Large Language Models and Causal Inference in Collaboration: A Survey

Announcement Date	@2025년 4월 10일
Conference Name	NAACL 2025
Keywords	Causal Inference LLM

Causal Inference ?

🎯 인과 추론이란?

👉 **"무언가가 다른 무언가에 영향을 줬는지 알아보는 것"**이에요.

예를 들어:

- **질문:** 감기에 걸린 이유가 비를 맞아서일까?
- 이 질문의 답을 찾는 게 바로 **인과 추론**이에요!

🔍 상관관계 vs 인과관계

많이 헷갈리는 개념인데 쉽게 구분해 볼게요:

개념	설명	예시
상관관계	A와 B가 같이 일어나는 현상 (같이 변해요)	아이스크림 판매량이 늘면 익사 사고도 늘어요 (둘 다 여름에 늘어나기 때문)
인과관계	A가 B의 **원인**이 되는 관계	비를 맞으면 몸이 젖고, 젖은 몸이 감기에 걸릴 확률을 높여요 → **비 → 감기**

> 상관관계가 있다고 해서 꼭 인과관계가 있는 건 아니에요! 그래서 우리가 '진짜 원인'을 찾기 위해 인과 추론을 하는 거죠.

🧪 어떻게 인과를 추론할까?

1. **실험** (가장 정확함):

- 집단을 나눠서 한쪽엔 약을 주고, 다른 쪽은 안 줘요.
- 이후 결과를 비교하면 약이 효과가 있는지 알 수 있어요.

2. **관찰 데이터** (실험 못할 때):
   - 그냥 관찰만으로는 원인을 단정하기 어려워요.
   - 그래서 수학적 모델, 그래프, 통계 기법 등으로 조심스럽게 인과 관계를 추정해요.

🧠 인과 추론에서 자주 쓰는 개념들 (간단 요약)

용어	쉬운 설명
처치 (Treatment)	원인이 될 수 있는 것 (예: 약 복용)
결과 (Outcome)	우리가 알고 싶은 결과 (예: 병이 나았는가)
반사실 (Counterfactual)	"약 안 먹었으면 어땠을까?" 같은 **상상 속 시나리오**
DAG 그래프	변수 간 인과 관계를 방향 있는 화살표로 나타낸 그래프
혼란변수 (Confounder)	A와 B 둘 다에 영향을 주는 숨겨진 제3의 변수 (헷갈리게 만듦)

https://yonghwankim-dev.tistory.com/222 - DAG 그래프

---

1. Introduction

- 대형 언어 모델(LLMs)은 다양한 작업에서 뛰어난 성능을 보이며, 멀티모달 모델로 확장되며 활용 가능성이 크게 증가하고 있음
- LLM의 핵심 역량은 추론 능력이며, 이를 향상시키기 위한 다양한 연구가 활발히 진행되고 있음
- 인과 추론은 NLP 모델의 정확성, 공정성, 견고성, 설명 가능성을 향상시키는 데 유용하며, 최근에는 LLM과의 통합이 주목받고 있음

- 본 연구에서는 인과 추론이 LLM 성능 향상에 어떻게 기여하는지와, 반대로 LLM이 인과 추론을 어떻게 지원할 수 있는지를 탐색

---

2. Causal Inference and Large Language Models

2.1 Evolution of the Large Language Models

- 대형 언어 모델(LLMs)은 Transformer 구조를 기반으로 발전했으며, 언어 생성과 이해 능력을 획기적으로 향상시켰음
- 모델 크기나 학습 데이터 규모를 늘릴수록, 다양한 하위 작업에서 더 우수한 성능을 보이는 경향이 확인
- 성능 향상을 위한 주요 기술로는 In-Context Learning(ICL), Chain-of-Thought(CoT) prompting, Instruction tuning 등이 사용됨
- 최근에는 이미지와 언어를 함께 처리할 수 있는 멀티모달 LLM(VLM/MLLM)도 등장하여 적용 영역이 확장되고 있음

2.2 Introduction of Causal Inference

- 인과 추론은 변수 간의 원인과 결과 관계를 식별하고 추정하는 이론적 틀로, 처치 효과(treatment effect) (**"어떤 개입(처치, treatment)이 결과에 얼마나 영향을 줬는가?"** 를 측정하는 것) 를 계산하는 데 중점을 둠
- 대표적인 접근인 잠재 결과 모델(potential outcomes framework)은 실제로 관측할 수 없는 반사실 데이터를 가정하여 인과 효과를 추정
- 인과 구조를 시각화하거나 설명하기 위해 구조 방정식 모델(SEM)과 인과 그래프(DAG)가 자주 사용됨
- 이러한 이론들은 LLM과 접목되어, 인과 관계 추론 및 자동화된 원인 분석에 응용될 수 있는 기초가 됨

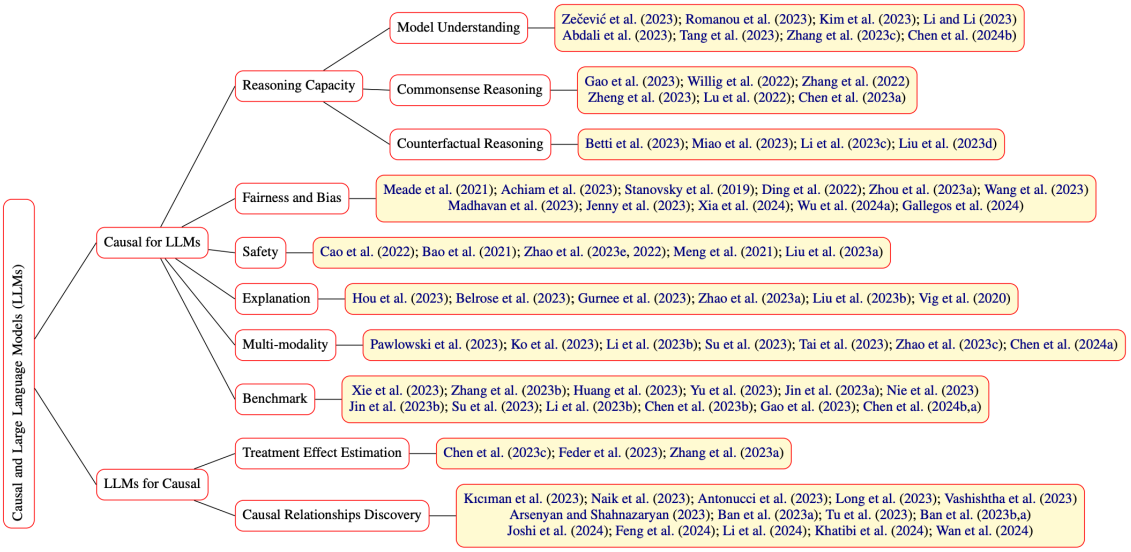

Figure 1: An overview of the interplay between causal inference frameworks and LLMs.

Causal and Large Language Models (LLMs)

```
|
├── Causal for LLMs      🔄 인과 추론이 LLM을 어떻게 향상시키는가
|
└── LLMs for Causal      🔄 LLM이 인과 추론을 어떻게 보조하는가
```

3. Causal Inference for Large Language Models

Causal for LLMs

> 인과 추론이 LLM의 성능 및 특성 향상에 어떻게 기여하는지를 다루는 분기입니다.

🔷 Reasoning Capacity (추론 능력)

- **Model Understanding (모델 이해)**

  LLM이 단순히 훈련 데이터를 기억하는 것이 아닌 **인과 구조를 이해하는지 평가**

  ➡ Zečević et al., Kim et al., Chen et al.

- **Commonsense Reasoning (상식 추론)**

  일상적 인과 관계를 통한 추론 능력

➡ Gao et al., Lu et al., Zhang et al.

- **Counterfactual Reasoning (반사실 추론)**

  "만약 ~였다면?" 같은 조건을 설정하고 모델의 인과 판단 능력을 평가

  ➡ Betti et al., Li et al.

🔷 Fairness and Bias (공정성 및 편향)

- 인과 그래프 기반 방법으로 사회적 편향(성별, 인종 등)을 **모델에서 식별하고 제거**

  ➡ Meade et al., Zhou et al., Jenny et al.

🔷 Safety (안전성)

- LLM이 불안정하거나 헛소리를 할 때, 이를 인과적으로 분석하여 **문제 원인 파악 및 방어 가능**

  ➡ Zhao et al., Cao et al.

🔷 Explanation (설명 가능성)

- 모델이 **왜 이런 출력을 했는가?**에 대해 **인과 그래프나 개입 기법을 통해 설명력 확보**

  ➡ Hou et al., Gurnee et al., Vig et al.

🔷 Multi-modality (다중모달 처리)

- 텍스트+이미지 같은 멀티모달 입력에서 **텍스트 중심 인과성만 의존하지 않도록** 조정

  ➡ Ko et al., Li et al., Chen et al.

🔷 Benchmark (평가 벤치마크)

- 위 모든 분야에 대해 인과성을 중심으로 설계된 **벤치마크 세트들**

  ➡ CRAB, CaLM, CLOMO, etc.

---

4. Large Language Models for Causal Inference

LLMs for Causal

> 반대로, LLM이 인과 추론 자체를 돕는 경우를 다루는 분기입니다.

4.1 Treatment Effect Estimation (처치 효과 추정)

LLM은 자연어 기반 정보와 반사실 시나리오 생성을 통해 처치 효과 추정에 활용될 수 있습니다.

예를 들어, DISCO와 같은 기법은 LLM이 인과 구조를 학습하고 개입의 결과를 추론하도록 돕습니다.

다만, 프롬프트 편향이나 외부 지식의 신뢰도 문제 등 한계도 존재합니다.

4.2 Causal Relationship Discovery (인과 관계 발견)

LLM은 변수 간 인과 관계를 텍스트 설명을 통해 추출하거나 DAG 구조를 직접 예측하는 데 활용됩니다.

이 방식은 전통적인 통계적 접근보다 직관적이고 빠르며, 다양한 도메인 지식과 결합이 가능합니다.

하지만 정확성과 해석 가능성 문제로 인해 **보조 도구로의 활용**이 권장됩니다.

4.3 Multimodal Causal Inference (멀티모달 인과 추론)

멀티모달 인과 추론은 이미지, 텍스트 등 다양한 모달 정보를 통합해 인과 관계를 추론하는 방식입니다.

LVLM 기반 모델은 의료, 시각 질의응답 등에서 원인과 결과 간 복잡한 상호작용을 파악하는 데 유용합니다.

아직 초기 단계지만, 산업적 활용 가능성이 높아 적극적인 연구가 진행 중입니다.

---

5. Conclusion

- LLMs와 인과 추론의 융합은 상호 보완적인 가능성을 지니며, 앞으로 더욱 정교한 프롬프트 설계, 반사실 생성, 그리고 멀티모달 인과 추론 분야에서 발전할 것임
- 특히 인과 추론 기반 LLM 성능 평가, 편향 제거, 인간과의 협력적 추론 설계는 핵심적인 연구 방향으로 제시됩니다.
- 궁극적으로 LLM의 신뢰성과 해석력을 강화하는 데 있어 인과 추론은 중요한 역할을 하게 될 것입니다.