LEARNING TO DEFER FOR CAUSAL DISCOVERY WITH IMPERFECT EXPERTS

Oscar Clivio[1,*] **Divyat Mahajan**[2,3] **Perouz Taslakian**[4,†] **Sara Magliacane**[5,†]

Ioannis Mitliagkas[2,3] **Valentina Zantedeschi**[4,6,‡] **Alexandre Drouin**[2,4,6,‡]

[1] University of Oxford [2] Mila – Quebec Artificial Intelligence Institute [3] Université de Montréal

[4] ServiceNow Research [5] University of Amsterdam [6] Université Laval

ABSTRACT

Integrating expert knowledge, e.g. from large language models, into causal discovery algorithms can be challenging when the knowledge is not guaranteed to be correct. Expert recommendations may contradict data-driven results, and their reliability can vary significantly depending on the domain or specific query. Existing methods based on soft constraints or inconsistencies in predicted causal relationships fail to account for these variations in expertise. To remedy this, we propose L2D-CD, a method for gauging the correctness of expert recommendations and optimally combining them with data-driven causal discovery results. By adapting learning-to-defer (L2D) algorithms for pairwise causal discovery (CD), we learn a deferral function that selects whether to rely on classical causal discovery methods using numerical data or expert recommendations based on textual meta-data. We evaluate L2D-CD on the canonical Tübingen pairs dataset and demonstrate its superior performance compared to both the causal discovery method and the expert used in isolation. Moreover, our approach identifies domains where the expert's performance is strong or weak. Finally, we outline a strategy for generalizing this approach to causal discovery on graphs with more than two variables, paving the way for further research in this area.

1 INTRODUCTION

Causal discovery is a fundamental task in artificial intelligence and data science, where the goal is to identify the unknown causal relationships between a set of random variables from their observations (Spirtes et al., 2001; Pearl, 2009; Peters et al., 2017). This typically involves inferring statistical dependencies in the data (Spirtes et al., 2001), leading to identification of the causal graph up to a Markov Equivalence Class (MEC), whose members all imply the same statistical dependencies (Verma & Pearl, 1990; Spirtes et al., 2001). Hence, we are not guaranteed a unique solution unless we make further distributional or functional assumptions to reduce the set of potential solutions (Shimizu et al., 2006; Peters et al., 2014).

In addition to exploiting statistical dependencies, the use of large language models (LLMs) has recently gained significant attention in the causal discovery literature. This growing interest is largely driven by the promising performance of LLMs in various causal discovery studies (Willig et al., 2022; Jin et al., 2024b). Typically, approaches query an LLM about the causal relationships between two or more variables, by prompting them with the variables' textual names, and optionally with additional information about the variables or an overall context (Abdulaal et al., 2023; Kıcıman et al., 2023; Willig et al., 2023; Long et al., 2023; Khatibi et al., 2024; Darvariu et al., 2024; Long et al., 2024). Other works attempt to use LLMs for conditional independence testing, which is then used for constraint-based causal discovery (Jin et al., 2024a; Cohrs et al., 2024).

[*] Work done while interning at ServiceNow Research, Mila and Université de Montréal.

[†] Equal contribution.

[‡] Equal supervision.

Thus, there has been a rising interest in combining knowledge-based causal discovery from LLMs with traditional statistical causal discovery (Choi et al., 2022; Long et al., 2024). This research direction builds on significant work on combining more general expert knowledge with statistical causal discovery, notably to narrow down the set of potential solutions (Constantinou et al., 2023). Previous work has considered various types of expert knowledge: (i) required directed edges (Meek, 2013; de Campos & Castellano, 2007; Li & Beek, 2018) (ii) (partial) orderings of the variables (Scheines et al., 1998; de Campos & Castellano, 2007; Andrews et al., 2020; Brouillard et al., 2022), (iii) ancestral constraints (Li & Beek, 2018; Chen et al., 2016), and (iv) the negation of previous constraints (Meek, 2013; de Campos & Castellano, 2007; Chen et al., 2016; Li & Beek, 2018).

However, several challenges arise when using LLMs as experts for causal discovery. First, while LLMs are mostly effective at providing knowledge on ancestral and ordering relationships, they are far less reliable in determining direct causal edges. This is because queries on causal relationships between two variables in natural language, such as "does X cause Y?", typically fail to distinguish between direct and indirect effects, as this distinction depends on what other variables are observed (Ban et al., 2023; Kıcıman et al., 2023). Hence, to integrate knowledge from LLMs with statistical causal discovery, we should use causal structures (Magliacane et al., 2017) or algorithms that allow for order or ancestral constraints (Chen et al., 2016; Ban et al., 2023; Vashishtha et al., 2023). However, designing causal discovery algorithms with such constraints is generally more difficult, e.g. due to their non-decomposability (Chen et al., 2016). Further, LLMs might provide incorrect knowledge due to poor training data (Long et al., 2024), lack of specialized knowledge or sensitivity to variations in words used in prompts to refer to causal relationships (e.g. "cause", "influence", etc.) (Darvariu et al., 2024). This makes them "imperfect experts" (Long et al., 2023), necessitating methods that can accommodate incorrect knowledge.

Towards this, a common approach is to incorporate background knowledge as "soft constraints" rather than hard constraints, often using Bayesian priors or initializations in score-based causal discovery methods (Choi et al., 2022; da Silva et al., 2023; Darvariu et al., 2024). However, to the best of our knowledge, there is no consensus on how to select the appropriate prior and score, which are crucial for determining the final set of solutions and their correctness (Constantinou et al., 2023; Darvariu et al., 2024). Alternatively, Long et al. (2024) propose incorporating a model for the expert's correctness in the likelihood. However, this approach has limitations, such as using LLMs to query direct edges and assuming that imperfect experts make conditionally independent errors, an assumption that is often unrealistic as experts usually show systematic errors, e.g. they struggle on specific domains. Other methods involve LLM experts to estimate their own confidence (Zhang et al., 2023) or double check their own results (Ban et al., 2023); but these are still vulnerable to the expert's mistakes. Finally, some works propose detecting inconsistencies caused by the incorrect expert knowledge (de Campos & Castellano, 2007; Chen et al., 2023), however not all incorrect expert knowledge would cause such inconsistencies so this approach may not be sufficient to detect them. Crucially, none of these methods accounts for the features in each individual causal query, or more generally identifies which *specific* knowledge returned by LLMs is correct and which is not, which is critical as LLMs might have better predictions on some domains such as common sense or insurance knowledge (Darvariu et al., 2024), and struggle on others such as physics (Brown et al., 2020) or specialized medical knowledge (Darvariu et al., 2024).

Contributions. In contrast to previous work, we propose to directly *learn and predict* which ancestral relationships returned by imperfect experts are correct, and which are not, informing the causal discovery process. To this aim, we build on the rich literature on *learning to defer* (L2D) (Madras et al., 2018; Mozannar & Sontag, 2021) that consists in learning a deferral function: for any instance with features and an outcome, this function selects whether to predict the outcome using either a given machine learning model or an external black-box expert based on the features. While the deferral function and the model are most often learned jointly, recent works propose to learn the deferral function in a post-hoc manner, i.e. using already trained models (Narasimhan et al., 2022; Mao et al., 2024). We build on this methodology to learn whether to use an imperfect expert or a statistical causal discovery method to decide on the ancestry between two variables. Pairwise causal predictions can then be aggregated to form a topological order using techniques outlined in the extensive literature on ranking from noisy pairwise comparisons (Bradley & Terry, 1952; Feige et al., 1994; Rajkumar & Agarwal, 2014; Ren et al., 2021).

Our key contributions are summarized below:

1. We propose a learning-to-defer approach to integrate background knowledge in pairwise causal discovery; notably, we show how this reduces to a modified form of classification, allowing the use of any off-the-shelf classifier for this purpose. We call the resulting method **L2D-CD**.

2. We show that this approach improves causal direction predictions in the canonical Tübingen pairs (Mooij et al., 2016). The L2D-CD combination of an imperfect expert and a statistical causal discovery method outperforms each method alone and a simple deferral baseline that defers at random, showing the importance of *learning* to defer.
3. We describe how to extend this approach to graphs with 3 variables or more, building on the literature on ranking from pairwise comparisons.

2 METHODOLOGY

We consider the task of determining causal relationships between two causal variables, identified by their names (u, v). We denote the corresponding observational data with N samples for these variables as (x_u, x_v) where $x_u, x_v \in \mathbb{R}^N$. Further, we assume access to metadata C – some textual context/description that provides extra information about the relationship between these variables. We represent the combined numerical data and textual description associated with the causal variables as $x = (C, u, v, x_u, x_v)$.

Let $y = \mathbb{I}_{u \to v}$, where $\mathbb{I}$ is the indicator function, denote a binary label for the causal relationship, i.e. $y = 1$ if u causes v and $y = 0$ otherwise. We assume access to the following predictors for causal relationships:

- *Expert Predictor.* Query an expert, typically an LLM, to obtain the causal relationship, where the expert uses only the textual description for prediction, i.e., $h_1(x) = \text{Expert}(C, u, v)$.
- *Causal Discovery Methods.* Train a causal discovery method using a numerical/observational dataset to predict causal relationships, i.e., $h_0(x) = CD(x_u, x_v)$.

Our goal is to construct a predictor $h^\star(x)$ of y that optimally combines the predictions from the expert $h_1(x)$ and the causal discovery method $h_0(x)$.

2.1 BACKGROUND ON LEARNING TO DEFER

To combine multiple predictors, we use techniques from the literature on learning to defer (L2D) (Madras et al., 2018; Mozannar & Sontag, 2021; Mao et al., 2024). Consider the task of binary prediction for input x and labels y, with values in $\mathcal{X}$ and $\mathcal{Y}$ respectively, where $\mathcal{Y}$ is finite. Given the base predictor $h(x) = h_0(x)$ and n_e expert predictors $\{h_j(x)\}_{j=1}^{n_e}$, the goal is to learn a deferral function $r(x) = \arg\max_{j \in [0, n_e]} r_j(x)$ that chooses between the different predictors for the sample x; e.g. $r(x) = 0$ means that r chooses the base predictor h and $r(x) = j$ means that r chooses the j-th expert predictor h_j. Critically, only one of the base predictors and any expert predictor is chosen for a given instance x. As a result, the combined predictor $h^\star(x)$ can be constructed as $h^\star(x) = \sum_{j=0}^{n_e} \mathbb{I}_{r(x)=j} h_j(x)$.

In order to learn the deferral function, we use the loss objective from Mao et al. (2024):

$$L_{\text{def}}(h, r, x, y) = \mathbb{I}_{h(x) \neq y} \mathbb{I}_{r(x)=0} + \sum_{j=1}^{n_e} c_j(x, y) \mathbb{I}_{r(x)=j}, \tag{1}$$

where $c_j(x, y) \in [0, 1]$ denotes the cost of predicting y from instance x associated with expert predictor $h(x)$, while the cost of the base predictor h is the standard 0-1 loss. Essentially, L_{def} is the loss incurred by the only base or expert predictor that r chooses on instance x; indeed, only one of $\mathbb{I}_{r(x)=j}$ for $j = 0, 1, \ldots, n_e$ is one and all others are zero. This leads to the following optimization problem,

$$\arg\min_r \mathbb{E}_{x,y}\big[L_{\text{def}}(h, r, x, y)\big]. \tag{2}$$

L2D typically learns both h and r jointly (Madras et al., 2018); however in our setting both causal direction predictors are pre-fitted, so we choose the version of Mao et al. (2024) where r is learned *after* h has been learned. We refer to this setup as *post-hoc L2D* and to the two steps of learning h then r as *two-stage L2D*. In the rest of the manuscript, unless specified otherwise, L2D refers to post-hoc L2D. Since the above loss is not differentiable for learning the deferral function r, the literature typically considers surrogate losses; for example Mao et al. (2024) use the following surrogate loss:

$$L^h_{\text{surr}}(r, x, y) = \mathbb{I}_{h(x)=y} l_2(r, x, 0) + \sum_{j=1}^{n_e} \bar{c}_j(x, y) l_2(r, x, j) \tag{3}$$

where $\bar{c}_j(x,y) = 1 - c_j(x,y)$ and $l_2(r,x,y)$ is a surrogate loss using soft assignment functions $r_j(x)$ for multi-class classification problem with classifier $r(x)$. If we set l_2 to the logistic classification loss, we obtain the following:

$$L^h_{\text{surr}}(r,x,y) = -\mathbb{I}_{h(x)=y} \log \frac{1}{1+\sum_{k=1}^{n_e} e^{-r_k(x)}} - \sum_{j=1}^{n_e} \bar{c}_j(x,y) \log \frac{e^{-r_j(x)}}{1+\sum_{k=1}^{n_e} e^{-r_k(x)}}. \tag{4}$$

Thus, L2D can be instantiated to our setting as $n_e = 1$, $x = (C, u, v, x_u, x_v)$, $y = \mathbb{I}_{u\to v}$, $h_1(x) = \text{Expert}(C, u, v)$, $h(x) = h_0(x) = CD(x_u, x_v)$. It is natural to use $c_1(x,y) = \mathbb{I}_{h_1(x)\neq y}$; however, this setup is excluded by Mao et al. (2024), since it violates an assumption used to prove desirable theoretical properties of the L2D surrogate loss. Nevertheless, as we will show next, these properties still extend to $c_1(x,y) = \mathbb{I}_{h_1(x)\neq y}$ when y is binary, and this choice of $c_1(x,y)$ allows us to reduce L2D to standard classification when $n_e = 1$.

2.2 Consistency bounds for the L2D loss with a 0-1 cost function and binary labels

L2D typically uses a surrogate loss; however does minimizing it actually minimize the original loss? To assess this, we employ the concept of $\mathcal{H}$-consistency bounds (Awasthi et al., 2022; Mao et al., 2024). Let $\mathcal{H}$ be a hypothesis class, ℓ be a non-negative function over (h,x,y) where $h \in \mathcal{H}$ is the predictor, x the features and y the label. Let $\mathcal{E}_\ell(h) := \mathbb{E}_{x,y}[\ell(h,x,y)]$ and $\mathcal{E}^\star_\ell(\mathcal{H}) := \min_{h\in\mathcal{H}} \mathcal{E}_\ell(h)$. Further, denote $\mathcal{M}_\ell(\mathcal{H}) := \mathcal{E}^\star_\ell(\mathcal{H}) - \mathbb{E}_x\left[\inf_{h\in\mathcal{H}} \mathbb{E}_{y|x}[\ell(h,x,y)]\right]$ the minimizability gap of $\mathcal{H}$ and ℓ, which is non-negative and vanishes when $\mathcal{E}^\star_\ell(\mathcal{H})$ coincides with the Bayes error of ℓ. Then, an $\mathcal{H}$-consistency bound of a surrogate loss ℓ_s with respect to an original function ℓ_o is a bound of the form

$$\forall h \in \mathcal{H},\ \mathcal{E}_{\ell_o}(h) - \mathcal{E}^\star_{\ell_o}(\mathcal{H}) + \mathcal{M}_{\ell_o}(\mathcal{H}) \leq \Gamma\left(\mathcal{E}_{\ell_s}(h) - \mathcal{E}^\star_{\ell_s}(\mathcal{H}) + \mathcal{M}_{\ell_s}(\mathcal{H})\right),$$

where Γ is a non-decreasing concave function such that $\Gamma(0) = 0$. Such a bound is desirable as it implies Bayes-consistency of ℓ_s w.r.t. ℓ_o, while also quantifying how improvements in $\mathcal{E}_{\ell_s}(h)$ translate to improvements in $\mathcal{E}_{\ell_o}(h)$ for $h \in \mathcal{H}$. From now on, we use "consistency bound" to refer to an $\mathcal{H}$-consistency bound without specifying the hypothesis class $\mathcal{H}$.

Theorem 6 of Mao et al. (2024) shows that when the surrogate loss used to learn the predictor h and the surrogate loss for the deferral function r have consistency bounds with respect to the multi-class 0-1 loss, then the L2D surrogate loss L_{surr} has a generalized form of consistency bound with respect to the original L2D loss L_{def} in Equation 1. However, this result relies on the assumption that

$$\underline{c}_j \leq \bar{c}_j(x,y) \leq \bar{c}_j,\ \ \forall j, x, y$$

for some $\underline{c}_j > 0$ and $\bar{c}_j \leq 1$, which does not apply when $c_j(x,y) = \mathbb{I}_{h_j(x)\neq y}$ for which $\forall j,\ \ c_j(x,y) = 0$ for some x, y. However, as we show next, the L2D loss does have such a consistency bound in this case, if one further assumes a binary y. We provide the proof in Appendix A.

Lemma 2.1 *Assume that y is binary and $\forall j = 1,\ldots,n_e,\ \ c_j(x,y) = \mathbb{I}_{h_j(x)\neq y}$. Denote $\mathcal{H}$ and $\mathcal{R}$ hypothesis classes for the base predictor h and the deferral function r, respectively. Assume that l_1 has a $\mathcal{H}$-consistency bound w.r.t. the binary 0-1 loss with concave function Γ_1, and l_2 has a $\mathcal{R}$-consistency bound w.r.t. the multiclass (with n_e+1 classes) 0-1 loss with concave function Γ_2. Then, for all $h \in \mathcal{H}$ and $r \in \mathcal{R}$,*

$$\begin{aligned}&\mathcal{E}_{L_{def}}(h,r) - \mathcal{E}^*_{L_{def}}(\mathcal{H},\mathcal{R}) + \mathcal{M}_{L_{def}}(\mathcal{H},\mathcal{R})\\ &\leq \Gamma_1(\mathcal{E}_{\ell_1}(h,r) - \mathcal{E}^*_{\ell_1}(h) + \mathcal{M}_{\ell_1}(\mathcal{H})) + n_e\Gamma_2\left(\mathcal{E}_{L^h_{surr}}(r) - \mathcal{E}^*_{L^h_{surr}}(r) + \mathcal{M}_{L^h_{surr}}(\mathcal{R})\right)\end{aligned}$$

where the n_e factor can be removed when Γ_2 is linear. In particular, if $\mathcal{H}$ is a singleton $\{h_0\}$, then this reduces to a $\mathcal{R}$-consistency bound of $L^{h_0}_{surr}$ w.r.t. $L^{h_0}_{def}(r,x,y) := L_{def}(h_0,r,x,y)$.

Notably, Lemma 2.1 will ensure that one can minimize the surrogate loss in Equation 3 with guarantees on the combined predictor $h^\star(x)$ in our pairwise causal discovery setting

2.3 Post-hoc L2D with a 0-1 cost function and a single expert can be reduced

TO STANDARD CLASSIFICATION

We now show that post-hoc L2D applied to our setting can be reduced to binary classification. Indeed, assume $c_j(x, y) = \mathbb{I}_{h_1(x)\neq y}$ and let

$$\mathcal{D}_\circ = \{(x, y) \mid \mathbb{I}_{h(x)\neq y} = \mathbb{I}_{h_j(x)\neq y} \ \forall j = 1, \dots, n_e\}$$

which is the set of pairs of features and labels on which all predictors are equally correct or wrong. Then it turns out that for any $(x, y) \in \mathcal{D}_\circ$, $L_{\text{def}}(h, r, x, y) = \mathbb{I}_{h(x)\neq y}$ does not depend on r. In particular, noting $\mathcal{D}_\circ^{\complement}$ the complement of $\mathcal{D}_\circ$ this leads to

$$\mathbb{E}_{x,y}\big[L_{\text{def}}(h, r, x, y)\big] = p(\mathcal{D}_\circ^{\complement})\mathbb{E}_{x,y}\big[L_{\text{def}}(h, r, x, y)\big|(x, y) \notin \mathcal{D}_\circ\big] + p(\mathcal{D}_\circ)\mathbb{E}_{x,y}\big[\mathbb{I}_{h(x)\neq y}\big|(x, y) \in \mathcal{D}_\circ\big]$$

Thus, we target the alternative loss $\mathbb{E}_{x,y}\big[L_{\text{def}}(h, r, x, y)\big|(x, y) \notin \mathcal{D}_\circ\big]$ instead of the original loss $\mathbb{E}_{x,y}\big[L_{\text{def}}(h, r, x, y)\big]$. Further, when $n_e = 1$, for any $(x, y) \notin \mathcal{D}_\circ$, then $\mathbb{I}_{h(x)\neq y} = \mathbb{I}_{h_1(x)=y}$. Thus,

$$L_{\text{def}}(h, r, x, y) = \mathbb{I}_{h_1(x)=y}\mathbb{I}_{r(x)=0} + \mathbb{I}_{h_1(x)\neq y}\mathbb{I}_{r(x)=1} = \mathbb{I}_{r(x)\neq \mathbb{I}_{h_1(x)=y}}$$

is the 0-1 classification loss for features x and label $\mathbb{I}_{h_1(x)=y}$, which is the indicator that $h_1(x)$ returns the correct prediction y. This amounts to learning to predict when the expert is correct or, equivalently, when the base predictor is wrong. As a result, in this setting, post-hoc L2D reduces to binary classification over samples not in $\mathcal{D}_\circ$. This allows us to use any off-the-shelf classification method for training, while directly optimizing the expected surrogate loss without removing samples in $\mathcal{D}_\circ$ would generally only be feasible through automatic differentiation, as done in the implementation of (Mao et al., 2024).

2.4 PROPOSED APPROACH: L2D-CD

Based on the above discussion, our training procedure to obtain a fitted deferral function $\hat{r}$ from a set of m examples $(x_i, y_i)_{i=1}^m$ when $n_e = 1$ is thus to:

1. Compute the set $S = \{i \mid \mathbb{I}_{h(x_i)\neq y_i} \neq \mathbb{I}_{h_1(x_i)\neq y_i}\}$. Note that when y is binary, we generally have $\mathcal{D}_\circ = \{(x, y) \mid x \in \mathcal{X}_\circ\}$ where $\mathcal{X}_\circ := \{x \mid h(x) = h_j(x) \ \forall j = 1, \dots, n_e\}$, so here S can also be computed as $S = \{i \mid h(x_i) \neq h_1(x_i)\}$.
2. Obtain $\hat{r}$ by fitting any binary classification method to the set $(x_i, y'_i)_{i\in S}$ where $y'_i = \mathbb{I}_{h_1(x_i)=y_i}$.

For pairwise causal discovery, we apply this procedure to $x = (C, u, v, x_u, x_v)$, $y = \mathbb{I}_{u\to v}$, $h_1(x) = \text{Expert}(C, u, v)$, $h(x) = h_0(x) = CD(x_u, x_v)$. We call the resulting method **L2D-CD**. Thus, intuitively L2D-CD learns which causal discovery predictor, statistical-based method or metadata-based expert, returns the correct causal direction when the two predictors differ in their predictions.

3 RESULTS ON TÜBINGEN DATASETS

We now apply the approach from Section 2.4 to the Tübingen pairs (Mooij et al., 2016), a canonical benchmark for pairwise causal discovery.

Division of Tübingen pairs by domain: In order to design synthetic experts having strengths in certain domains and, potentially, weaknesses in others, we manually assign a domain to each Tübingen pair. After excluding multivariate pairs (52-55, 71, 105), Tübingen pairs are each assigned to five different domains: Climate/Environment, Economics/Finance, Biology, Medicine, Physics. Further, a training set and a testing set are formed by stratified sampling w.r.t. the domains at 50/50 proportions. Table 3 in Appendix B details the pairs present in each domain as well as in each of the training and testing sets.

Causal discovery methods: We consider three causal discovery (CD) methods that can be applied on cause-effect pairs: LiNGAM (more specifically, Direct-LiNGAM (Shimizu et al., 2011)), RECI (Blöbaum et al., 2018) and bQCD (Tagasovska et al., 2020). Notably, LiNGAM assumes linear structural equations with non-Gaussian noise, while RECI and bQCD rely on a postulate of independence of causal mechanisms (Peters et al., 2017); this means that they may also be misspecified, thus "imperfect", if such assumptions are violated, further strengthening the motivation of combining them with experts.

Synthetic experts: We design synthetic experts having a pre-defined probability $p_d \in [0, 1]$ of returning the correct answer for each domain $d \in \{$Climate/Environment, Economics/Finance, Biology, Medicine, Physics$\}$, i.e. for each pair $i \in d$, the expert returns the correct causal direction with probability p_d and the incorrect direction with probability $1 - p_d$. We consider two types of experts:

Information for pairs0008:

`https://archive.ics.uci.edu/ml/datasets/Abalone`

1. **Title of Database:** Abalone data
2. **Sources:**
   a) **Original owners of database:** Marine Resources Division Marine Research Laboratories - Taroona Department of Primary Industry and Fisheries, Tasmania GPO Box 619F, Hobart, Tasmania 7001, Australia *(contact: Warwick Nash, +61 02 277277, wnash@dpi.tas.gov.au)*
   b) **Donor of database:** Sam Waugh (`Sam.Waugh@cs.utas.edu.au`) Department of Computer Science, University of Tasmania GPO Box 252C, Hobart, Tasmania 7001, Australia
   c) **Date received:** December 1995
3. **Attribute Information:** Given is the attribute name, attribute type, the measurement unit, and a brief description.

Name	Data Type	Meas.	Description
x: Rings	Integer		+1.5 gives the age in years
y: Height	Continuous	mm	with meat in shell

Figure 1: Example ground-truth-free textual description (D'_i) for a Tübingen pair (here $i = 8$). Formatting was added here to help the reader but raw text is given to the LLM.

- ϵ-experts : for a given $\epsilon \in (0, 0.5)$, define:

$$p_{\text{Biology}} = p_{\text{Economics/Finance}} = p_{\text{Physics}} = 1 - \epsilon \text{ and } p_{\text{Climate/Environment}} = p_{\text{Medicine}} = \epsilon \quad (5)$$

  Notably, the expert is considered as "good" on domains d where $p_d = 1 - \epsilon > 0.5$, and "bad" on domains d where $p_d = \epsilon < 0.5$. Domains in Equation 5 are chosen in order to assure a roughly equal share between pairs in a good domain (54) and pairs in a bad domain (48). In the following, we refer to these experts as "$\epsilon = \dots$" where "$\dots$" refers to the selected value of ϵ. We considered $\epsilon = 0.05$, $\epsilon = 0.1$, and $\epsilon = 0.2$.

- p-experts : here we consider deterministic experts, i.e. $p_d \in \{0, 1\} \ \forall d$. However, unlike the previous experts, we change values of p_d by domain d. To maintain consistency with the previous ϵ-experts where three domains are good and two are bad, we select assignments of p_d's such that three p_d's are equal to 1 and two are equal to 0. In the following, we refer to these experts as "ABC" were A, B, C refer to the initials of the domains d where $p_d = 1$

LLM experts : Finally, we consider experts implemented using OpenAI GPT models (OpenAI, 2024). For every Tübingen pair i with original description D_i, we manually remove ground-truth mentions in D_i to obtain a ground-truth-free description D'_i (see Figure 1 for an example); and prompt the OpenAI model as follows :

- System part of the prompt : *You will be given a text describing two columns in a dataset. The text will be delimited by backticks as in a code block. The first column is also referred to as "x" and the second column as "y". Based on the text description between backticks, is it more likely that 1) x causes y, or that 2) y causes x? Please choose one and only one of these two options.*
- User part of the prompt : the text description D'_i between two sets of three backticks and two spaces, as in a code block. An example of such D'_i is given in Figure 1.

Then, we parse the causal direction from the answer. We considered GPT-4o and GPT-4o-mini with default hyperparameters; due to stochasticity we varied the seed which was assigned values 0, 1, . . . , 19.

L2D-CD model: L2D-CD was implemented according to Section 2.4, where for each Tübignen pair i, the variables remain constant as $(u_i, v_i) = (\texttt{First column (x)}, \texttt{Second column (y)})$, the numerical variables are the two columns $(X_u, X_v) = (X_1, X_2)$, and the context C_i is the ground-truth-free text description D'_i. For simplicity, we exclude numerical variables X_u, X_v from the input to the deferral function, so with constant (u, v)'s across Tübingen pairs our deferral function becomes:

$$r(x_i) = \text{Classifier}\left(\text{Embedding}(D'_i)\right)$$

where "Embedding" refers to a model that converts the text description to a numerical vector embedding, and "Classifier" refers to any classifier from the space of this embedding. In practice, we observe small training samples (from 11 to 36) after restriction to the set S as in Section 2.4, motivating us to use random forests (Breiman, 2001) as classifiers; we resorted to the scikit-learn implementation (Pedregosa et al., 2011). For the embedding, we use OpenAI's `text-embedding-3-small` model and perform dimensionality reduction according to the recommended practice on the OpenAI website. To select hyperparameters for the L2D method, we perform hyperparameter search on a grid of the following hyperparameters : 10, 50, 100 for the random forest's `n_estimators`; 2, 5 for its `min_samples_split`; 5, 10, 15, 20, 50 for the size of the embedding.

After evaluating these hyperparameters across all possible pairs of experts and CD methods under consideration, and across 20 random training seeds, we select `n_estimators` = 100 and `min_samples_split` = 5 for the random forest, and 50 dimensions for the embedding model. Evaluation is done using leave-one-out (LOO) cross-validation on the training set, using the loss from Equation 1 and the training procedure from 2.4. As a metric for selection, LOO sample-wise losses are averaged for each expert, CD method and random training seed, then for each of the three aforementioned types of experts, then across these three types; this is done in order to balance performance across such types. L2D-CD with the selected hyperparameters is then retrained for each expert, CD method and random seed using the full training set.

Baselines: In order to assess whether the description-wise heterogeneity of the L2D deferral function r impacts performance compared to randomly deferring to either method, we introduce a baseline. In this baseline, the probability of deferring to the expert is set as the fraction of correct expert predictions on the same set S used to train L2D-CD's classifier. The causal direction of the testing set is then predicted as a Bernoulli of this probability. Given that baselines randomly sample causal predictions, in contrast to L2D predictions which are deterministic, we consider 20 random seeds for the sampling.

L2D-CD consistently improves accuracy compared to the expert and CD alone: For each pair of expert and CD method, Table 1 presents the testing accuracies for the CD method alone, for the expert alone, and for the L2D-CD and baseline combinations of these two methods. L2D-CD almost always improves over all other methods, showing its capability to combine accurate predictions from the expert and the CD method. Notably, it outperforms both synthetic experts and real-world LLM-based ones. While Tübingen pairs are known to have been memorized by LLMs since they are available online (Kıcıman et al., 2023), L2D-CD still generally improves on the LLM experts, regardless of whether memorization occurs or not. On the other hand, the baseline's average accuracy interpolates between those of the CD method and the expert, showing the necessity of instance-dependent predictions.

L2D-CD can identify strong and weak domains for the expert: While the accuracy results indicate improved performance of L2D-CD on all pairs, one might wonder whether L2D-CD can identify domains where the expert is strong and those where the expert is weak. To assess this, we focus on synthetic ϵ-experts and p-experts where performance is controlled through domain-wise probabilities p_d. Let $r(x; \text{CD}, \text{Expert}, \epsilon')$ be the deferral function corresponding to a given causal discovery method CD, a given expert Expert and a random variable ϵ' capturing sources of uncertainty, such as the expert seed for stochastic experts, the training seed for random forests, or the sampling seed for baselines. Then, the probability that r chooses Expert on domain d can be computed as

$$p(\text{Expert chosen by r}|\text{Expert}, d) = \mathbb{E}_{X,\epsilon',\text{CD}}[1_{\{r(X;\text{CD},\text{Expert},\epsilon')=\text{Expert}\}} \mid X \in d]$$

where we average over the CD method, the samples in the domain and uncertainty from the expert and the deferral function fitting. We say that r is **domain-consistent** for Expert if, on average, it defers to Expert more frequently on any domain d_+ where Expert is strong than any domain d_- where Expert is weak:

$$H^1_{d_+,d_-,r,\text{Expert}} : p(\text{Expert chosen by r}|\text{Expert}, d_+) > p(\text{Expert chosen by r}|\text{Expert}, d_-) \tag{6}$$

Notably, this would enable setting a threshold separating strong and weak domains for the expert. For experts with domain-wise probabilities p_d, strong domains are defined as $p_d > 0.5$ and weak domains $p_d < 0.5$.

Table 1: Hold-out accuracies by possible combinations of causal discovery (CD) method and expert, averaged by random seeds w.r.t. stochastic expert predictions, for random forest training and baseline sampling predictions (20 seeds for each). Associated standard errors are shown after "±"; Notably, the standard error is always zero for CD methods and for p-experts, as they are deterministic.

CD	Expert	CD Acc	Expert Acc	L2D-CD Acc	Baseline Acc
LiNGAM	EMP	0.442 ± 0.000	0.538 ± 0.000	$\mathbf{0.661 \pm 0.005}$	0.489 ± 0.011
	CMP	0.442 ± 0.000	0.635 ± 0.000	$\mathbf{0.700 \pm 0.004}$	0.554 ± 0.010
	CEP	0.442 ± 0.000	$\mathbf{0.692 \pm 0.000}$	0.691 ± 0.006	0.603 ± 0.007
	CEM	0.442 ± 0.000	0.673 ± 0.000	$\mathbf{0.749 \pm 0.005}$	0.595 ± 0.009
	BMP	0.442 ± 0.000	0.481 ± 0.000	$\mathbf{0.648 \pm 0.006}$	0.466 ± 0.013
	BEP	0.442 ± 0.000	0.538 ± 0.000	$\mathbf{0.681 \pm 0.004}$	0.486 ± 0.009
	BEM	0.442 ± 0.000	0.519 ± 0.000	$\mathbf{0.626 \pm 0.007}$	0.476 ± 0.012
	BCP	0.442 ± 0.000	0.635 ± 0.000	$\mathbf{0.730 \pm 0.004}$	0.540 ± 0.008
	BCM	0.442 ± 0.000	0.615 ± 0.000	$\mathbf{0.687 \pm 0.003}$	0.532 ± 0.009
	BCE	0.442 ± 0.000	0.673 ± 0.000	$\mathbf{0.731 \pm 0.004}$	0.568 ± 0.009
bQCD	EMP	0.692 ± 0.000	0.538 ± 0.000	$\mathbf{0.788 \pm 0.005}$	0.623 ± 0.009
	CMP	0.692 ± 0.000	0.635 ± 0.000	$\mathbf{0.854 \pm 0.003}$	0.656 ± 0.012
	CEP	0.692 ± 0.000	0.692 ± 0.000	$\mathbf{0.857 \pm 0.005}$	0.688 ± 0.013
	CEM	0.692 ± 0.000	0.673 ± 0.000	$\mathbf{0.779 \pm 0.002}$	0.683 ± 0.009
	BMP	0.692 ± 0.000	0.481 ± 0.000	$\mathbf{0.775 \pm 0.005}$	0.620 ± 0.011
	BEP	0.692 ± 0.000	0.538 ± 0.000	$\mathbf{0.796 \pm 0.005}$	0.633 ± 0.010
	BEM	0.692 ± 0.000	0.519 ± 0.000	$\mathbf{0.734 \pm 0.005}$	0.618 ± 0.006
	BCP	0.692 ± 0.000	0.635 ± 0.000	$\mathbf{0.856 \pm 0.003}$	0.668 ± 0.012
	BCM	0.692 ± 0.000	0.615 ± 0.000	$\mathbf{0.722 \pm 0.004}$	0.657 ± 0.009
	BCE	0.692 ± 0.000	0.673 ± 0.000	$\mathbf{0.753 \pm 0.002}$	0.676 ± 0.011
RECI	EMP	0.654 ± 0.000	0.538 ± 0.000	$\mathbf{0.772 \pm 0.004}$	0.590 ± 0.009
	CMP	0.654 ± 0.000	0.635 ± 0.000	$\mathbf{0.740 \pm 0.003}$	0.652 ± 0.010
	CEP	0.654 ± 0.000	0.692 ± 0.000	$\mathbf{0.846 \pm 0.004}$	0.684 ± 0.010
	CEM	0.654 ± 0.000	0.673 ± 0.000	$\mathbf{0.785 \pm 0.004}$	0.682 ± 0.008
	BMP	0.654 ± 0.000	0.481 ± 0.000	$\mathbf{0.768 \pm 0.004}$	0.585 ± 0.011
	BEP	0.654 ± 0.000	0.538 ± 0.000	$\mathbf{0.883 \pm 0.003}$	0.593 ± 0.008
	BEM	0.654 ± 0.000	0.519 ± 0.000	$\mathbf{0.776 \pm 0.002}$	0.588 ± 0.009
	BCP	0.654 ± 0.000	0.635 ± 0.000	$\mathbf{0.824 \pm 0.004}$	0.652 ± 0.010
	BCM	0.654 ± 0.000	0.615 ± 0.000	$\mathbf{0.767 \pm 0.006}$	0.642 ± 0.007
	BCE	0.654 ± 0.000	0.673 ± 0.000	$\mathbf{0.844 \pm 0.003}$	0.663 ± 0.010
LiNGAM	$\epsilon = 0.05$	0.442 ± 0.000	0.535 ± 0.001	$\mathbf{0.676 \pm 0.001}$	0.484 ± 0.002
	$\epsilon = 0.1$	0.442 ± 0.000	0.531 ± 0.001	$\mathbf{0.672 \pm 0.002}$	0.483 ± 0.002
	$\epsilon = 0.2$	0.442 ± 0.000	0.512 ± 0.002	$\mathbf{0.648 \pm 0.002}$	0.476 ± 0.002
bQCD	$\epsilon = 0.05$	0.692 ± 0.000	0.535 ± 0.001	$\mathbf{0.795 \pm 0.001}$	0.632 ± 0.002
	$\epsilon = 0.1$	0.692 ± 0.000	0.531 ± 0.001	$\mathbf{0.793 \pm 0.001}$	0.630 ± 0.002
	$\epsilon = 0.2$	0.692 ± 0.000	0.512 ± 0.002	$\mathbf{0.785 \pm 0.001}$	0.625 ± 0.002
RECI	$\epsilon = 0.05$	0.654 ± 0.000	0.535 ± 0.001	$\mathbf{0.873 \pm 0.002}$	0.593 ± 0.002
	$\epsilon = 0.1$	0.654 ± 0.000	0.531 ± 0.001	$\mathbf{0.863 \pm 0.003}$	0.593 ± 0.002
	$\epsilon = 0.2$	0.654 ± 0.000	0.512 ± 0.002	$\mathbf{0.815 \pm 0.005}$	0.592 ± 0.002
LiNGAM	GPT4o	0.442 ± 0.000	$\mathbf{0.751 \pm 0.001}$	0.747 ± 0.001	0.662 ± 0.002
	GPT4o-mini	0.442 ± 0.000	0.755 ± 0.001	$\mathbf{0.771 \pm 0.001}$	0.669 ± 0.002
bQCD	GPT4o	0.692 ± 0.000	0.751 ± 0.001	$\mathbf{0.795 \pm 0.002}$	0.739 ± 0.003
	GPT4o-mini	0.692 ± 0.000	0.755 ± 0.001	$\mathbf{0.820 \pm 0.001}$	0.746 ± 0.003
RECI	GPT4o	0.654 ± 0.000	0.751 ± 0.001	$\mathbf{0.773 \pm 0.002}$	0.733 ± 0.002
	GPT4o-mini	0.654 ± 0.000	0.755 ± 0.001	$\mathbf{0.795 \pm 0.001}$	0.742 ± 0.002

Table 2: Domain consistency of each combination method for each synthetic expert

	$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.2$	BCE	BCM	BCP	BEM
L2D-CD	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Baseline	No	No	No	No	No	No	No

	BEP	BMP	CEM	CEP	CMP	EMP
L2D-CD	Yes	No	Yes	Yes	Yes	Yes
Baseline	No	No	No	No	No	No

We exclude LLMs for analysis as (i) they do not have such ground-truth probabilities, and (ii) their domain-wise accuracies are all above 0.5, thus are strong as in the previous definition on all domains. We describe our approach to evaluate domain consistency using hypothesis testing in Appendix C. Table 2 presents whether domain consistency holds depending on whether r is L2D-CD or the baseline and on the synthetic expert. Notably, we can see that the L2D-CD is domain-consistent for almost every synthetic expert, while the baseline is never domain-consistent for any synthetic expert. This shows that L2D-CD can capture the strengths and weaknesses of experts, while deferral based on a constant probability (unsurprisingly) cannot.

4 Extension to graphs with 3 variables or more

To extend L2D-CD to graphs of 3 variables or more, we propose building on methods for ranking from pairwise comparisons (Braverman & Mossel, 2007; Ailon, 2011; Rajkumar & Agarwal, 2014; Mao et al., 2017; Falahatgar et al., 2018; Ren et al., 2021), where a ranking over V is a function π that is bijective from V to $\{1, \ldots, |V|\}$. We let $\pi(u) < \pi(u')$ indicate that u is ranked before u' for any $u, u' \in V$. While these methods differ in their exact problem formulation, they all amount to learning a ranking $\hat{\pi}(u)$ over items u in a set V from samples (u_i, u_i', y_i), where $y_i \in \{-1, 1\}$ indicates a comparison between u_i and u_i', with $y_i = 1$ indicating that u_i is deemed as ranked before u_i' and $y_i = -1$ as the converse. Thus, the learning algorithm $\mathcal{A}$ attempts to find a ranking π that best fits the individual comparisons; the latters are allowed to be contradictory. Thus we propose generalizing our method to causal discovery on more than two variables as follows.

Graph notations: Let G be a graph having nodes V_G, where every node is represented by its name, and edges E_G. Define $\Sigma(V_G, E_G)$ as a matrix of pairwise ancestries derived from E_G, where (u, u') is equal to 1 if u precedes u' in E_G, -1 if u' prcedes u, and 0 if no ancestry relationship between u and u' exists. Additionally, let C_G be the graph's textual context, and $X_G = (X_u)_{u \in V_G}$ its numerical data.

Problem definition: Let u, v be the textual names of nodes, X the full numerical data of a graph, and C a textual context. Assume access to an expert Expert(C, u, v) and a causal discovery oracle CD$(u, v; X)$, both of which determine the ancestry relationship between u and v. The causal discovery oracle applies the causal discovery method to X and, like the expert, may return the absence of ancestry.

Training: Assume that we have a training set of graphs $\mathcal{G}_{\text{train}} = (V_{G_i}, E_{G_i}, C_{G_i}, X_{G_i})_{i=1,\ldots,m_{\text{train}}}$. Then, train a deferral function $r(C, u, v, X)$ between the expert and the causal discovery oracle using L2D-CD where the training sample is composed of each pair of nodes in each graph and its ancestry status, i.e. the training set is formed by $(C_{G_i}, u, u', X_{G_i})_{u,u' \in V_{G_i}, i=1,\ldots,m_{\text{train}}}$ as features, and $(\Sigma_{u,u'}(V_{G_i}, E_{G_i}))_{u,u' \in V_{G_i}, i=1,\ldots,m_{\text{train}}}$ as the corresponding labels.

Inference: For every hold-out graph G, where we have access to V_G, C_G, X_G, sample pairs of nodes $u, u' \in V_G$, infer their ancestry relationship $\hat{y}_{u,u'}$ using L2D-CD, decide on how to handle $y_{u,u'} = 0$, e.g. discard; apply $\mathcal{A}$ to the resulting comparisons to obtain a topological ordering $\hat{\pi}$; optionally deduce an edge set $\hat{E}_G$ using an edge pruning method (Bühlmann et al., 2014).

Algorithms $\mathcal{A}$ for ranking from pairwise preferences typically benefit from convergence to a ground-truth or optimal ranking in a moderately scaling number of steps. Key challenges will be establishing convergence to any of the multiple topological orderings allowed by a causal graph depending on the accuracy of the deferral function, and handling the absence of ancestry in pairwise comparisons.

5 CONCLUSION

We have designed a procedure leveraging learning-to-defer to combine two pairwise causal discovery methods, one conventional method and one expert. Experiments on Tübingen pairs showed that the combined method generally improves over each separate method, and so for both synthetic and real-world LLM-based experts. The learnt deferral function can also identify the expert's strong and weak domains. Note that our methodology can also be applied to human knowledge. An inherent limitation of the L2D-CD approach is the need for a training set, while an improvable one is that our training procedure does not generalize straightforwardly to more than two methods.We also did not yet implement our strategy to generalize L2D-CD for bivariate causal discovery to more general causal discovery, which is future work.

ACKNOWLEDGEMENTS

We sincerely thank Philippe Brouillard for helpful comments on the manuscript. O.C. acknowledges support from the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1) and Novo Nordisk for doctoral studies, as well as from Mitacs for the internship at ServiceNow Research. I.M. acknowledges support by an NSERC Discovery grant (RGPIN-2019-06512), and a Canada CIFAR AI chair. D.M. acknowledges support via FRQNT doctoral training scholarship for his graduate studies.

REFERENCES

Ahmed Abdulaal, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C Castro, Daniel C Alexander, et al. Causal modelling agents: Causal graph discovery through synergising metadata-and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2023.

Nir Ailon. An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity, 2011.

Bryan Andrews, Peter Spirtes, and Gregory F Cooper. On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge. In *International Conference on Artificial Intelligence and Statistics*, pp. 4002–4011. PMLR, 2020.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, pp. 1117–1174. PMLR, 2022.

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

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57 (1):289–300, 1995.

Roger L Berger and Jason C Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996.

Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pp. 900–909. PMLR, 2018.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Mark Braverman and Elchanan Mossel. Noisy sorting without resampling, 2007.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Philippe Brouillard, Perouz Taslakian, Alexandre Lacoste, Sebastien Lachapelle, and Alexandre Drouin. Typing assumptions improve identification in causal discovery, 2022.

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, 2020.

Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. 2014.

Eunice Yuh-Jie Chen, Yujia Shen, Arthur Choi, and Adnan Darwiche. Learning bayesian networks with ancestral constraints. *Advances in Neural Information Processing Systems*, 29, 2016.

Lyuzhou Chen, Taiyu Ban, Xiangyu Wang, Derui Lyu, and Huanhuan Chen. Mitigating prior errors in causal structure learning: Towards llm driven prior knowledge, 2023.

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

Kai-Hendrik Cohrs, Gherardo Varando, Emiliano Diaz, Vasileios Sitokonstantinou, and Gustau Camps-Valls. Large language models for constrained-based causal discovery. *arXiv preprint arXiv:2406.07378*, 2024.

Anthony C. Constantinou, Zhigao Guo, and Neville K. Kitson. The impact of prior knowledge on causal structure learning, 2023.

Tiago da Silva, Eliezer Silva, Adèle Ribeiro, António Góis, Dominik Heider, Samuel Kaski, and Diego Mesquita. Human-in-the-loop causal discovery under latent confounding using ancestral gflownets. *arXiv preprint arXiv:2309.12032*, 2023.

Victor-Alexandru Darvariu, Stephen Hailes, and Mirco Musolesi. Large language models are effective priors for causal graph discovery, 2024.

Luis M de Campos and Javier G Castellano. Bayesian network learning algorithms using structural restrictions. *International Journal of Approximate Reasoning*, 45(2):233–254, 2007.

Moein Falahatgar, Ayush Jain, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. The limits of maxing, ranking, and preference learning. In *International conference on machine learning*, pp. 1427–1436. PMLR, 2018.

Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

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: Assessing causal reasoning in language models, 2024a.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation?, 2024b.

Elahe Khatibi, Mahyar Abbasian, Zhongqi Yang, Iman Azimi, and Amir M Rahmani. Alcm: Autonomous llm-augmented causal discovery framework. *arXiv preprint arXiv:2405.01744*, 2024.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality, 2023.

Andrew Li and Peter Beek. Bayesian network structure learning with side constraints. In *International conference on probabilistic graphical models*, pp. 225–236. PMLR, 2018.

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

Stephanie Long, Tibor Schuster, and Alexandre Piché. Can large language models build causal graphs?, 2024.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer, 2018.

Sara Magliacane, Tom Claassen, and Joris M. Mooij. Ancestral causal inference, 2017.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. *Advances in neural information processing systems*, 36, 2024.

Cheng Mao, Jonathan Weed, and Philippe Rigollet. Minimax rates and efficient algorithms for noisy sorting, 2017.

Christopher Meek. Causal inference and causal explanation with background knowledge, 2013.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert, 2021.

Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya K Menon, Ankit Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. *Advances in Neural Information Processing Systems*, 35: 29292–29304, 2022.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International conference on machine learning*, pp. 118–126. PMLR, 2014.

Wenbo Ren, Jia Liu, and Ness B. Shroff. On sample complexity upper and lower bounds for exact ranking from noisy comparisons, 2021.

Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33 (1):65–117, 1998.

Donald J Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, 15:657–680, 1987.

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.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12 (Apr):1225–1248, 2011.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

Natasa Tagasovska, Valérie Chavez-Demoulin, and Thibault Vatter. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery, 2020.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*, 2023.

Thomas Verma and J. Pearl. Equivalence and synthesis of causal models. *Probabilistic and Causal Inference*, 1990.

Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality?, 2022.

Moritz Willig, Matej Zecevic, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *preprint*, 8, 2023.

Yanming Zhang, Brette Fitzgibbon, Dino Garofolo, Akshith Kota, Eric Papenhausen, and Klaus Mueller. An explainable ai approach to large language model assisted causal model auditing and development, 2023.

A PROOF OF LEMMA 2.1

The result follows from following the proof of Theorem 6 from Mao et al. (2024), except that we change the upper bound of $\mathcal{C}_{L_{\text{def}}}(h, r, x) - \inf_{r \in \mathcal{R}} \mathcal{C}_{L_{\text{def}}}(h, r, x)$, where for any loss $\ell(g, x, y)$ where g is a predictor, $\mathcal{C}_\ell(g, x) := \mathbb{E}_{y|x}[\ell(g, x, y)]$. We aim to prove that

$$\mathcal{C}_{L_{\text{def}}}(h, r, x) - \inf_{r \in \mathcal{R}} \mathcal{C}_{L_{\text{def}}}(h, r, x) \leq \begin{cases} \Gamma_2\left(\mathcal{C}_{L_h}(r, x) - \inf_{r \in \mathcal{R}} \mathcal{C}_{L_h}(r, x)\right) & \text{if } \Gamma_2 \text{ is linear} \\ n_e \Gamma_2(\mathcal{C}_{L_h}(r, x) - \inf_{r \in \mathcal{R}} \mathcal{C}_{L_h}(r, x)) & \text{otherwise} \end{cases}$$

Let $\mathcal{X}_\circ := \{x \quad | \quad h(x) = h_j(x) \ \ \forall j = 1, \ldots, n_e\}$ and $x \in \mathcal{X}$. If $x \in \mathcal{X}_\circ$ then it turns out that the bound holds since the LHS is zero, as $L_{\text{def}}(h, r, x, y) = \mathbb{I}_{h(x) \neq y}$ does not depend on r, and the RHS is non-negative. Now assume $x \notin \mathcal{X}_\circ$. Then, with the convention that $h_0 = h$ and $c_0(x, y) = \mathbb{I}_{h(x) \neq y}$, for any y, $\sum_{j=0}^{n_e} \bar{c}_j(x, y) \geq 1$, as each each $\bar{c}_j(x, y)$ is binary and, from y being binary, all $\bar{c}_j(x, y)$'s being zero would imply that $\forall j = 0, \ldots, n_e, \ \ h_j(x) = 1 - y$, which contradicts $x \notin \mathcal{X}_\circ$. Similarly, $\sum_{j=0}^{n_e} \bar{c}_j(x, y) \leq n_e$ as at least one $\bar{c}_j(x, y)$ should be zero (this does not require y being binary). Then, we can repeat the steps of the original proof of Mao et al. (2024) to obtain the upper-bound, using $1 \leq \mathbb{E}_{y|x}\left[\sum_{j=0}^{n_e} \bar{c}_j(x, y)\right] \leq n_e$. This completes the general consistency bound.

For the part where $\mathcal{H}$ is a singleton h_0, this follows from the argument of the Γ_1 function being zero as a result of the singleton.

B TÜBINGEN PAIRS BY DOMAIN AND TRAINING OR TESTING SET

Table 3: Tübingen pairs, denoted by their numerical identifiers, for each domain and training/testing set.

	Training set	Testing set
Biology	7, 9, 70, 78, 79, 90, 92	5, 6, 8, 10, 11, 80, 89, 91
Climate/Environment	1, 3, 4, 13, 15, 19, 21, 42, 48, 50, 72, 77, 82, 83, 94, 95	2, 14, 16, 20, 43, 44, 45, 46, 49, 51, 69, 73, 81, 87, 93, 96
Economics/Finance	12, 47, 57, 58, 60, 61, 62, 63, 67, 68, 86	17, 56, 59, 64, 65, 66, 74, 75, 76, 84, 99
Medicine	18, 22, 34, 36, 39, 40, 88, 107	23, 24, 33, 35, 37, 38, 41, 85
Physics	26, 28, 30, 31, 32, 97, 103, 104	25, 27, 29, 98, 100, 101, 102, 106, 108

C HYPOTHESIS TESTING TO ASSESS DOMAIN CONSISTENCY

To take account of uncertainty, for fixed r and Expert, we assessed Equation 6 by performing a statistical test with $H^1_{d_+, d_-, r, \text{Expert}}$ as the alternative hypothesis and

$$H^0_{d_+, d_-, r, \text{Expert}} : p(\text{Expert chosen by r}|\text{Expert}, d_+) \leq p(\text{Expert chosen by r}|\text{Expert}, d_-)$$

as the null hypothesis for each strong/weak domain pair (d_+, d_-). This was done using Fisher's exact test over the sets of binary values I_{d^+} and I_{d^-} where

$$I_d := (1_{\{r(x_i; \text{CD}, \text{Expert}, \epsilon') = \text{Expert}\}})_{i \in \mathcal{T}_d, \text{CD}, \epsilon'},$$

with $\mathcal{T}_d$ denoting the intersection of the testing set and the domain d. This yields a p-value $\text{pval}(d_+, d_-, r, \text{Expert})$. Then, we assess domain consistency of r for Expert by computing a p-value $\text{pval}(r, \text{Expert})$ for the null hypothesis $H^0_{r, \text{Expert}} = \bigcup_{d_+, d_-,} H^0_{d_+, d_-, r, \text{Expert}}$ and the alternative hypothesis $H^1_{r, \text{Expert}} = \bigcap_{d_+, d_-,} H^1_{d_+, d_-, r, \text{Expert}}$. This is a classical instance of an intersection-union test (Berger & Hsu, 1996) and for its p-value we can take $\text{pval}(r, \text{Expert}) = \max_{d_+, d_-} \text{pval}(d_+, d_-, r, \text{Expert})$ (Schuirmann, 1987). In the end, domain consistency of r for Expert is defined as $\text{pval}_{\text{corrected}}(r, \text{Expert}) < 0.05$, where we adjust all p-values $\text{pval}(r, \text{Expert})$ jointly using Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) to obtain $\text{pval}_{\text{corrected}}(r, \text{Expert})$.