# 因果推理数据集概览

## Causal-Reasoning-Dataset-Collection

Lin Ren / 任林

18/11/2025

Southeast University
Institute of Cognitive Intelligence (COIN)

# Outline / 总目录

# 数据集总览 / Dataset Overview

| 类别 | 数据集 | 任务形式 | 评价侧重点 |
|---|---|---|---|
| **Causal Inference** | cladder | 单分类（两类） | 区分相关/干预/反事实能力 |
| | copa | 单分类（两类） | 语义常识下的因果方向判别 |
| **Causal Discovery** | corr2cause | 多分类（三类） | 相关与因果的区分能力 |
| | crab | 排序 | 因果强度与成对一致性 |
| | crass | 多分类（四类） | 反事实场景下的因果判断 |
| | e_care | 单分类（两类） | 临床因果选择（原因/结果） |
| | pain | 单分类（两类） | 临床因果真假判定 |
| **Additional Causal Tasks** | moca | 单分类（两类） | 社会规范下的因果归因 |
| | tram | 单分类（两类） | 现代语境的原因/结果选择 |

注：此处“单分类（两类）”与“多分类（三/四类）”均为单选题，区别仅在于标签类别数量的多少。

# 常识相关性 / Commonsense Relevance

| 数据集 | 任务形式 | 备注 |
|---|---|---|
| 近似纯符号 / 结构化推理（弱常识依赖） | | |
| cladder | 单分类（两类） | Pearl 因果阶梯上的结构化因果推理 |
| corr2cause | 多分类（三类） | 相关 vs 因果的关系类型识别 |
| 显式依赖语义常识 / 领域背景 | | |
| copa | 单分类（两类） | 日常语境下的因果方向判别 |
| crab | 排序 | 因果强度排序与成对一致性 |
| crass | 多分类（四类） | 反事实场景下的常识性因果判断 |
| e_care | 单分类（两类） | 临床场景下的因果选择（原因/结果） |
| pain | 单分类（两类） | 临床因果陈述真伪判断 |
| moca | 单分类（两类） | 道德与常识相关的因果归因 |
| tram | 单分类（两类） | 现代语境下的原因/结果选择 |

上半部分数据集主要考察「给定结构/符号下的因果推理」，下半部分数据集则显式依赖语义常识、临床知识或社会规范。

Part. 1

# 因果推断 / Causal Inference

# Outline

# cladder: Causal Ladder Reasoning (1/2)

任务定义 / **Task Definition**

- 测试语言模型在 Pearl 因果阶梯三个层级上的推理能力：

$$\text{Association} \rightarrow \text{Intervention} \rightarrow \text{Counterfactual}$$

- 每条样本由前提 $x$ 与因果询问 $q$ 组成，要求判断查询陈述是否成立。

输入 / 输出格式

- 输入：$x =$ context, $q =$ causal query
- 输出标签集合：

$$Y = \{\texttt{yes}, \texttt{no}\} \quad \text{或} \quad Y = \{A, B\}$$

形式化定义

- 判定函数：

$$f(x, q) = y, \quad y \in Y$$

- 每条样本带有整数阶梯标签 rung $\in \{1, 2, 3\}$，大致对应相关性 (association)、干预 (intervention)、反事实 (counterfactual)，并结合 `query_type`（如 `marginal` / `ate` / `det-counterfactual` 等）做细粒度评估。

# cladder: Causal Ladder Reasoning (2/2)

**示例 / Example**

**Prompt**

“Husband has a direct effect on wife and alarm clock … If we disregard the mediation effect through wife, would husband positively affect alarm clock?”

---

**Label**

`yes`

标签与阶梯

- 任务类型：相关性 / 干预 / 反事实三类因果问题。
- 选项数量：2 个 (yes/no 或 A/B)。

# cladder: Dataset Statistics (3/3)

整体规模与分布

- 样本总数：$N = 10112$。
- 每条样本带有阶梯标签 rung $\in \{1, 2, 3\}$ 以及细粒度 `query_type`。

| | Total | Rung 1 | Rung 2 | Rung 3 |
|---|---|---|---|---|
| Size | | | | |
| # Samples | 10,112 | 3,160 | 3,160 | 3,792 |
| Question | | | | |
| # Sentences/Sample | 6.01 | 5.88 | 5.37 | 6.65 |
| # Words/Sample | 80.9 | 73.43 | 76.95 | 90.42 |
| # Nodes/Graph | 3.52 | 3.5 | 3.5 | 3.54 |
| # Edges/Graph | 3.38 | 3.3 | 3.3 | 3.5 |
| Answer | | | | |
| Positive Class (%) | 50 | 50 | 50 | 50 |
| Explanations | | | | |
| # Sentences/Sample | 9.11 | 9.1 | 8.1 | 9.96 |
| # Words/Sample | 47.95 | 49.87 | 32.8 | 58.97 |

Table 1: Statistics of our CLADDER dataset v1.5.

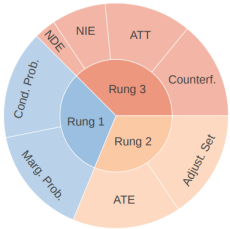


Figure 3: Distributions of query types in our 10K data.

# copa: Everyday Causal QA (1/2)

任务定义

- 给定前提句 $p$ 和询问方向（`cause`/`effect`），在两个候选句中选出更合理的原因或结果：

$$y \in \{A, B\}$$

输入 / 输出格式

- 输入：$p = \text{Premise},\ q \in \{\texttt{cause?}, \texttt{effect?}\},\ \text{Options} = \{A, B\}$
- 输出：$Y = \{A, B\}$

形式化定义

- 决策函数：

$$f(p, q, A, B) = y, \quad y = \arg\max_{c \in \{A,B\}} P(c \mid p, q)$$

# copa: Everyday Causal QA (2/2)

示例

**Premise**

“The office was closed.”

**Asks-for**

`cause`

**Options**

`A) It was a holiday.  B) It was summer.`

**Gold**

`A`

选项与标签

- 选项数量：2 (A/B)，输出集合 $Y = \{A, B\}$。
- 可分别统计 `cause` / `effect` 两种设问的准确率。

Part. 2

# 因果发现 / Causal Discovery

# Outline

# corr2cause: Correlation vs. Causation (1/2)

任务定义

- 输入为相关性叙述 $x$ 与因果假设 $h$，模型需判断二者关系：

$$y \in \{\texttt{entailment}, \texttt{neutral}, \texttt{contradiction}\}$$

输入 / 输出格式

- 输入：$x =$ correlation statement, $h =$ causal hypothesis
- 输出：$Y = \{\text{entailment}, \text{neutral}, \text{contradiction}\}$

形式化定义

- 自然语言推理式三分类：

$$f(x, h) = y, \quad y \in Y$$

# corr2cause: Correlation vs. Causation (2/2)

**示例**

**Premise**
“Suppose there is a closed system of 2 variables, A and B. … A correlates with B.”

---

**Hypothesis**
“A directly affects B.”

---

**Relation**
`neutral`

标签与选项

- 三类标签: `entailment`, `neutral`, `contradiction`。
- 考察模型是否能抗拒“把相关当因果”的倾向。

# crab: Graded and Pairwise Causality (1/2)

任务定义

- 给定结果事件 $e$ 与候选因果事件集合 $C = \{c_1, \ldots, c_k\}$，评估因果强度或比较哪一个 cause 更强。

输入 / 输出格式

- Graded：输入 $(e, C)$，输出每个 $c_i$ 的强度等级

$$y_i \in \{\texttt{High}, \texttt{Medium}, \texttt{Low}, \texttt{None}\}$$

- Pairwise：输入 $(e, c_i, c_j)$，输出

$$y \in \{c_i \succ c_j,\ c_j \succ c_i,\ \texttt{tie}\}$$

形式化定义

- Graded: $f_{\text{graded}}(e, c_i) = s_i$
- Pairwise: $f_{\text{pair}}(e, c_i, c_j) = y$

# crab: Graded and Pairwise Causality (2/2)

**示例 (graded_causality)**

**Effect**
“The U.K. broke its national record for the highest temperature ever registered.”

---

**Candidates**
多个与全球变暖、极端高温相关的事件，各自带有人类因果得分（如 `score_c = 90`）。

---

**Task**
模型需选择对该效果最具因果解释力的事件。

标签与评估

- Graded: 4 级强度标签; Pairwise: 成对比较的一致性指标等。

# crass: Counterfactual Causal QA (1/2)

任务定义

- 给定实际事件 $e$ 和一个反事实询问 $q$，从四个选项中选择最合理的反事实结果：

$$y \in \{1, 2, 3, 4\}$$

输入 / 输出格式

- 输入：$e = \text{actual event},\ q = \text{counterfactual question},\ \text{Options} = \{o_1, o_2, o_3, o_4\}$
- 输出：$Y = \{1, 2, 3, 4\}$

形式化定义

- 决策函数：

$$f(e, q, o_1, \ldots, o_4) = y, \quad y = \arg\max_i P(o_i \mid e, q)$$

# crass: Counterfactual Causal QA (2/2)

**示例**

**Input**

“A woman opens a treasure chest. What would have happened if the woman had not opened the treasure chest?”

---

**Options**

```
(1) The treasure chest would have been open. (2) That is not possible.
(3) The treasure chest would have remained closed. (4) I don't know.
```

---

**Gold**

```
(3)
```

选项与标签

- 4 选项多选题，包含“不可能”“我不知道”等不确定性表达。

# e_care: Clinical Causal Judgement (1/2)

任务定义

- 临床语境的因果选择任务：给定前提句 $p$ 与设问方向（`cause/effect`），在两个候选断言中选择更合理的原因或结果。

输入 / 输出格式

- 输入：$p = \text{Premise},\ q \in \{\texttt{cause?}, \texttt{effect?}\},\ \text{Hypotheses} = \{h_1, h_2\}$
- 输出：$Y = \{1, 2\}$（或记为 $\{A, B\}$）

形式化定义

- 决策函数：

$$f(p, q, h_1, h_2) = y, \quad y = \arg\max_{c \in \{1,2\}} P(h_c \mid p, q)$$

# e_care: Clinical Causal Judgement (2/2)

示例

**Premise**

“This child caught roseola.”

---

**Ask-for**

`effect`

---

**Hypotheses**

`(1) He succeeded via conjuring up a flower.`
`(2) The child had a fever which followed by a rash.`

---

**Gold**

`(2)`

标签与难点

- 标签集合：二选一（1/2 或 A/B）；同时覆盖 `cause`/`effect` 两种设问。
- 易混淆点：非因果相关（共病/伴随症状）与真实因果的区分。

# pain: Clinical Causal Judgement (1/2)

任务定义

- 医学因果判别：判断一个临床因果断言（如“疾病 $\rightarrow$ 症状”或“治疗 $\rightarrow$ 结局”）是否为真。

输入 / 输出格式

- 输入：$x =$ clinical causal statement
- 输出：$y \in \{\texttt{true}, \texttt{false}\}$

形式化定义

- 二分类函数：

$$f(x) = y, \quad y \in \{\texttt{true}, \texttt{false}\}$$

# pain: Clinical Causal Judgement (2/2)

示例

**Query**

“L L4 Radikulopati causes L Mediala kn00e4ledsbesv00e4r. … Answer with true or false.”

---

**Answer**

`true` (`Answer = 1.0`)

标签与难点

- 标签集合：True / False；分布不均衡，更贴近真实临床统计。
- 易混淆点：共病或相关症状被误判为因果关系。

Part. 3

# 其他因果任务 / Additional Causal Tasks

# Outline

# moca: Moral and Commonsense Causality (1/2)

任务定义

- 在复杂社会/道德情景中，判断某个主体是否应被归因为造成结果的“原因”（因果责任判断）。
- 标签为二分类：

$$y \in \{\texttt{yes}, \texttt{no}\}$$

输入 / 输出格式

- 输入：长故事文本 $s$、关注主体 $a$、结果事件 $e$。
- 输出：$Y = \{\texttt{yes}, \texttt{no}\}$。

形式化定义

- 归因函数：

$$f(s, a, e) = y, \quad y \in Y$$

# moca: Moral and Commonsense Causality (2/2)

**示例**

**Story**

“Lauren and Jane share a weak computer. If two people are logged on at the same time, it usually crashes … Jane disobeys the policy and also logs on at 9:00 am. The computer crashed.”

---

**Question**

“Did Jane cause the computer to crash?”

---

**Answer**

`Yes`

标签与分析

- 关注“违反规范”与客观因果链之间的权衡。

# tram: Modern COPA-style Causal Choice (1/2)

任务定义

- `tram` 为二选一因果选择任务，给定前提 $p$ 与提问方向（原因/结果），从两个备选句中选择更合理者：

$$y \in \{A, B\}$$

- 在 COPA 设定基础上，语料更贴近现代生活，语境更复杂、句子更长。

输入 / 输出格式

- 输入：$p = \text{Premise},\ q \in \{\texttt{cause?}, \texttt{effect?}\},\ \text{Options} = \{A, B\}$
- 输出：$Y = \{A, B\}$

形式化定义

- 决策函数：

$$f_{\text{tram}}(p, q, A, B) = y, \quad y = \arg\max_{c \in \{A,B\}} P(c \mid p, q)$$

# tram: Modern COPA-style Causal Choice (2/2)

**示例**

**Premise**
“The roads were slippery this morning.”

---

**Question**
“What’s the more plausible CAUSE?”

---

**Options**
`A) There was a heatwave.  B) It had snowed overnight.`

---

**Gold**
`B`

差异与引用

- `tram` 题量更大、语境更现代，可与 COPA 对比迁移表现。

Part. 4

# 小结

# Outline

# 数据集小结

| 类别 | 数据集 | 任务形式 | 是否为符号推理 |
|---|---|---|---|
| **Causal Inference** | cladder | 单分类（两类） | 是 |
| | copa | 单分类（两类） | 否 |
| **Causal Discovery** | corr2cause | 多分类（三类） | 是 |
| | crab | 排序 | 否 |
| | crass | 多分类（四类） | 否 |
| | e_care | 单分类（两类） | 否 |
| | pain | 单分类（两类） | 否 |
| **Additional Causal Tasks** | moca | 单分类（两类） | 否 |
| | tram | 单分类（两类） | 否 |

Part. 5

# 参考文献

# Outline

# 参考文献


- Jin, Z., Chen, Y., Leeb, F., et al. (2023). CLADDER: Assessing causal reasoning in language models. *NeurIPS*.
- Roemmele, M., Bejan, C. A., & Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symp. on Logical Formalizations of Commonsense Reasoning*.
- Jin, Z., Liu, J., Lyu, Z., et al. (2023). Can large language models infer causation from correlation? *Findings of ACL 2023*.
- Romanou, A., Montariol, S., Paul, D., et al. (2023). CRAB: Assessing the strength of causal relationships between real-world events. *Findings of EMNLP 2023*.
- Frohberg, J., & Binder, F. (2022). CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *LREC 2022* (pp. 2126–2140).
- Du, L., Ding, X., Xiong, K., et al. (2022). e-CARE: A new dataset for exploring explainable causal reasoning. *Findings of ACL 2022*.
- Pain / e-CARE（同源临床因果任务）：参见上述 e-CARE 数据集条目（Du et al., 2022），Pain 任务为同源临床因果推理场景扩展。
- Nie, A., et al. (2023). MoCa: Measuring human–language model alignment on causal and moral judgment tasks. *NeurIPS*, 36.
- Wang, Y., & Zhao, Y. (2023). TRAM: Benchmarking temporal reasoning for large language models. *arXiv*:2310.00835.