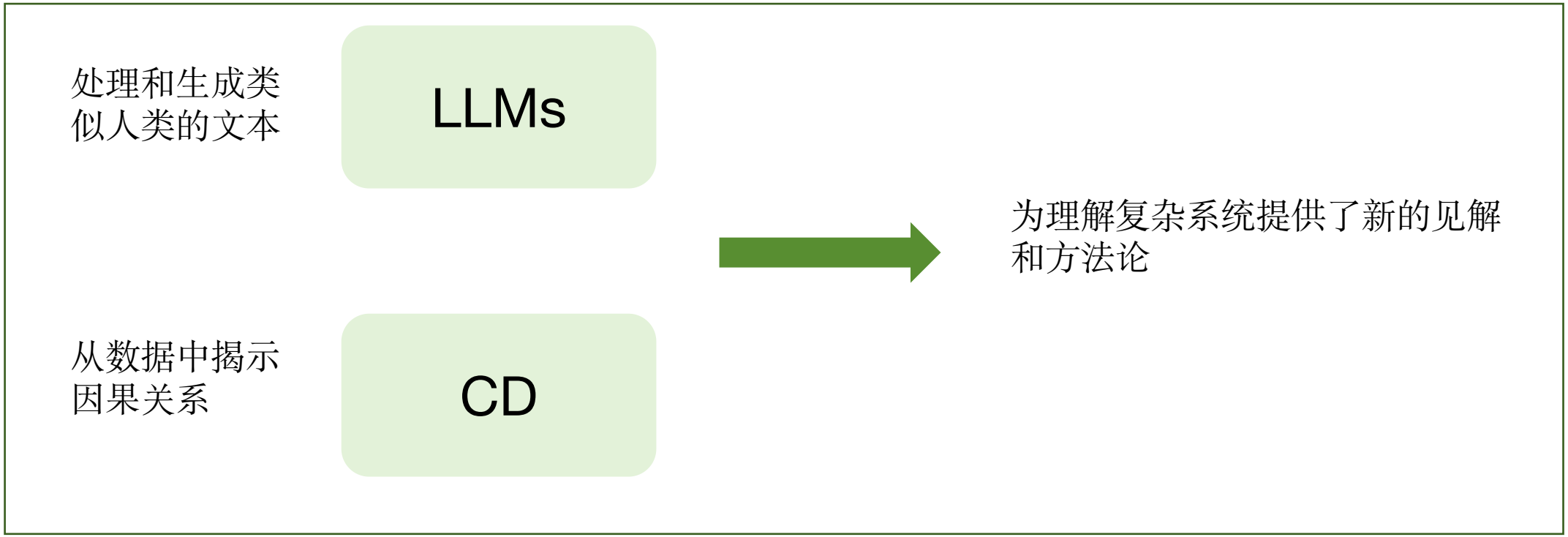
LLMs & 因果发现（CD）

汇报人：王子天

日期：2024.3.15

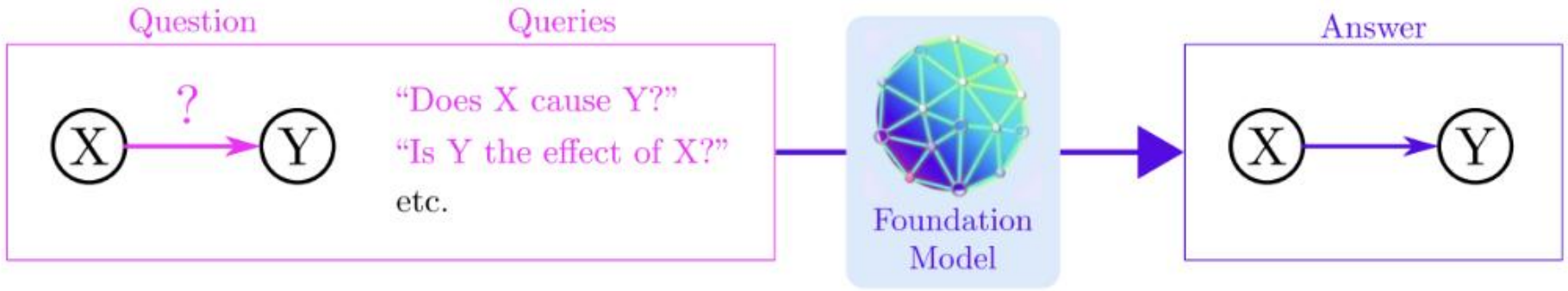
Bridging Causal Discovery and Large Language Models: A Comprehensive Survey of Integrative Approaches and Future Directions

- 研究院校：University of Virginia & University of Alberta



探索方向--Pair-wise Discovery

- 直接查询LLMs关于变量对之间因果关系。
- 这个查询针对每对变量在两个可能的方向上进行（因果和反因果学习），以聚合一个综合的准确性度量。

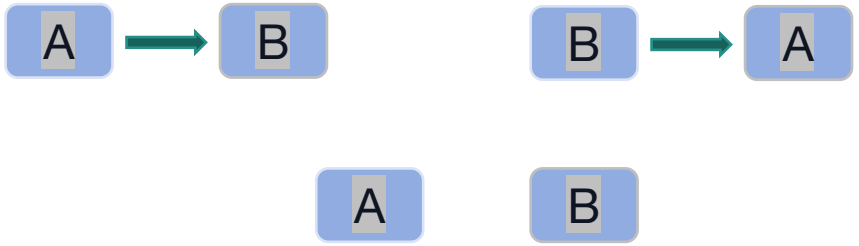


缺点：如果仅依赖于成对发现，可能会忽略变量如何相互作用的更广泛背景，这可能导致不完整或误导性的结论

探索方向-- Full Graph Discovery

- 因果结构学习（CSL），或称全图发现，用于识别系统内不同因素之间相互影响。
- 与简单确定变量对之间的因果连接的任务不同，CSL旨在绘制出整个因果关系网络。

Iterated Pair-wise Discovery. 当面对一组变量时，一个自然的方法是生成每一对可能的变量的目录，并反复进行成对分析，以获取最终的图。



缺点：

1.LLMs可能无法有效地识别变量之间的直接因果链接

2.计算负担大

现有方法--基于知识的因果发现

- 与传统的依靠数据分析统计模式的因果发现技术不同，LLMs引入了基于知识的因果发现的概念。

问题设置：给定一组描述性文本T = {t1, t2, . . . , tn}，用于一组变量X = {x1, x2, . . . , xn}，LLMs可以在充分理解每个变量的概念后执行因果发现。将确定的因果陈述表示为S = {(xi, xj)}，其中(xi, xj)表示xi导致xj。

Tasks	**Prompt**
Pairwise Discovery	”Which is more likely to be true: (A) lung cancer causes cigarette smoking, or (B) cigarette smoking causes lung cancer?”
Conditional Independence Set Test	As an expert in a specific field, you’re asked to assess the statistical independence between two variables, potentially conditioned on another variable set. Your response, based on theoretical knowledge, should be a binary guess (YES or NO) and the probability of its correctness, formatted as: [ANSWER (PROBABILITY%)]. For example, [YES (70%)] or [NO (30%)].
Full Graph Discovery	As a domain expert, analyze cause-and-effect relationships among variables with given abbreviations and values. Interpret each variable and present the causal relationships as a directed graph, using edges to denote direct causality, e.g., $x_{i1} \rightarrow x_{j1}, \ldots, x_{im} \rightarrow x_{jm}$.

现有方法--数据驱动的因果发现

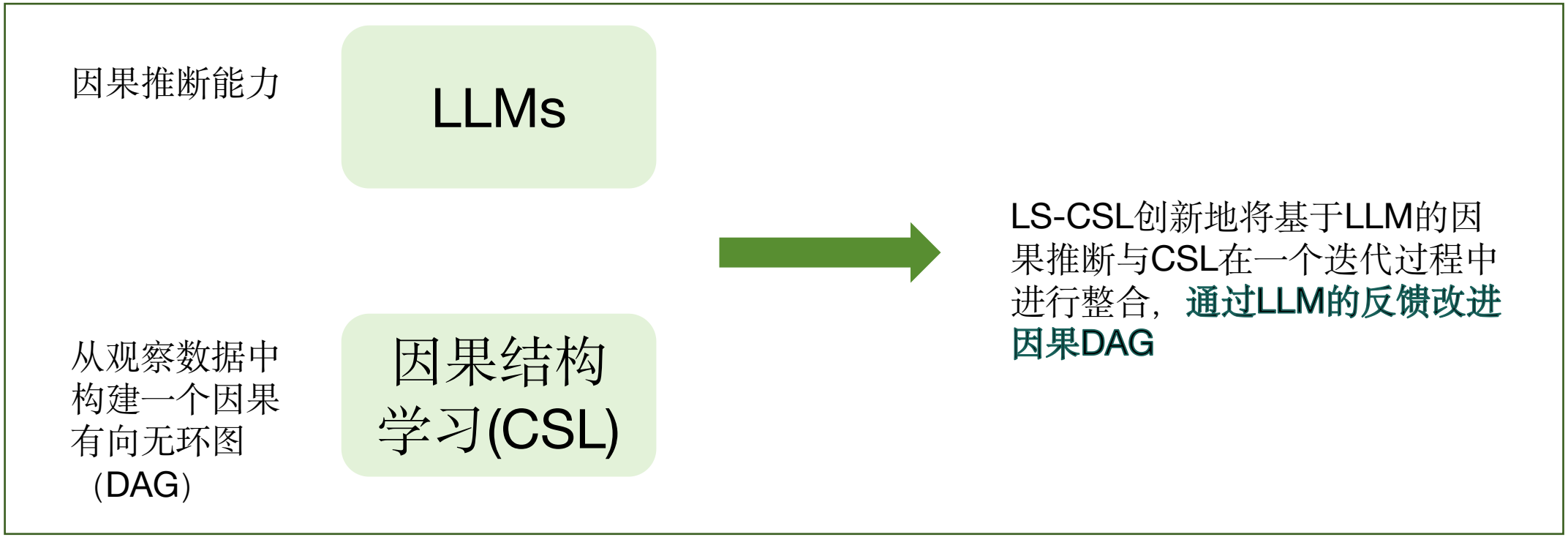
- Constraint-based
- 算法理念：利用条件独立性测试来排除不太可能的变量对的因果关系，从而更容易找到最终的因果图。

- Score-based
- 算法理念：确定最大化特定评分标准的有向无环图（DAG）。

最新的进展，由[Ban等人，2023a]提出的迭代LLM监督的CSL框架（ILS-CSL）
将LLM能力与因果结构学习相结合迈出了重要的一步。

Causal Structure Learning Supervised by Large Language Model

- 研究院校：中科院　提出了迭代LLM监督的CSL（ILS-CSL）框架



ILS-CSL算法

Algorithm 1 LLM supervised CSL

Require: Observed data, $\mathbf{D}$; Textual descriptions, $\mathbf{T}$
Ensure: Causal DAG, $\mathcal{G}$

Initialize the set of structural constraints, $\lambda \leftarrow \{\}$
repeat
  $\mathcal{G} \leftarrow \arg\max_{\mathcal{G}} \sigma(\mathcal{G}; \mathbf{D})$, s.t. $\mathcal{G} \in \text{DAG}, \mathcal{G} \models \lambda$
  for $X_i \to X_j \in E(\mathcal{G})$ **do**
    $c \leftarrow$ LLM infers causality between X_i and X_j based on $\mathbf{T}$
    if c is $X_i \leftarrow X_j$ **then**
      $\lambda \leftarrow \lambda \cup \{X_j \to X_i\}$
    end if
    if c is $X_i \nleftrightarrow X_j$ **then**
      $\lambda \leftarrow \lambda \cup \{X_i \not\to X_j, X_j \not\to X_i\}$
    end if
  end for
until no new constraints are added
return $\mathcal{G}$

- 对于每条有向边Xi → Xj和相关的文本描述T = {tf，ti，tj}，LLM被提示为：

```
You are an expert on t_f. There are two
factors:X_i : t_i, X_j : t_j.
Which cause-and-effect relationship is more likely
for following causal statements for V1 and V2?
A.changing V1 causes a change in V2.
B.changing V2 causes a change in V1.
C.changes in V1 and in V2 are not correlated.
D.uncertain.
Provide your final answer within the tags
<Answer>A/B/C/D</Answer>.
Analyze the statement:X_i X_j.
```

- tf：研究领域；ti和tj分别描述了Xi和Xj。
- 根据LLM对此提示的响应，可以得到A、B、C或D中的一个答案。

ILS-CSL算法

Algorithm 1 LLM supervised CSL

Require: Observed data, $\mathbf{D}$; Textual descriptions, $\mathbf{T}$
Ensure: Causal DAG, $\mathcal{G}$

Initialize the set of structural constraints, $\lambda \leftarrow \{\}$
repeat
  $\mathcal{G} \leftarrow \arg\max_{\mathcal{G}} \sigma(\mathcal{G}; \mathbf{D}),$ s.t. $\mathcal{G} \in \text{DAG}, \mathcal{G} \models \lambda$
  for $X_i \to X_j \in E(\mathcal{G})$ **do**
    $c \leftarrow$ LLM infers causality between X_i and X_j based on $\mathbf{T}$
    if c is $X_i \leftarrow X_j$ **then**
      $\lambda \leftarrow \lambda \cup \{X_j \to X_i\}$
    end if
    if c is $X_i \nleftrightarrow X_j$ **then**
      $\lambda \leftarrow \lambda \cup \{X_i \not\to X_j, X_j \not\to X_i\}$
    end if
  end for
until no new constraints are added
return $\mathcal{G}$

答案	约束
B（反向）	指定存在Xj → Xi
C（无因果关系）	指定Xi ↮ Xj
D（不确定）或A（正确）	不指定约束

- 指定已经从数据中发现的边的存在通常不会增强CSL，并可能无意中导致错误。
- 例如，如果真实结构是Xi ⇝ Xj而不是直接的，Xi ↛ Xj，LLM很容易推断出Xi引起Xj，因为它在区分直接因果关系和间接因果关系的短板。如果我们指定Xi → Xj，则会引入一个错误的边。

ILS-CSL算法

Algorithm 1 LLM supervised CSL

Require: Observed data, $\mathbf{D}$; Textual descriptions, $\mathbf{T}$
Ensure: Causal DAG, $\mathcal{G}$

Initialize the set of structural constraints, $\lambda \leftarrow \{\}$
repeat
    $\mathcal{G} \leftarrow \arg\max_{\mathcal{G}} \sigma(\mathcal{G}; \mathbf{D}),\ \text{s.t. } \mathcal{G} \in \text{DAG}, \mathcal{G} \models \lambda$
    for $X_i \to X_j \in E(\mathcal{G})$ **do**
        $c \leftarrow$ LLM infers causality between X_i and X_j based on $\mathbf{T}$
        if c is $X_i \leftarrow X_j$ **then**
            $\lambda \leftarrow \lambda \cup \{X_j \to X_i\}$
        end if
        if c is $X_i \nleftrightarrow X_j$ **then**
            $\lambda \leftarrow \lambda \cup \{X_i \not\to X_j, X_j \not\to X_i\}$
        end if
    end for
until no new constraints are added
return $\mathcal{G}$

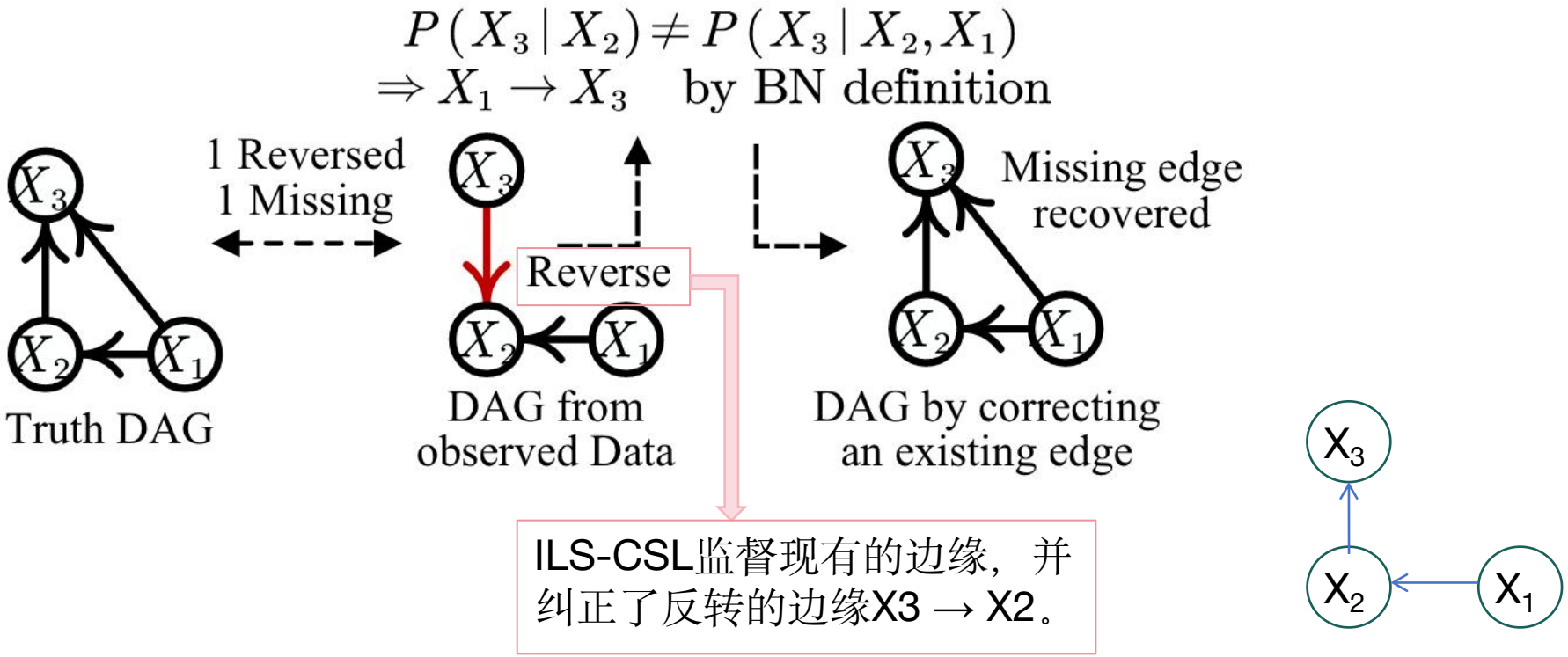
- 从LLM监督中获得的结构约束λ，将它们整合到CSL过程的下一次迭代中，采用硬约束或软约束方法。如果没有指定新的约束，则该过程终止。

	优点	缺点
硬约束方法	从准确的约束中获得更大的好处	有误差敏感性的风险
软约束方法	提供了一定程度上抵抗潜在不准确性的能力	可能并不总是遵循所有正确的约束

关键问题分析

关键问题：先验约束是否可以间接影响和修正那些不直接受此知识支配的边缘？

- 应用的先验知识纠正因果结构的能力



因果	描述
额外因果（p_e）	给定一个因果陈述（X1，X2），如果真实的因果DAG既不包含路径X1 ⇝ X2也不包含路径X2 ⇝ X1，那么它是一个额外因果的实例。
反转因果（p_r）	给定一个因果陈述（X1，X2），如果真实的因果DAG包含路径X2 ⇝ X1，那么它是一个反转因果的实例。
反转直接因果（p_r^d）	给定一个因果陈述（X1，X2），如果真实的因果DAG有一个边X2 → X1，那么它是一个反转直接因果的实例。
缺失直接因果（p_m^d）	如果真实的因果DAG中存在边X1 → X2或X2 → X1，但推断出X1和X2没有因果关系，那么它是一个缺失直接因果的实例。
正确的现有因果（p_c）	给定一个因果陈述（X1，X2），如果路径X1 ⇝ X2在真实的因果DAG中存在，那么它是一个正确的现有因果的实例。

定义了基于LLM的因果推断中的五种情况

关键问题分析

关键问题：估计并比较了ILS-CSL产生的错误约束数量与现有方法中所有成对变量的完全推断产生的错误约束数量。

假设：1）当满足相应的结构时，五种因果的概率是相同的；2）真实DAG和学习到到DAG都是稀疏的。

$\gamma_1 \binom{N}{2}$ 真实DAG中没有连接路径的节点对的数量

学习到的因果DAG中的边数

$$E_{\text{full}} = (p_e\gamma_1 + p_r(1-\gamma_1))\binom{N}{2} \qquad E_{\text{ours}} \leq \left((p_r^d + p_m^d)z_1 + p_m^d z_2 + (p_r + p_c P_{R|E})z_3\right)\gamma_2 N$$

额外因果+反转因果

正确发现的边上的错误约束包括反转和缺失的直接因果

从错误边推断因果产生的错误约束：在反转边上缺失直接因果+在额外边上额外推断的直接因果

z1：正确识别的边的比例，z2：反转的边的比例，z3：不存在于真实DAG中的额外边的比例

$P_{R|E}$：对于学习DAG中的额外边X1 → X2，ground truth存在一个反转路径X2 ⇝ X1的概率

$$E_{\text{ours}} \approx 0.10N, E_{\text{full}} \approx 0.36\binom{N}{2}, \frac{E_{\text{ours}}}{E_{\text{full}}} \approx \frac{1}{1.8(N-1)}$$

相对于完全的成对变量推断，ILS-CSL通过减少约1.8(N - 1)倍的不完美LLM推断引起的错误约束数量，从而显着降低了错误数量

GPT-4在这些样本上在不同数据集上的准确率以及反转推断的比例，类型1和2的真实答案是A，类型3的真实答案是C。

Dataset	Alarm	Asia	Insurance	Mildew	Child	Cancer	Water	Barley
Direct causality (Acc_1 /Rev_1)	1.00 / 0.00	1.00 / 0.00	0.85 / 0.05	0.95 / 0.05	1.00 / 0.00	1.00 / 0.00	0.95 / 0.05	0.70 / 0.05
Indirect causality (Acc_2 /Rev_2)	0.65 / 0.15	1.00 / 0.00	0.95 / 0.05	1.00 / 0.00	0.50 / 0.40	1.00 / 0.00	0.50 / 0.50	0.30 / 0.30
No causality (Acc_3)	0.60	0.80	0.35	0.10	0.50	0.00	0.45	0.50
Qualitative causality(Acc_4 / Rev_4)	0.72 / 0.12	1.00 / 0.00	0.92 / 0.05	0.99 / 0.01	0.70 / 0.24	1.00 / 0.00	0.67 / 0.33	0.36 / 0.26

成对变量名称	描述
直接边（直接因果）	抽样具有真实DAG中直接边Xi → Xj的成对变量
间接路径（间接因果）	抽样没有直接边但具有一个有向路径的成对变量，Xi ↛ Xj，Xi ⇝ Xj。
未连接（无因果）	抽样没有任何路径的成对变量，Xi /⇝ Xj，Xj /⇝ Xi。
定性因果	Acc4 = (Acc1 × |E|+Acc2 × |P|)/(|E|+ |P|) |E|和|P|分别表示真实因果DAG中的边缘数量和间接路径数量

1) Extra causality: $p_e = 1 - \text{Acc}_3 = 0.56$
2) Reversed causality: $p_r = \text{Rev}_4 = 0.15$
3) Reversed direct causality: $p_r^d = \text{Rev}_1 = 0.03$
4) Missing direct causality: $p_m^d = 1 - \text{Acc}_1 - \text{Rev}_1 = 0.05$
5) Correct existing causality: $p_c = \text{Acc}_4 = 0.75$

GPT-4推断的主要错误源于额外因果，这是因为一些直觉相关的概念在具有特定条件的实验中可能不会产生真正的因果关系。

实验

- RQ1：ILS-CSL能否提升基于数据的CSL基线并超越现有的基于LLM驱动的CSL方法？

Dataset	Cancer		Asia		Child		Insurance	
N	250	1000	250	1000	500	2000	500	2000
MINOBSx	0.75±0.22	0.46±0.29	0.52±0.32	0.31±0.07	0.38±0.08	0.21±0.04	0.46±0.05	0.29±0.02
+sepLLM-hard	0.13 -83%	0.00 -100%	0.27 -48%	0.04 -87%	0.42 +11%	0.31 +48%	0.91 +98%	0.60 +107%
+ILS-CSL-hard	0.50±0.22 -33%	0.29±0.29 -37%	0.42±0.37 -19%	0.15±0.15 -52%	0.25±0.06 -34%	0.07±0.03 -67%	0.42±0.03 -9%	0.28±0.06 -3%
CaMML	0.75±0.00	0.62±0.14	0.58±0.29	0.27±0.05	0.25±0.03	0.09±0.04	0.69±0.04	0.61±0.15
+sepLLM-soft	0.50 -33%	0.33 -47%	0.02 -97%	0.00 -100%	0.19 -24%	0.04 -56%	1.00 +45%	0.82 +34%
+ILS-CSL-soft	0.75±0.00 +0%	0.33±0.20 -47%	0.23±0.09 -60%	0.15±0.18 -44%	0.17±0.05 -32%	0.04±0.00 -56%	0.47±0.04 -32%	0.47±0.11 -23%

Dataset	Alarm		Mildew		Water		Barley	
N	1000	4000	8000	32000	1000	4000	2000	8000
MINOBSx	0.21±0.06	0.14±0.04	0.50±0.02	0.46±0.05	0.77±0.07	0.61±0.04	0.56±0.04	0.40±0.03
+sepLLM-hard	0.27 +29%	0.19 +36%	0.88 +76%	0.47 +2%	1.01 +31%	0.84 +38%	0.62 +11%	0.65 +62%
+ILS-CSL-hard	0.09±0.03 -57%	0.08±0.02 -43%	0.43±0.00 -14%	0.33±0.18 -28%	0.68±0.05 -12%	0.56±0.02 -8%	0.54±0.02 -4%	0.38±0.02 -5%
CaMML	0.24±0.05	0.18±0.06	1.20±0.10	1.30±0.12	0.88±0.08	0.81±0.04	0.96±0.07	0.96±0.10
+sepLLM-soft	0.13 -46%	0.07 -61%	1.07 -11%	1.30 +0%	0.89 +1%	0.73 -10%	0.98 +2%	0.98 +2%
+ILS-CSL-soft	0.08±0.01 -67%	0.06±0.01 -67%	1.01±0.07 -16%	1.26±0.05 -3%	0.70±0.02 -20%	0.63±0.04 -22%	0.90±0.06 -6%	0.83±0.06 -14%

结果以标准化的SHD（较低的值更好）呈现

- RQ2：在不同的基础算法中，ILS-CSL能否持续提高因果结构的质量？软约束和硬约束哪个更好？

使用不同的评分函数，BDeu和BIC评分，以及搜索算法，MINOBSx和HC

Dataset	Cancer		Asia		Child		Insurance	
N	250	1000	250	1000	500	2000	500	2000
HC-BDeu	0.58±0.13	0.33±0.26	0.56±0.27	0.23±0.17	0.57±0.12	0.49±0.18	0.69±0.06	0.68±0.09
+ILS-CSL-hard	0.50±0.22 -14%	0.29±0.29 -12%	0.46±0.33 -18%	0.15±0.15 -35%	0.24±0.07 -58%	0.10±0.02 -80%	0.45±0.06 -35%	0.34±0.04 -50%
+ILS-CSL-soft	0.50±0.22 -14%	0.29±0.29 -12%	0.44±0.30 -21%	0.15±0.15 -35%	0.26±0.06 -54%	0.11±0.03 -78%	0.50±0.08 -28%	0.35±0.04 -49%
MINOBSx-BDeu	0.75±0.22	0.46±0.29	0.52±0.32	0.31±0.07	0.38±0.08	0.21±0.04	0.46±0.05	0.29±0.02
+ILS-CSL-hard	0.50±0.22 -33%	0.29±0.29 -37%	0.42±0.37 -19%	0.15±0.15 -52%	0.25±0.06 -34%	0.07±0.03 -67%	0.42±0.03 -9%	0.28±0.06 -3%
+ILS-CSL-soft	0.50±0.22 -33%	0.29±0.29 -37%	0.42±0.37 -19%	0.15±0.15 -52%	0.25±0.04 -34%	0.08±0.04 -62%	0.41±0.03 -11%	0.26±0.04 -10%
HC-BIC	0.92±0.29	0.62±0.34	0.48±0.36	0.31±0.29	0.53±0.07	0.38±0.16	0.76±0.05	0.72±0.06
+ILS-CSL-hard	0.92±0.29 +0%	0.42±0.34 -32%	0.33±0.25 -31%	0.19±0.17 -39%	0.26±0.07 -51%	0.07±0.03 -82%	0.60±0.03 -21%	0.41±0.03 -43%
+ILS-CSL-soft	0.92±0.29 +0%	0.42±0.34 -32%	0.35±0.26 -27%	0.21±0.19 -32%	0.27±0.08 -49%	0.07±0.05 -82%	0.62±0.06 -18%	0.42±0.03 -42%
MINOBSx-BIC	1.00±0.25	0.62±0.21	0.46±0.23	0.27±0.05	0.34±0.06	0.18±0.04	0.62±0.05	0.55±0.05
+ILS-CSL-hard	0.92±0.29 -8%	0.38±0.26 -39%	0.42±0.40 -9%	0.12±0.08 -56%	0.24±0.08 -29%	0.06±0.02 -67%	0.55±0.03 -11%	0.39±0.08 -29%
+ILS-CSL-soft	0.92±0.29 -8%	0.38±0.26 -39%	0.35±0.26 -24%	0.15±0.12 -44%	0.25±0.05 -26%	0.06±0.02 -67%	0.55±0.03 -11%	0.41±0.09 -25%

Dataset	Alarm		Mildew		Water		Barley	
N	1000	4000	8000	32000	1000	4000	2000	8000
HC-BDeu	0.65±0.12	0.64±0.09	0.79±0.11	0.99±0.07	0.76±0.07	0.64±0.08	0.80±0.06	0.65±0.06
+ILS-CSL-hard	0.12±0.02 -82%	0.08±0.01 -88%	0.46±0.01 -42%	0.22±0.02 -78%	0.64±0.02 -16%	0.55±0.03 -14%	0.69±0.06 -14%	0.57±0.06 -12%
+ILS-CSL-soft	0.30±0.05 -54%	0.25±0.06 -61%	0.43±0.00 -46%	0.47±0.04 -53%	0.64±0.01 -16%	0.56±0.03 -12%	0.76±0.04 -5%	0.62±0.03 -5%
MINOBSx-BDeu	0.21±0.06	0.14±0.04	0.50±0.02	0.46±0.05	0.77±0.07	0.61±0.04	0.56±0.04	0.40±0.03
+ILS-CSL-hard	0.09±0.03 -57%	0.08±0.02 -43%	0.43±0.00 -14%	0.33±0.18 -28%	0.68±0.05 -12%	0.56±0.02 -8%	0.54±0.02 -4%	0.38±0.02 -5%
+ILS-CSL-soft	0.09±0.02 -57%	0.07±0.01 -50%	0.47±0.01 -6%	0.37±0.02 -20%	0.68±0.04 -12%	0.56±0.02 -8%	0.55±0.03 -2%	0.38±0.02 -5%
HC-BIC	0.68±0.05	0.59±0.10	0.90±0.06	0.91±0.13	0.76±0.04	0.70±0.03	0.87±0.05	0.80±0.08
+ILS-CSL-hard	0.22±0.04 -68%	0.12±0.04 -80%	0.58±0.01 -36%	0.46±0.04 -49%	0.69±0.02 -9%	0.61±0.03 -13%	0.76±0.02 -13%	0.69±0.06 -14%
+ILS-CSL-soft	0.41±0.04 -40%	0.35±0.11 -41%	0.71±0.01 -21%	0.57±0.02 -37%	0.69±0.02 -9%	0.61±0.03 -13%	0.82±0.04 -6%	0.74±0.09 -8%
MINOBSx-BIC	0.32±0.08	0.15±0.04	0.74±0.01	0.73±0.09	0.82±0.03	0.77±0.03	0.79±0.04	0.58±0.03
+ILS-CSL-hard	0.16±0.07 -50%	0.09±0.03 -40%	0.58±0.01 -22%	0.45±0.03 -38%	0.69±0.03 -16%	0.62±0.01 -19%	0.73±0.03 -8%	0.55±0.03 -5%
+ILS-CSL-soft	0.19±0.06 -41%	0.10±0.01 -33%	0.73±0.01 -1%	0.64±0.04 -12%	0.70±0.02 -15%	0.64±0.02 -17%	0.76±0.02 -4%	0.56±0.03 -3%

- RQ3:LLM监督因果发现的过程是如何详细展开的？

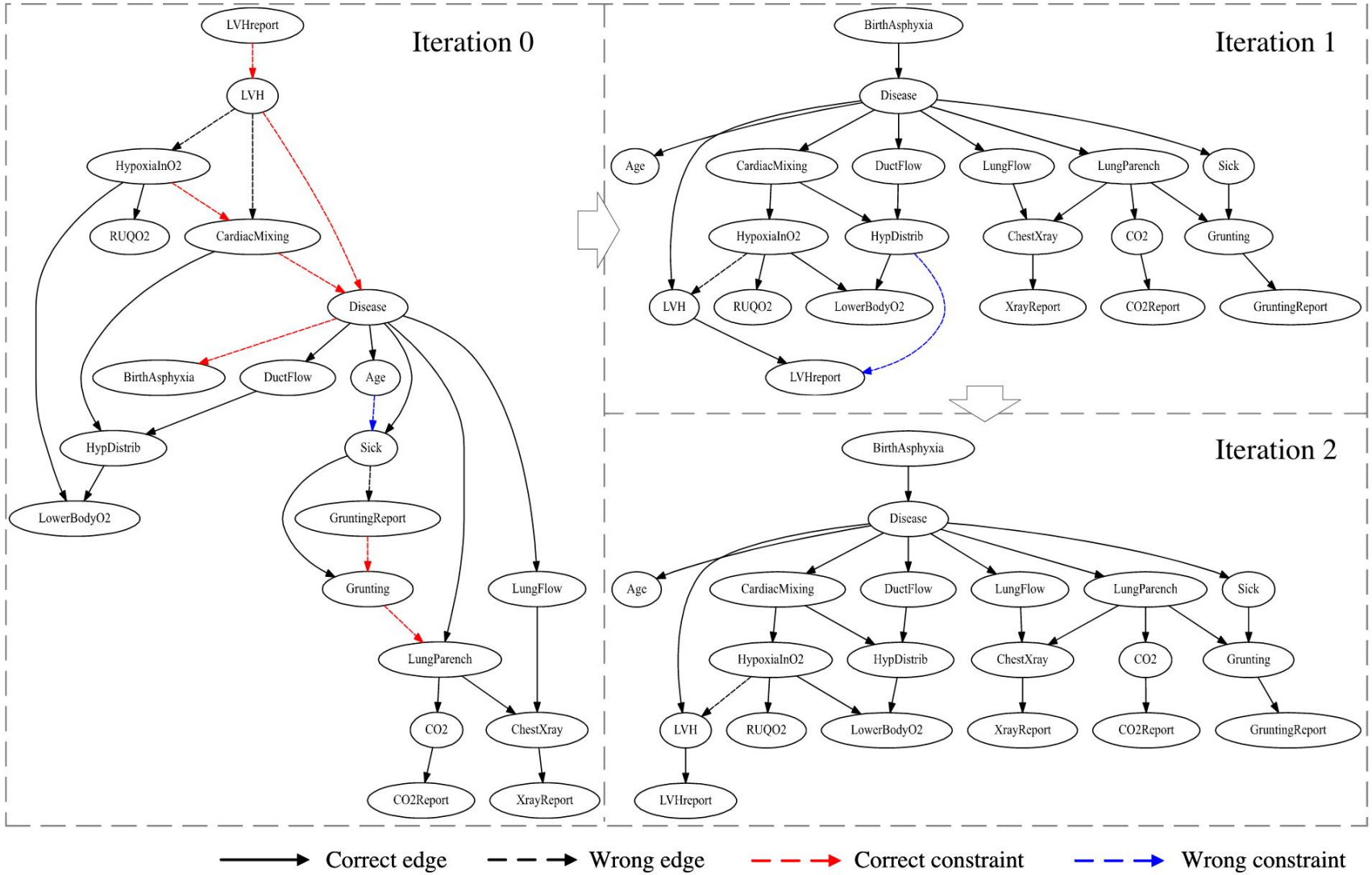


Fig. 4: Visualized process of HC-BDeu+ILS-CSL-hard on a set of observed data of *Child*, 2000 samples. The SHD of iterations are: 12 for Iteration 0, 3 for Iterations 1 and 2.

最初，HC (BDue)从纯观测数据(迭代0)中学习因果DAG，其边由LLM监督，导致LLM在不一致的推断边上产生边缘约束(彩色箭头)。

这些约束可以改进本地结构（红色箭头），也可能因错误推断而带来损害（蓝色箭头）。随着迭代的进行，错误边缘（虚线箭头）逐渐减少。

局限性 LLMs&CD

- 符号和数学理解

大型语言模型在高级符号推理方面存在局限性，研究人员观察到大型语言模型在理解必要以找到现有因果解决方案的假设时可能会犯错，限制源自语言模型的基础架构，其主要定制是用于识别模式而不是进行算法或逻辑操作。因此，这些模型容易出现计算和逻辑错误，可能会影响结果的完整性。

- 幻觉
- 鲁棒性

在预测因果图的任务中，简单地改变提问方式可能会导致生成的因果图显著不同。

- 因果关系的理解

LLMs无法严格按照正式的因果推理实践解决因果问题。

causal parrot

- [Zeˇcevi´c等，2023]提出的基准数据集提供了明确的基准真实因果图和相关的自然语言语句，直接将LLMs暴露于文本“因果事实的相关性”，他们引入了一种称为“元结构因果模型”的形式化方法，用于编码关于其他因果模型的因果事实。中心推测是，LLMs通过利用因果查询和事实陈述之间的相关性，使它们显得具有因果关系，“模仿鹦鹉”的行为。

未来方向

- Specialized LLMs
- Causal LLM Agent
- New Benchmark Dataset