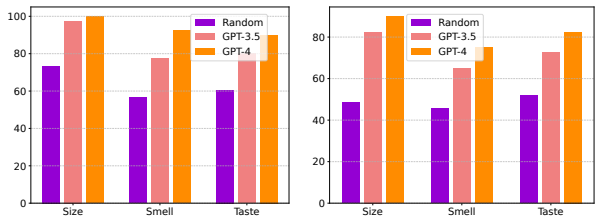
Thus, in this section, we construct the first benchmark under our setting, called AppleGastronome for causal discovery from unstructured data. We examine the capabilities of the predominant LLMs such as GPT-3.5-turbo (OpenAI, 2022), GPT-4-turbo (OpenAI, 2023), LLaMA2-chat (Touvron et al., 2023b), as well as Mistral-Medium (Jiang et al., 2024) in realizing COAT. For LLaMA2, we mainly evaluate the 70b variant, as we find the 13b and 7b variants can not follow the instructions well in AppleGastronome.

4.1. The AppleGastronome Benchmark

Benchmark construction. In 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 and 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. More details about the construction of the AppleGastronome are left in Appendix B.1. Finally, we have the AppleGastronome benchmark $\mathcal{D}$ with $(\boldsymbol{x}_i, y_i)$ as the review and the score of the i-th gastronome, respectively. We aim to leverage COAT to identify the underlying factors that cause the rating score.

Evaluation. In AppleGastronome benchmark, since we do not have an external tool to fetch the data of the proposed factors, we employ two LLMs Ψ and Ψ_s to propose the useful factors along with the annotation criteria, and parse the unstructured review data into the structured tabular data, respectively. We evaluate both factor proposal and recovery of the underlying causal graph based on the annotated data.

Factor proposal baselines. Since most of the existing causal discovery approaches start with structured data, we



(a) Apple attributes Acc. (b) Preference matchness Acc.

Figure 5. Annotation accuracies of GPT-4 and GPT-3.5 for apple attributes and preference matchness in AppleGastronome.

mainly compared COAT with the alternative use of using LLMs (Kiciman et al., 2023). Specifically, for the comparison of causal factor proposal, we mainly employ two different uses of LLMs, i.e., **Meta** that uses LLMs to directly output the factors given the context; **Data** that provides the same examples as the first round in COAT to the LLM.

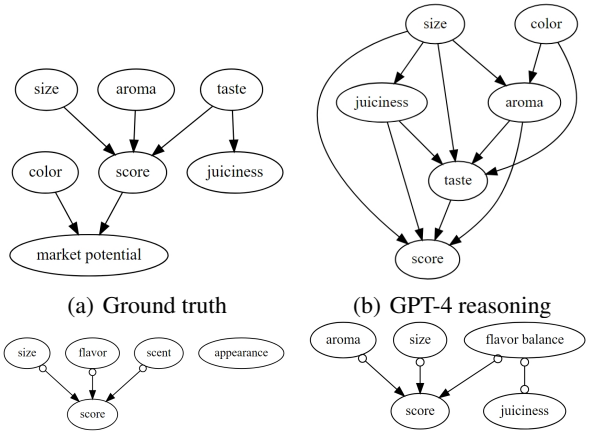
Causal relation inference baselines. For the causal relation reasoning, we follow the previous works Kiciman et al. (2023) that prompts LLMs to reason for the causal direction of each pair of the discovered variables by **Data** baseline.

4.2. Analysis with AppleGastronome Benchmark

Can LLMs be an effective factor proposer? The causal discovery results are given in Table 3. It can be found that, compared to other uses of LLMs, COAT obtain significant improvements regardless of which LLM is used. In contrast, directly using LLMs to reason about the causal relations results in a high sensitivity to the capabilities of LLMs. When incorporating a relatively weak backbone such as LLaMA2, both **Meta** and **Data** can not find too much meaningful information than that with a stronger LLM backbone such as GPT-3.5 or GPT-4. Interestingly, when fed the examples to LLMs, all methods can generically obtain improvements. Therefore, we will directly perform the causal relation inference based on the factors proposed by **Data**.

Can LLMs be an effective factor annotator? Moreover, since LLMs are also used to annotate the data according to the annotation guidelines proposed by Ψ, we analyze the capabilities of LLMs in terms of annotation accuracy and whether the annotation will bring additional confounders that hinder the causal discovery. We mainly evaluate two LLMs GPT-3.5 and GPT-4. Fig. 5 shows the annotation accuracies of GPT-3.5 and GPT-4 when given proper guidelines. It can be found that both LLMs are generally good at annotating subjective attributes. When it comes to objective human preferences, the performance of GPT-3.5 will decrease while being relatively high

Will LLMs introduce additional confounders in annotating factors? In addition, since the annotated results by



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

(c) GPT-3.5 COAT (d) GPT-4 COAT

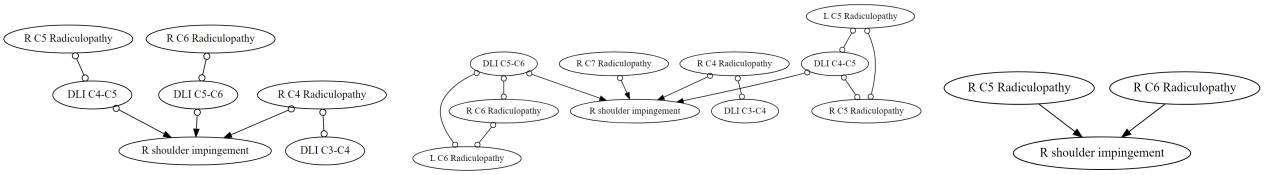
Figure 6. The discovered causal graphs in AppleGastronome. Compared to the ground truth results, directly adopting LLMs to reason about the causal relations can easily elicit lots of false positive causal relations. In contrast, the relations recovered by COAT have a high precision as well as the recall. The directed edge between "taste" and "juiciness" can not be recovered by COAT is because of the limitations of FCI.

LLMs will involve additional noises, or even additional confounders, we also conduct independence tests among the annotation noises and the features. Table 4 presents the independence testing results between the annotation noises and the annotated features, and the noises themselves. It can be found that, the introduced noises are independent of the attributes, therefore, will not introduce much additional interference to the causal discovery procedure.

Can COAT reliably recover the causal relationships? We present quantitative and qualitative results in Table 3 and Fig. 6, respectively. Compared to directly adopting LLMs to reason the causal relations, COAT significantly boosts the causal relation recovery. When considering the F1 score, even the weakest LLM LLaMA2 can outperform all the other LLMs when incorporated into COAT. GPT-3.5 is also boosted by COAT to find better causal relations. The results serve as strong evidence for the effectiveness of COAT.

5. Empirical Study with Realistic Benchmarks

After examining the capabilities of COAT in AppleGastronome, we are further motivated to challenge COAT in a more complex setting from neuronpathic panic diagnosis (Tu et al., 2019). Tu et al. (2023) have conducted tests with GPT-3.5 on the dataset and find the performance of direct reasoning with GPT-3.5 is rather limited. In this study, we construct a more realistic and challenging benchmark based on the dataset.



(a) Faithful ground truth (b) GPT-4 COAT (c) GPT-4 reasoning

Figure 7. The discovered causal graphs in Neuropathic.

5.1. The Neuropathic Benchmark

Benchmark construction. In the Neuropathic benchmark, we convert the dataset into a clinical diagnosis task. In the original dataset, 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. More details about the construction of the Neuropathic are left in Appendix B.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 dataset (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.

Table 2. 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.

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
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
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
LLAMA2-13B	META	1	3	6	0.88	0.40
	DATA	3	6	4	0.95	0.75
	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
	COAT	3	6	2	0.97	0.86
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

5.2. Empirical Results on Neuropathic Benchmark

Factor proposal. The quantitive results on Neuropathic benchmark are given in Table 2. Similarly, we can find that COAT consistently outperforms all of the baselines regardless of which LLMs are incorporated. In particular, COAT can boost the weakest backbone LLaMA2-7b to be better than any other LLMs.

Causal relation recovery. Fig. 7(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 Fig. 7, when using LLMs to perform the reasoning, LLMs cannot identify the faithfulness issues. In contrast, COAT can imply faithful causal insights.

6. Conclusions

In this position paper, we advocate a new paradigm to incorporate LLMs towards building a causal foundation model for broader applications of various causal methods. In particular, we argue that LLMs and CDs can be mutually beneficial to mitigate the reliance on high-quality human-annotated causal variables for causal discovery. We presented a new framework called COAT to verify our vision in four challenging real-world case studies.

Looking forward, we envision future studies can be focused on two parts: (i) From the theory sides, how can we extend the previous identifiability theories to accommodate the inputs of LLMs? (ii) From the methodology side, can we design more useful feedback from CDs, and build a more efficient COAT to identify the whole causal graphs? With more sophisticated COAT implementations in the future, the discovered causal graphs will be a foundation for all causal or causality-inspired methods, that empower the exploration of the world and better living of humankind with general artificial intelligence.

Impact Statements

This work focuses on fully leveraging the rich knowledge learned by LLMs during pre-training to facilitate causal discovery from unstructured data, with the hope of empowering broader applications and social benefits. Besides, this paper does not raise any ethical concerns. This study does not involve any human subjects, practices to data set releases, potentially harmful insights, methodologies and applications, potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity issues.

References

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(7):171–234, 2010. (Cited on page 4)

Anonymous. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2024. (Cited on pages 1, 3 and 12)

Ban, T., Chen, L., Wang, X., and Chen, H. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint*, arXiv:2306.16902, 2023. (Cited on pages 3 and 12)

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *arXiv preprint*, arXiv:2309.12288, 2023. (Cited on pages 2, 3 and 12)

Betker, J., Goh, G., Jing, L., TimBrooks, Wang, J., Li, L., LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and Ramesh, A. Improving image generation with better captions. 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf. (Cited on page 3)

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. (Cited on pages 1, 3, 5 and 12)

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S. M., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint*, arXiv:2303.12712, 2023. (Cited on pages 1, 2, 3, 5 and 12)

Choi, K., Cundy, C., Srivastava, S., and Ermon, S. Lmpriors: Pre-trained language models as task-specific priors. *arXiv preprint*, arXiv:2210.12530, 2022. (Cited on pages 3 and 12)

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint*, arXiv:2110.14168, 2021. (Cited on pages 3 and 12)

Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., and Yao, H. Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges. *arXiv preprint*, arXiv:2311.03287, 2023. (Cited on pages 2, 3 and 12)

Dong, X., Huang, B., Ng, I., Song, X., Zheng, Y., Jin, S., Legaspi, R., Spirtes, P., and Zhang, K. A versatile causal discovery framework to allow causally-related hidden variables. *arXiv preprint*, arXiv:2312.11001, 2023. (Cited on pages 3 and 12)

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., Welleck, S., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., Sanyal, S., Ren, X., Ettinger, A., Harchaoui, Z., and Choi, Y. Faith and fate: Limits of transformers on compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. (Cited on pages 3 and 12)

Glymour, C., Zhang, K., and Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019. (Cited on pages 3, 5 and 12)

Gupta, S., Childers, D., and Lipton, Z. C. Local causal discovery for estimating causal effects. In *Conference on Causal Learning and Reasoning*, volume 213, pp. 408–447, 2023. (Cited on page 4)

Huang, J., Gu, S., Hou, L., Wu, Y., Wang, X., Yu, H., and Han, J. Large language models can self-improve. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1051–1068, 2023. (Cited on page 5)

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de Las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet,

T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts. *arXiv preprint*, arXiv:2401.04088, 2024. (Cited on page 6)

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., LYU, Z., Blin, K., Adauto, F. G., Kleiman-Weiner, M., Sachan, M., and Schölkopf, B. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. (Cited on pages 1, 3 and 12)

Jin, Z., Liu, J., Lyu, Z., Poff, S., Sachan, M., Mihalcea, R., Diab, M. T., and Schölkopf, B. Can large language models infer causation from correlation? *arXiv preprint*, arXiv:2306.05836, 2023b. (Cited on pages 1, 3 and 12)

Kiciman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint*, arXiv:2305.00050, 2023. (Cited on pages 1, 3, 7 and 12)

Lampinen, A. K., Chan, S. C., Dasgupta, I., Nam, A. J. H., and Wang, J. X. Passive learning of active causal strategies in agents and language models. In *Advances in Neural Information Processing Systems*, 2023. (Cited on pages 3 and 12)

Lee, P., Bubeck, S., and Petro, J. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239, 2023. (Cited on page 2)

Liu, B., Ash, J. T., Goel, S., Krishnamurthy, A., and Zhang, C. Transformers learn shortcuts to automata. In *The Eleventh International Conference on Learning Representations*, 2023a. (Cited on pages 3 and 12)

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *arXiv preprint*, arXiv:2307.03172, 2023b. (Cited on page 4)

Long, S., Piché, A., Zantedeschi, V., Schuster, T., and Drouin, A. Causal discovery with language models as imperfect experts. *arXiv preprint*, arXiv:2307.02390, 2023a. (Cited on pages 3 and 12)

Long, S., Schuster, T., and Piché, A. Can large language models build causal graphs? *arXiv preprint*, arXiv:2303.05279, 2023b. (Cited on pages 3 and 12)

LYU, Z., Jin, Z., Mihalcea, R., Sachan, M., and Schölkopf, B. Can large language models distinguish cause from effect? In *UAI 2022 Workshop on Causal Representation Learning*, 2022. (Cited on page 12)

OpenAI. Chatgpt. https://chat.openai.com/chat/, 2022. (Cited on pages 1, 3, 6 and 12)

OpenAI. Gpt-4 technical report, 2023. (Cited on pages 1, 3, 6 and 12)

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. (Cited on pages 3 and 12)

Pearl, J. Probabilistic reasoning in intelligent systems - networks of plausible inference. In *Morgan Kaufmann series in representation and reasoning*, 1991. (Cited on pages 4, 5 and 6)

Pearl, J. and Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X. (Cited on pages 3, 4 and 12)

Pearl, J. and Robins, J. M. Causal diagrams for epidemiologic research. *Epidemiology*, 10 1:37–48, 1999. (Cited on pages 3 and 12)

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., and Chen, H. Reasoning with language model prompting: A survey. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2023. (Cited on page 5)

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 8748–8763, 2021. (Cited on page 3)

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. *arXiv preprint*, arXiv:2302.04761, 2023. (Cited on page 5)

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Towards causal representation learning. *arXiv preprint*, arXiv:2102.11107, 2021. (Cited on pages 1, 3 and 12)

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press, 2000. ISBN 978-0-262-19440-2. (Cited on pages 1, 2, 3, 4 and 12)

Spirtes, P., Glymour, C., Scheines, R., and Tillman, R. Automated Search for Causal Relations: Theory and Practice. 2010. URL https://kilthub.cmu.edu/articles/journal_contribution/Automated_

Search_for_Causal_Relations_Theory_and_Practice/6490961. (Cited on pages 1, 3, 5 and 12)

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models. *arXiv preprint*, arXiv:2302.13971, 2023a. (Cited on page 1)

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, arXiv:2307.09288, 2023b. (Cited on page 6)

Tu, R., Zhang, K., Bertilson, B. C., Kjellström, H., and Zhang, C. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. In *Advances in Neural Information Processing Systems*, pp. 12773–12784, 2019. (Cited on pages 5, 7 and 8)

Tu, R., Ma, C., and Zhang, C. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis. *arXiv preprint*, arXiv:2301.13819, 2023. (Cited on pages 2, 3, 7 and 12)

Vowels, M. J., Camgoz, N. C., and Bowden, R. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Survey*, 55(4), 2022. ISSN 0360-0300. (Cited on pages 1, 3, 5 and 12)

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. (Cited on pages 2, 3, 5 and 12)

Willig, M., Zečević, M., Dhami, D. S., and Kersting, K. Can foundation models talk causality? In *UAI 2022 Workshop on Causal Representation Learning*, 2022. (Cited on pages 3 and 12)

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huan, X., and Gui, T. The rise and potential of large language model based agents: A survey. *arXiv preprint*, arXiv:2309.07864, 2023. (Cited on page 5)

Zečević, M., Willig, M., Dhami, D. S., and Kersting, K. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. (Cited on pages 1, 3 and 12)

Zhang, C., Bauer, S., Bennett, P., Gao, J., Gong, W., Hilmkil, A., Jennings, J., Ma, C., Minka, T., Pawlowski, N., and Vaughan, J. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint*, arXiv:2304.05524, 2023a. (Cited on pages 1, 3 and 12)

Zhang, Y., Fitzgibbon, B., Garofolo, D., Kota, A., Papenhausen, E., and Mueller, K. An explainable AI approach to large language model assisted causal model auditing and development. *arXiv preprint*, arXiv:2312.16211, 2023b. (Cited on page 12)

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., and Shi, S. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint*, arXiv:2309.01219, 2023c. (Cited on pages 1, 2, 3 and 12)

A. More Related Work

Causal discovery aims to discover the *unknown* causal relations from the observational data (Spirtes et al., 2010; 2000), which is critical to both real-world applications and scientific discoveries (Pearl & Robins, 1999; Pearl & Mackenzie, 2018). 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, and can easily suffer from the low quality of annotated data (e.g., latent confounders) (Dong et al., 2023). In this work, we show that incorporating LLMs that learn world knowledge from massive training data could effectively relieve the need for artificial causal factor annotation. Meanwhile, COAT also opens up a new line to learn causal representations with rich pre-trained knowledge as well as feedback from causal discovery algorithms (Schölkopf et al., 2021),

Reasoning with LLMs has achieved remarkable performance across a variety of tasks with few demonstrations of the samples (Brown et al., 2020; OpenAI, 2022; Ouyang et al., 2022; OpenAI, 2023). The strong capabilities of LLMs show that it is evident that LLMs could acquire and understand commonsense knowledge about the world (Bubeck et al., 2023). The power of LLMs can be further unlocked with suitable context as inputs (Wei et al., 2022). Nevertheless, LLMs have also been found to make mistakes in basic algorithmic reasoning (Cobbe et al., 2021; Dziri et al., 2023), easy to hallucinate nonfactual results (Zhang et al., 2023c), tend to learn shortcuts or dataset biases (Liu et al., 2023a; Berglund et al., 2023; Cui et al., 2023), and fall short in complex planning and reasoning tasks (Bubeck et al., 2023). The drawbacks of LLMs render it *risky* to rely on the reasoning results to derive any rigorous results. Therefore, we do not directly derive the results from LLMs. Rather, we merely leverage the learned world knowledge in LLMs to find useful causal factors by constructing proper instructions based on causal feedback.

Causal learning with LLMs has received lots of attention results from the community (Kiciman et al., 2023; Zhang et al., 2023a). Kiciman et al. (2023) find that LLMs can recover the pairwise causal relations very well. Lampinen et al. (2023) show that transformer-based agents can learn causal strategies passively if allowed intervention during tests. Choi et al. (2022); Long et al. (2023a); Ban et al. (2023); Anonymous (2024) propose to incorporate the causal discovery results by LLMs as a prior or constraint to improve the performance of data-driven causal discovery algorithms. However, Willig et al. (2022); Zečević et al. (2023) find that LLMs can not understand causality while simply retelling the causal knowledge contained in the training data. Zhang et al. (2023a); Jin et al. (2023b;a) find that LLMs can hardly provide satisfactory answers for discovering new knowledge or decision-making tasks. Although Long et al. (2023b) find that LLMs can build causal graphs with 3-4 nodes, Tu et al. (2023) find that the performance of LLMs in more complex causal discovery remains limited as LLMs can hardly understand new concepts and knowledge. The aforementioned debate implies the limitations in directly adopting the causal discovery results by LLMs, which motivates us to incorporate the existing causal discovery algorithms with rigorous guarantees instead of LLMs to learn the causal relations.

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. Zhang et al. (2023b) propose to leverage LLMs to audit causal discovery results in an explainable way.

The closest works to ours are Choi et al. (2022); Long et al. (2023a); Ban et al. (2023); Anonymous (2024) which also incorporates LLMs into the pipeline of causal discovery. Nevertheless, all of the existing combinations of LLMs and causal discovery still focus on artificially curated structured data and rely on the capability of LLMs to infer causal relations, therefore, limited in both the reliability and the utility of LLMs in causal learning.

B. More Details about Experiments

B.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 and 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 Fig. 8. The additional results on Relation Extraction are given in Table. 3.

Table 3. Causal relation extraction results in AppleGastronome.

LLM	METHOD	RELATION EXTRACTION		
		RECALL	PRECISION	F1
GPT-4	PAIR-WISE	0.75	0.27	0.40
	COAT	1.00	1.00	1.00
GPT-3.5	PAIR-WISE	0.80	0.40	0.53
	COAT	1.00	0.75	0.86
LLAMA2	PAIR-WISE	0.33	0.13	0.18
	COAT	0.67	0.80	0.73
MISTRAL	META	0.75	0.50	0.60
	COAT	1.00	1.00	1.00

Table 4. Independence tests of the annotation noises with annotated features and other noises AppleGastronome.

LLM	TEST OBJECT	T	P-VALUE
GPT-4	FEATURE	0.4803	0.0325
	FEATURE	0.2828	0.9997
GPT-3.5	NOISE	0.3030	0.1594
	NOISE	0.06181	0.9745

Examples of AppleGastronome are given in Fig. 9.

B.2. More Details on Prompts for AppleGastronome

The prompts for factor proposal are given in Fig. 10.

The prompts for factor annotation are given in Fig. 11,Fig. 12,Fig. 13.

The prompts for constructing feedback are given in Fig. 14.

B.3. More Details of Results on AppleGastronome

The detailed causal graph results are given from Fig. 15 to Fig. 19.

B.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" hroughout all experiments. Other parameters are kept as the default.

B.5. More Details on Constructing Neuropathic

In the Neuropathic benchmark, we convert the dataset into a clinical diagnosis task. In the original dataset, 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 Fig. 20.

Examples of Neuropathic are given in Fig. 21.

B.6. More Details of Results on Neuropathic

The detailed causal graph results are given from Fig. 22 to Fig. 26.

```
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:
- vibrant and uniformly colored
- 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 8. Illustration of prompts for generating AppleGastronome.

B.7. Additional Real-World Case Studies

Brain Tumor (kaggle/brain-tumor-classification-mri). As shown in Fig. 27: We use gpt-4-vision-preview to handle image samples. Each sample is a scanning MRI of a human brain. The interesting variable is the tumor type. Two visual factors appeared in the final causal graph. The proposed two factors, the contrast enhancement (brightness w.r.t. background) and the mass effect (pushing or displacing the surrounding brain tissue), are results of tumors' activities; and therefore can be used to determine the tumor type. These factors can be verified by the medical literature.

News & Stock (kaggle/stock-price-and-news-realted-to-it). As shown in Fig. 28: We analyze the potential factors behind the news and their relation to future return rates (using data before 2007-09). Each sample is the news about Microsoft at a certain date. The interesting variable is the stock's future return rate. Factor values are post-processed by moving averaging and rolling standardization with fixed time windows. The "innovation and technology focus" is identified as one potential causal factor of future return rates for the Microsoft stock and it yields the highest sharp ratio in the out-of-sample trading test using data after 2007-09.

Score: -4
Unfortunately, this batch of apples fails to meet quality standards. Their small size, extensive surface blemishes, and dry texture fall short of expectations. The off-putting musty scent suggests potential rot, and the sour taste overpowers any sweetness. Suppliers should beware; these apples could result in financial disappointment if offered to discerning customers.

This particular apple, while boasting a striking and consistent hue, falls short of expectations due to its unappealing small stature, an off-putting musty odor suggesting decay, and a flavor profile that tilts disappointingly toward sourness. The lack of juiciness further detracts from its desirability, making it a risky choice for suppliers and a potential loss-inducer in the market.

Score: -3
This apple, regrettably, fails to deliver on multiple fronts, its diminutive stature only overshadowed by an off-putting musty undertone that hints at decay. The fruit's unbalanced profile skewing toward sourness compounded by a disappointing dryness ensures a less than palatable experience. Given these attributes, its market performance is predictably underwhelming, posing a significant risk for monetary losses.

Figure 9. Illustration of examples in AppleGastronome.

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 10. 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
approciate facotr value based on the
criterion?

{factor_state}


**Your final output should follow this
template**:

**"The value is: ___."**
```

Figure 11. Illustration of the prompt for factor annotation (part 1).

```
## Output

### Part 1. analysis

In this part, feel free to write down
the process of your considerations.

Hint
- You need to abstract, identify, and
choose suitable values. You may write
down your hypothesis.
-  each value has one specific
criterion, you should choose the most
suitable one.
- **Your final output should follow this
template**:

    **"The value is: ___."**

direct copy helpful information from
sample that related to factor criterion,
if any.



'''
```

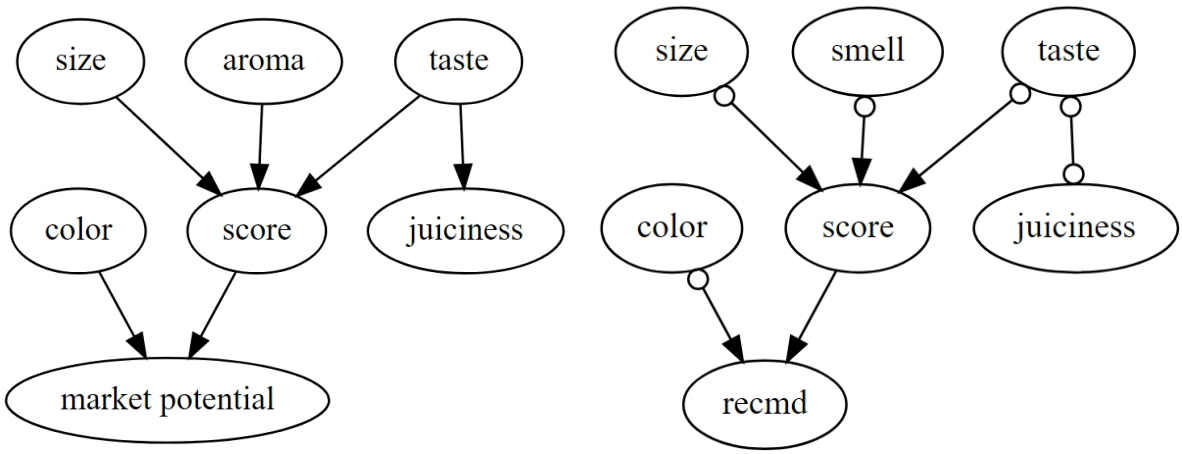
Figure 12. Illustration of the prompt for factor annotation (part 2).

```
### Part 4. Final output


In this part, you are required to report
the value


**Your final output should follow this
template**:

**"The value is: ___."**
```

Figure 13. Illustration of the prompt for factor annotation (part 3).

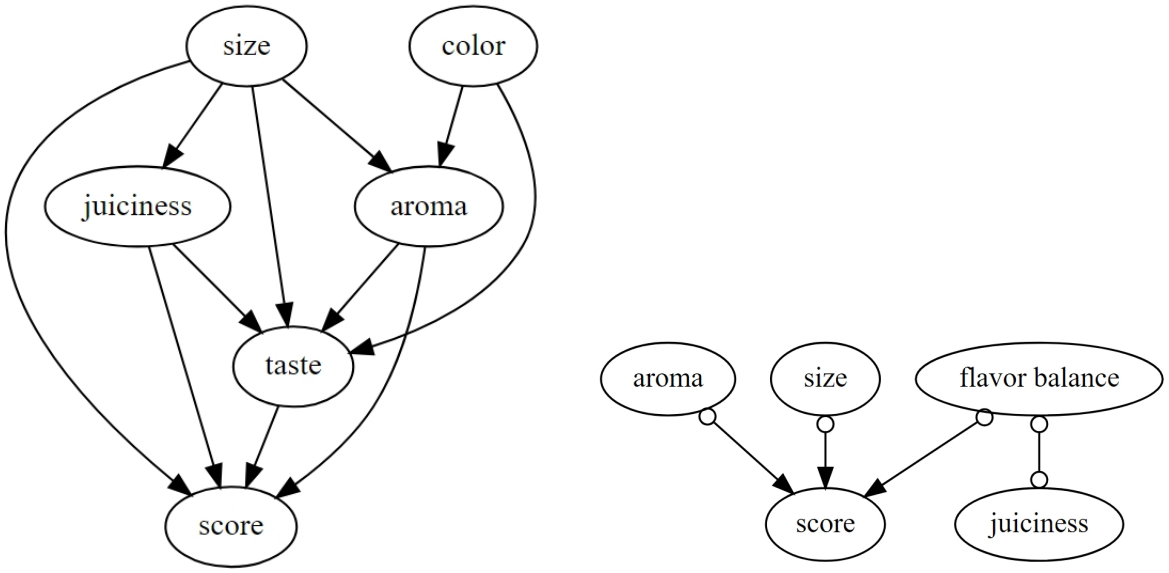
```
… <groups of samples> …
Part 1: Consideration.
In this part, feel free to write down the process of your
considerations.
Hint: You need to abstract, identify, and design suitable factors
with coresponding criteria, and each factor is only allowed to take
value from [-1, 0, 1] .
Part 2: Factor filtration.
You shoud decide whether to use each of the proposed factors by
following criteria:
- Data availability. The concrete value of each new factor could be
easily and clearly read off from most of samples in given data. If
this point is not satisfied, you should not use this factor.

- Informative. Each new factor should exhibit relative stability
within its respective group while manifesting significant variation
across the various groups. If this point is not satisfied, you
should not use this factor.
- Each Factor should focus on one concrete aspect and try to avoid
overlape. If this point is not satisfied, you should not use this
factor.
- Avoid overlapping with those existing factors:
  - size
  - aroma
  - flavor balance
  - juiciness
  - appearance
```

Figure 14. Illustration of the prompt for feedback.



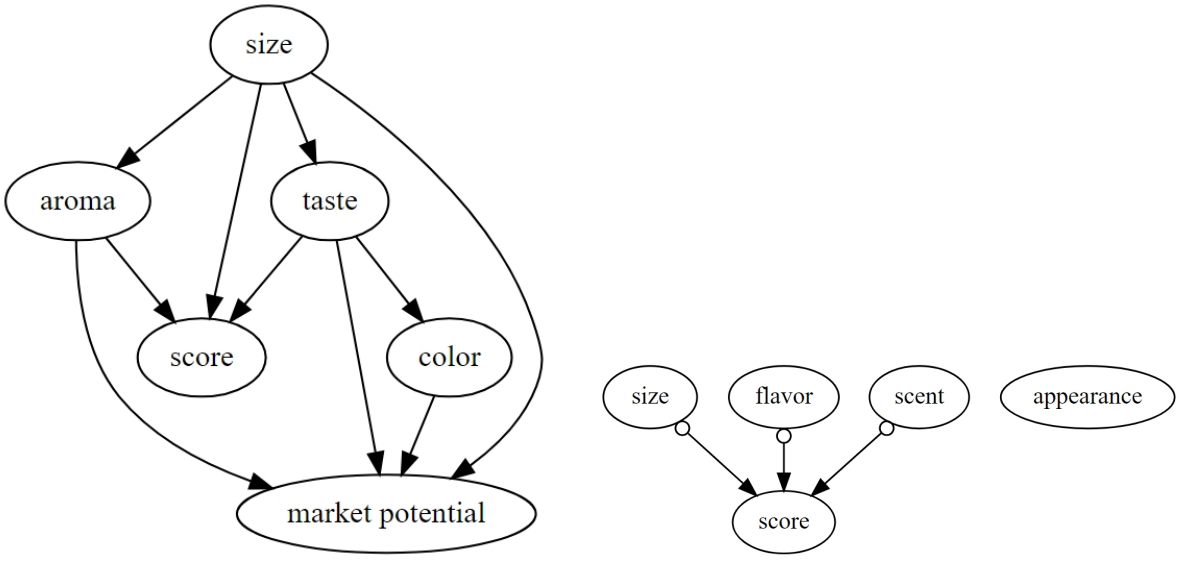
(a) Ground truth (b) FCI results

Figure 15. Ground truth and faithful (via FCI algorithm) causal graphs in AppleGastronome.



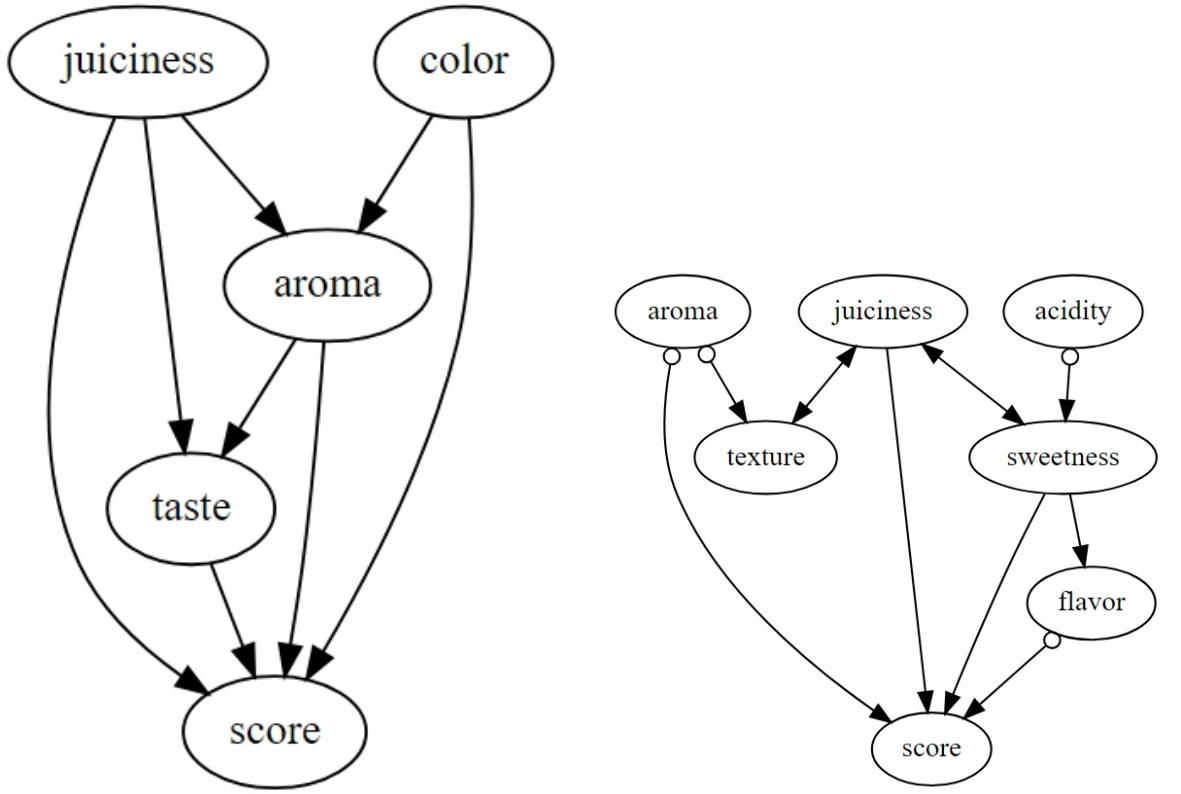
(a) GPT-4 reasoning (b) GPT-4 COAT

Figure 16. Causal graphs with GPT-4 in AppleGastronome.



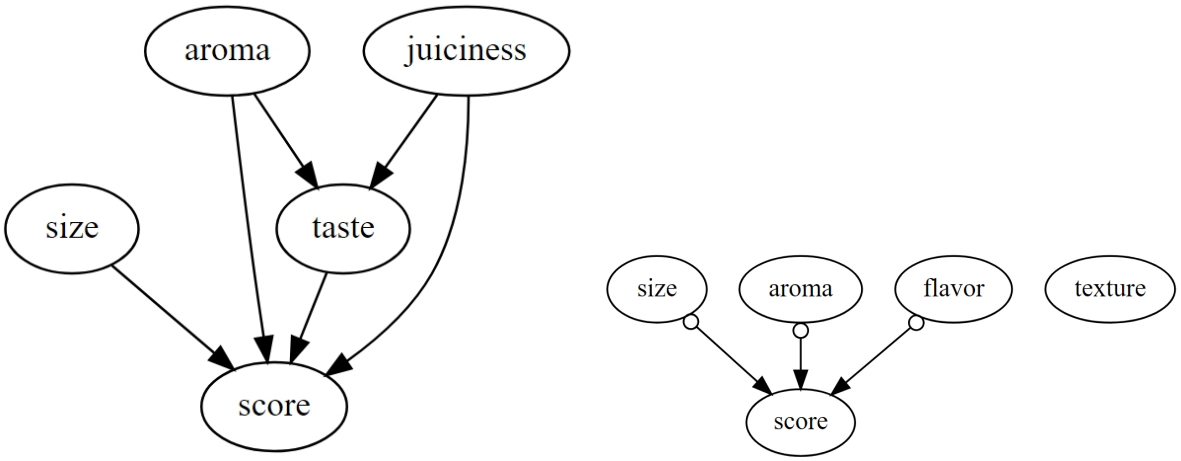
(a) GPT-3.5 reasoning (b) GPT-3.5 COAT

Figure 17. Causal graphs with GPT-3.5 in AppleGastronome.



(a) LLaMA2 reasoning (b) LLaMA2 COAT

Figure 18. Causal graphs with GPT-3.5 in AppleGastronome.



(a) Mistral reasoning (b) Mistral COAT

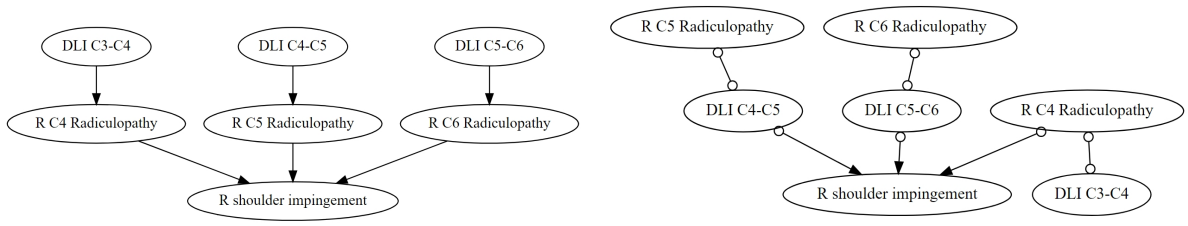
Figure 19. Causal graphs with GPT-3.5 in AppleGastronome.

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

Write a clinical notes based on
following informaiton:

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 20. Illustration of prompts for generating Neuropathic.

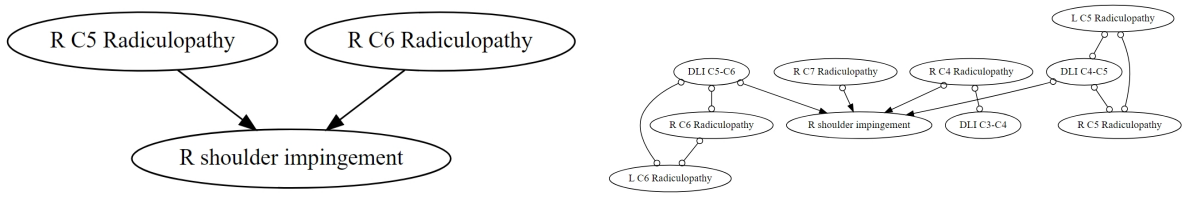
```
## Group with 'R shoulder impingement'=0
----
Patient ID: Neuro_00000
Age: 57
Gender: Male

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 21. Illustration of examples in AppleGastronome.



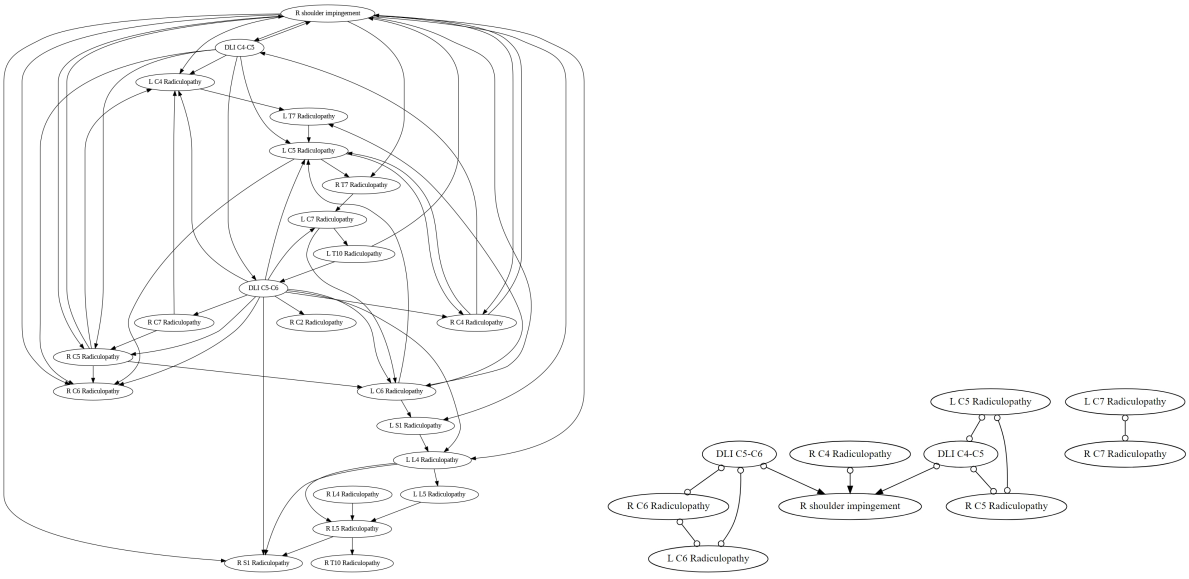
(a) Ground truth (b) FCI results

Figure 22. Ground truth and faithful (via FCI algorithm) causal graphs in Neuropathic.



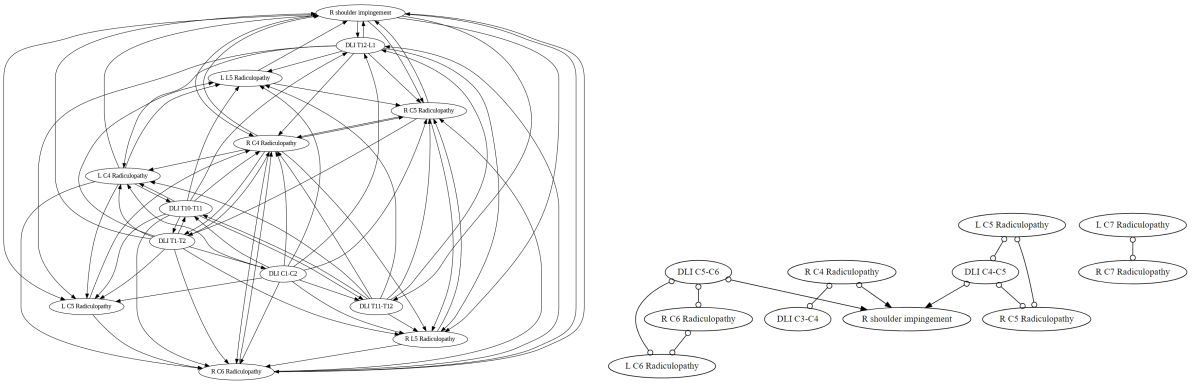
(a) GPT-4 reasoning (b) GPT-4 COAT

Figure 23. Causal graphs with GPT-4 in Neuropathic.



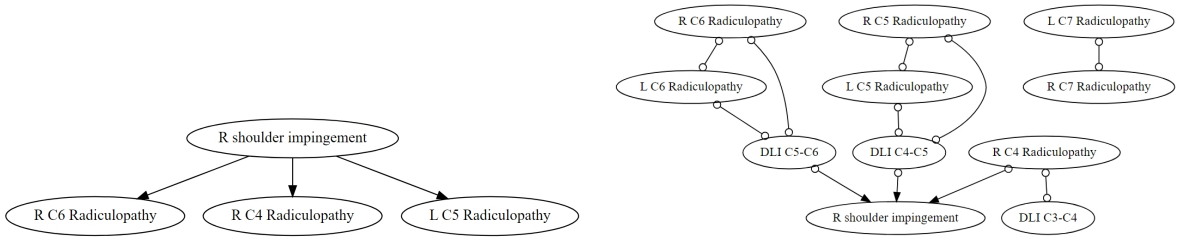
(a) GPT-3.5 reasoning (b) GPT-3.5 COAT

Figure 24. Causal graphs with GPT-3.5 in Neuropathic. (part 1)



(a) LLaMA2 reasoning (b) LLaMA2 COAT

Figure 25. Causal graphs with GPT-3.5 in Neuropathic. (part 2)



(a) Mistral reasoning (b) Mistral COAT

Figure 26. Causal graphs with GPT-3.5 in Neuropathic. (part 3)

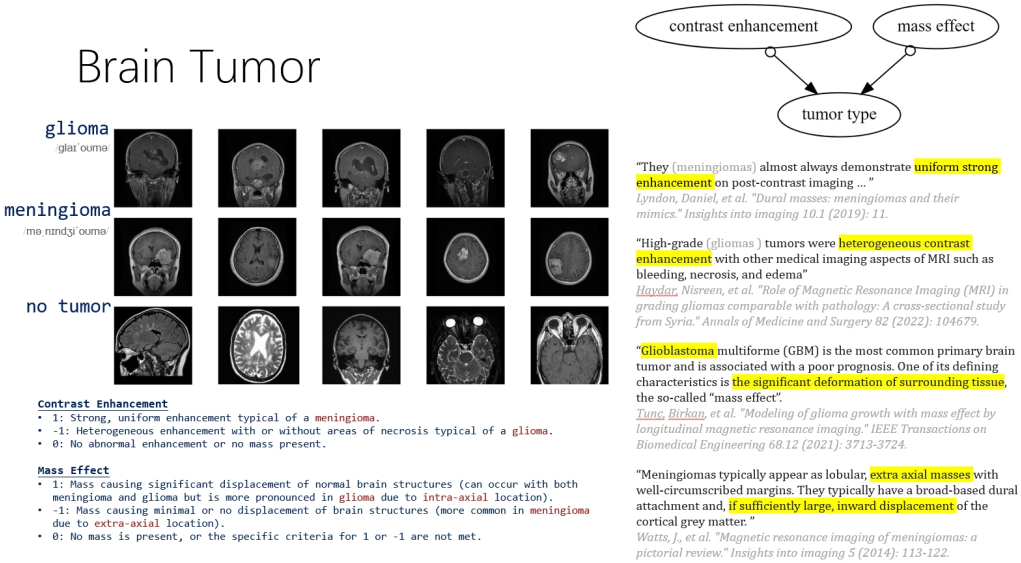
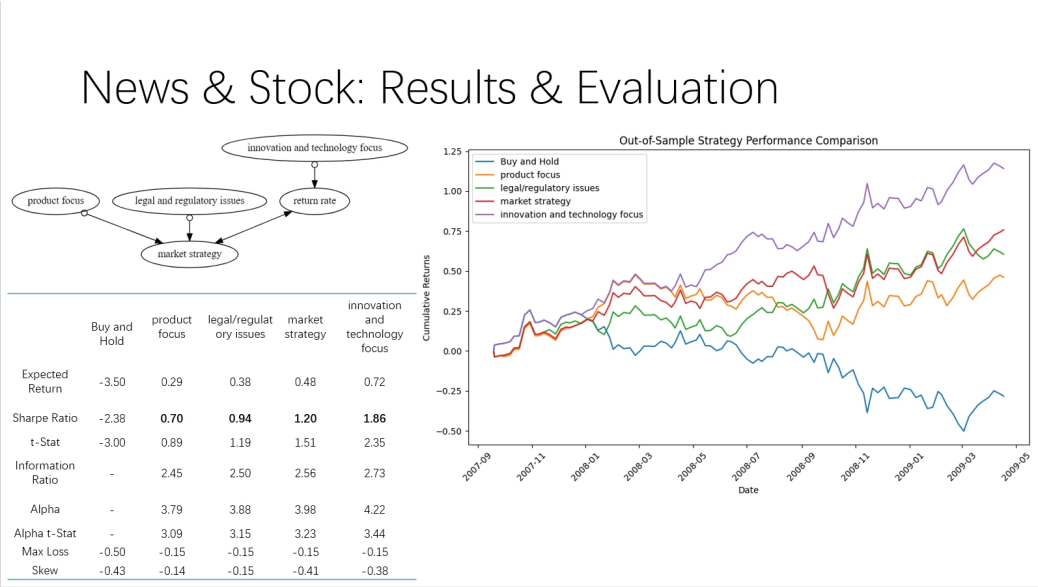


Figure 27. Visual Illustration for the Brain Tumor Case Study



	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
Alpha	-	3.79	3.88	3.98	4.22
Alpha 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

Figure 28. Visual Illustration for the News & Stock Case Study