# 대규모 언어모델과 인과추론의 융합 동향

**이현원[1], 홍원석[1], 최승훈[1], 최현수[1,*]**
[1]**서울과학기술대학교**
**{lee.hyunwon999, hongwonseok11, cshooon, *choi.hyunsoo}@seoultech.ac.kr**

# The Convergence Trends of Large Language Model and Causal Inference

**Lee Hyunwon[1], Hong Won-Seok[1], Choi Seunghun[1], and Choi Hyun-Soo[1,*]**
[1]**Seoul National University of Science and Technology**

**요 약**

본 논문은 최근 급성장 중인 대규모 언어모델(LLM)과 인과추론(Causal Inference)의 융합 연구를 체계적으로 분류하고 분석한다. Judea Pearl의 인과 사다리를 기반으로 LLM의 현재 위치와 인과적 한계를 확인하고 , 이론적 엄밀성(인과 질의 및 구조)과 방법론적 방향(작업 주체 및 목표)의 두 가지 범주로 구성된 4대 핵심 축을 제시한다. 이를 통해 기존 연구들의 인과성 적용 수준을 진단하고, LLM의 도메인 지식과 인과추론의 구조적 제약이 결합하여 상호 보완하는 연구 지형을 정의한다.

## 1. 서론

최근 LLM의 추론 능력을 고도화하거나, 인과 모델의 반사실 데이터 생성을 위해 두 도메인을 접목하는 시도가 급증하고 있다 [1][2]. 그러나 많은 연구가 엄밀한 인과 구조 없이 개념적으로만 인과성을 차용하고 있어 학술적 혼선을 야기한다. 특히 단순 상관관계를 인과성으로 과잉 주장하는 사례는 LLM의 신뢰성 확보에 걸림돌이 된다. 이에 본 논문은 LLM과 인과추론의 융합 연구를 다각도에서 분석하고, 두 도메인의 체계적 결합을 판별할 수 있는 기준을 제안한다.

## 2. 본론

### 2.1 이론적 엄밀성: 인과성 적용의 체계성

일부 융합 연구는 단순 통계적 상관관계만으로 인과적 성능 향상을 보고하여 용어 오남용의 비판을 받는다. 특히 모델 차원에서는 Pearl의 인과 사다리 1단계(관찰)만 수행하면서, 실제로는 2단계(개입)나 3단계(반사실) 추론이 가능하다고 과잉 주장하는 사례가 대표적이다 [3]. 이를 판별하기 위해 다음 두 축을 제안한다.

#### 2.1.1 축 1: 인과 질의의 명시성

해당 연구가 명시적인 인과 질의를 정의하고 있는지 분류한다. 2단계 개입 ($P(Y|do(X))$)이나 3단계 반사실 ($P(Y_{X=x'}|X=x, Y=y)$)에 해당하는 질의를 명시하지 않은 경우, LLM은 인과적 메커니즘을 차용하지 못한 채 상관관계 기반 방법으로 학습하게 된다.

#### 2.1.2 축 2: 구조적 가정의 도입 여부

구조적 인과 모형(SCM) 혹은 이에 준하는 가정을 명시적으로 도입했는지 확인한다. 방향성, 독립성 가정, 식별 가능성 등이 전개되지 않은 질의는 단순한 개념적 설명 도구에 불과하며, 엄밀한 가정과 제약조건을 통해서만 인과 질의의 식별(Identification) 작업을 수행할 수 있기 때문이다.

### 2.2 방법론적 방향: 상호보완적 결합 양상

LLM과 인과추론의 융합은 작업 주체와 상대 모델의 역할에 따라 방법론적 성격이 명확히 구분된다. 본 절에서는 두 도메인의 결합 양상과 목표를 기준으로 연구를 정리한다.

#### 2.2.1 축 3: 작업 주체에 따른 분류

**LLM 'for Causal Discovery'**는 전통적 데이터 기반 알고리즘(PC, GES, NOTEARS 등)이 변수 증가에 따라 탐색 공간이 급증하고 텍스트 메타데이터를 활용하지 못하는 한계를 LLM의 도메인 지식으로 보완하는 방식이다 [4]. 반면 **Causal Inference 'for LLM'**은 LLM에 인과 개념을 도입하여 추론 능력을 체계화하고, 허위 상관관계에 의한 성능 저하를 줄여 신뢰성과 설명 가능성을 높이는 데 주력한다 [5].

#### 2.2.2 축 4: 융합의 궁극적 목표

먼저 인과추론 모델이 주체가 되는 경우, LLM과의 융합을 통해 주로 **초기 인과 구조 발견, 상식에 기반한 잠재변수 추출, 반사실 데이터 생성 및 설명** 등 인과 파이프라인 내부의 특정 단계를 보조할 수 있다. 이들 또한 모델이 최종적으로 갖고 있는 목표인 ATE, CATE, ITE 등의 성능 향상을 위해 이를 사용한다. 반대로 LLM이 주체가 되는 경우, 인과추론 모델과의 융합 목표는 보통 **공정성, 강건성, 편향 및 환각 현상 감소** 등을 목표로 하며 인과 모형의 규제 또는 정규화 기법 등이 된다. 해당하는 목표의 성능을 향상시킴으로써, 결과적으로 예측 성능 혹은 일반화 성능 향상을 노린다.

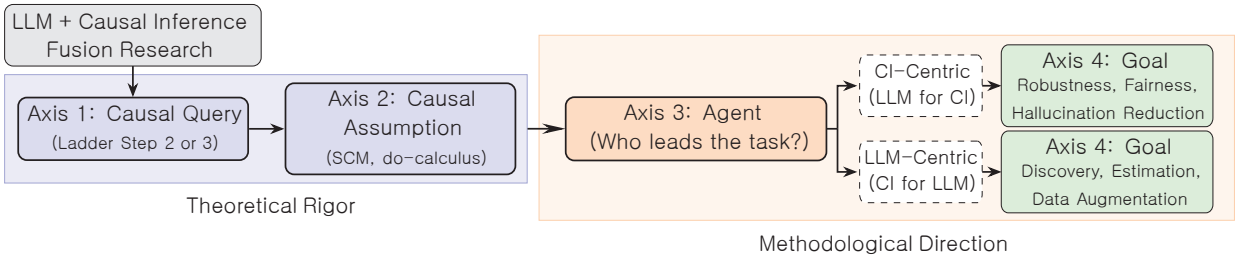


그림 1: LLM-인과추론 융합 연구 분류 프레임워크

### 2.3 제안 프레임워크 기반 연구 사례 분석

앞서 제시한 4대 축을 바탕으로 최신 융합 연구들을 분류한 결과는 표 1과 같다. 우선 **이론적 엄밀성** 측면에서, 대다수 연구가 인과 질의를 명시하고 있으나 구조적 인과 모형(SCM) 등 엄격한 제약을 도입한 사례는 상대적으로 적었다. Ma 등[2]과 Zhao 등[6]은 인과적 개념을 활용해 LLM의 편향 제거를 시도했으나 구조적 전개보다는 개념적 차용에 그친 반면, Zhou 등[7]은 환각 제어를 위해 주의 집중 메커니즘의 인과성을 분석하며 보다 구체적인 구조적 접근(Δ)을 시도하였다. **방법론적 방향**에 따라서는 작업 주체별 지향점이 명확히 구분되었다. 인과추론 모델이 주체인 연구[1]는 LLM의 풍부한 배경지식을 활용해 인과 발견의 탐색 효율을 높이는 데 집중하였다. 반대로 LLM이 주체인 연구들[2, 6, 8, 7]은 인과적 제약을 정규화 도구로 삼아 모델의 설명 가능성과 추론 강건성을 확보하는 것을 주요 목표로 삼았다. 이러한 분류 체계는 향후 LLM과 인과추론을 결합하고자 하는 연구자들에게 실질적인 가이드라인을 제공한다.

표 1: LLM과 인과추론 융합 연구 분류

| 구분 | 이론적 엄밀성 | | 방법론적 방향 | |
|---|---|---|---|---|
| | 인과 질의 여부 | 인과 구조 여부 | 작업 주체 | 융합 목표 |
| [1] | O | O | CI | Causal Discovery |
| [2] | O | X | LLM | Debiasing |
| [6] | O | X | LLM | Debiasing |
| [8] | X | X | LLM | Explainable Reasoning |
| [7] | O | △ | LLM | Hallucination Control |

## 3. 결론

본 논문에서는 LLM과 인과추론이 상호 한계를 보완할 수 있는 융합 연구의 발전 방향을 제시하였다. LLM의 지식이 인과 발견의 탐색 효율을 개선하는 양상과, 인과적 제약이 LLM의 강건성 및 공정성을 강화하는 과정을 4대 핵심 축으로 체계화하였다. 본 연구에서 제시한 분류 체계는 향후 인과적으로 엄밀하고 설명 가능한 AI 모델을 설계하려는 연구자들에게 실질적인 분석 틀로 기능할 것으로 기대된다.

## 사사

본 연구는 보건복지부의 재원으로 국립암센터 암정복추진연구개발사업 지원으로 이루어진 것임 (RS-2025-02215373)

## References

[1] Jin Li et al. “Revealing Multimodal Causality with Large Language Models”. In: *The Thirty-ninth Annual Conference on NeurIPS*. 2025.

[2] Bo Ma et al. “LLM4Rec: Large Language Models for Multimodal Generative Recommendation with Causal Debiasing”. In: *arXiv preprint arXiv:2510.01622* (2025).

[3] Luis Cavique. “Causality: the next step in artificial intelligence”. In: *Philosophy of artificial intelligence and its place in society*. IGI Global, 2023, pp. 1–17.

[4] Guangya Wan et al. “Large Language Models for Causal Discovery: Current Landscape and Future Directions”. In: *Proceedings of the Thirty-Fourth IJCAI*. Survey Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2025, pp. 10687–10695.

[5] Longxuan Yu et al. “CausalEval: Towards Better Causal Reasoning in Language Models”. In: *Proceedings of the 2025 Conference of the NAACL*. Apr. 2025, pp. 12512–12540.

[6] Shitian Zhao et al. “Causal-cog: A causal-effect look at context generation for boosting multi-modal language models”. In: *Proceedings of the IEEE/CVF Conference on CVPR*. 2024, pp. 13342–13351.

[7] Guanyu Zhou et al. “Mitigating Modality Prior-Induced Hallucinations in Multimodal Large Language Models via Deciphering Attention Causality”. In: *The Thirteenth ICLR*. 2025.

[8] Xinmeng Hou et al. “DriveAgent: Multi-Agent Structured Reasoning With LLM and Multimodal Sensor Fusion for Autonomous Driving”. In: *IEEE Robotics and Automation Letters* 10.11 (2025), pp. 12189–12196.