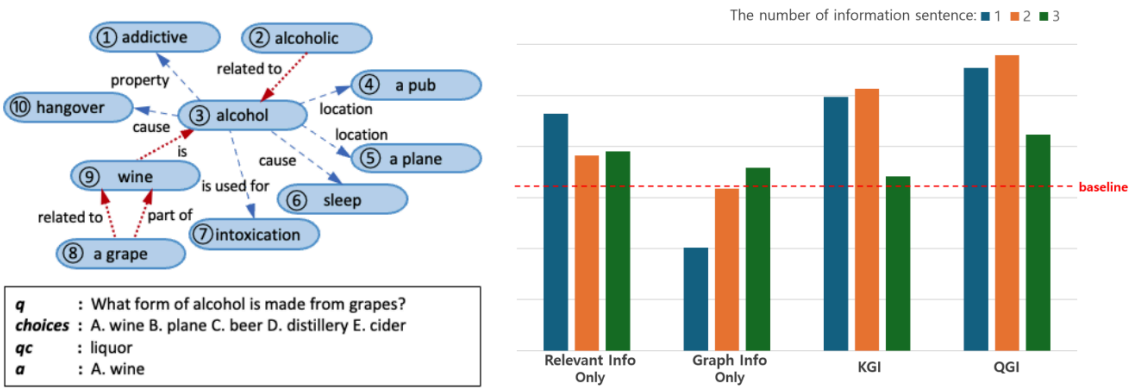
# Causal Reasoning in Large Language Models: A Knowledge Graph Approach

| | |
|---|---|
| Announcement Date | @2025년 4월 24일 |
| Conference Name | NeurIPS-CaLM 2024 |
| Keywords | Knowledge Graph LLM RAG |

## Abstract


- 대형 언어 모델(LLM)은 유사한 정보를 검색하거나 구조화된 프롬프트를 활용하여 성능을 향상시킵니다.
- 하지만 이 두 접근법 중 어떤 방식이 더 효과적인지는 명확하지 않으며, 본 논문은 이에 대한 해답을 제시하고자 합니다.
- 연구팀은 지식 그래프(KG)를 활용한 랜덤 워크 추론 방식을 제안하여 상식적 질문 응답(CommonsenseQA) 문제에 적용했습니다.
- 이 방식은 인과 관계에 기반한 구조화된 정보 흐름을 통해 LLM의 추론 능력을 향상시킵니다.
- 특히 무관해 보이는 문장도 KG 기반으로 활용할 경우 성능을 높이는 데 기여한다는 결과는 인과적 구조의 중요성을 시사합니다.




(a) KG example of QA (b) KG-Based Reasoning

Figure 1: (a) Illustration of a KG structure and an example of CommonsenseQA (Talmor et al., 2018). At the bottom, the question and its concept are represented as $q$ and $qc$ respectively, while the answer is denoted as $a$. (b) Performance comparison on commonsense QA. The dotted line (baseline) represents the performance when no additional context is provided through the prompt. The three bars represent the number of sentences provided as context in the prompt.

# (a) KG example of QA — 지식 그래프 기반 질문 예시

## 🔷 구성 요소 설명

- **노드(Node)**: 각 파란색 타원은 하나의 개념(예: *alcohol, wine, grape* 등)을 나타냅니다.
- **엣지(Edge)**: 노드 간의 선은 개념들 사이의 **관계(relation)**를 나타내며, 방향성과 관계 명칭이 표시되어 있어요.
  - 예: `alcohol → causes → sleep`
  - 관계 예시: `related to`, `part of`, `is`, `cause`, `property`, `location`

## 🔷 하단 질문 예시 (QA)

- **q (질문)**: What form of alcohol is made from grapes?

  → 포도(grape)로 만든 술의 형태는?
- **choices (선지)**: A. wine, B. plane, C. beer, D. distillery, E. cider
- **qc (질문 개념, question concept)**: liquor
- **a (정답)**: A. wine

## 🔷 설명

- 질문 개념인 "liquor"에 가장 유사한 노드는 **alcohol (노드 3)**입니다.

- 여기서 출발하여 랜덤 워크 기반으로 탐색한 결과:
  - **alcohol → part of → wine**
  - **wine → related to → grape**
- 이 경로를 통해 모델은 "**wine은 alcohol의 일종이며 포도와 관련이 있다**"는 인과적·연결적 정보를 추론할 수 있게 됩니다.

---

# (b) KG-Based Reasoning — 성능 비교 그래프

## 🔷 구성 요소

- **x축**: 프롬프트 구성 방식
  - **Relevant Info Only**: 유사 문장만 사용 (정보 검색 기반)
  - **Graph Info Only**: 인과 그래프 기반 추론만 사용
  - **KGI (Keyword + Graph Inference)**: 키워드 관련 정보 + 그래프 기반 추론
  - **QGI (Query + Graph Inference)**: 질문 자체에 관련된 정보 + 그래프 기반 추론
- **y축**: 모델의 정확도 (Accuracy)
- **막대 색상**:
  - 파랑 (1): 문장 1개 사용
  - 주황 (2): 문장 2개 사용
  - 초록 (3): 문장 3개 사용
- **빨간 점선 (baseline)**: 아무 정보도 추가하지 않았을 때의 정확도

## 🔷 해석

- *QGI (질문 관련 + 그래프 추론)**가 가장 좋은 성능을 보임.
- **Graph Info Only**도 어느 정도 효과가 있음 — 특히 문장 2~3개일 때 **정보가 질문과 직접 관련이 없어도** 성능이 향상됨.
- **Relevant Info Only**는 문장 수가 늘어날수록 효과가 떨어지는 경향이 있음.
- 흥미롭게도, 문장이 2개일 때 (주황색) 대부분의 방법에서 가장 좋은 성능을 보임 — 정보가 많다고 항상 좋은 건 아님.

# 2. Method

## 2.1 Problem Formulation

1. 이 연구는 주어진 질문과 질문 개념(q, qc)에 대해 올바른 답변 a를 선택하는 다지선다형 QA 문제를 다룹니다.
2. 질문 개념 qc는 질문의 핵심 키워드를 의미하며, 지식 그래프(KG)는 이 키워드와 가장 유사한 노드를 출발점으로 사용합니다.
3. 지식 그래프는 노드들(V)과 관계들(E)로 구성된 이종 그래프이며, 관계는 방향성을 가집니다.
4. 시작 노드(vqc)에서부터 n-hop 이웃들을 따라가면서 관련 정보를 수집합니다.
5. 이 과정을 통해 LLM이 구조화된 추론 과정을 따라갈 수 있도록 문맥을 제공합니다.

## 2.2 Retrieval-Based Prompting

1. 전통적인 RAG 방식은 질문과 유사한 문서를 검색하여 모델의 입력 프롬프트에 추가합니다.
2. 문서 검색은 질문과 문서의 임베딩 간 유사도(예: 코사인 유사도)를 기반으로 top-k 문서를 선택합니다.
3. 선택된 문서는 생성기 모델이 정답을 생성할 때의 참고 문맥이 됩니다.
4. 이 방식은 의미적으로 관련 있는 문장을 제공하지만, 인과적 추론 구조는 없습니다.
5. 본 논문에서는 이 방식과 KG 기반 추론 방식을 성능 비교합니다.

## 2.3 KG-Based Random-Walk Reasoning

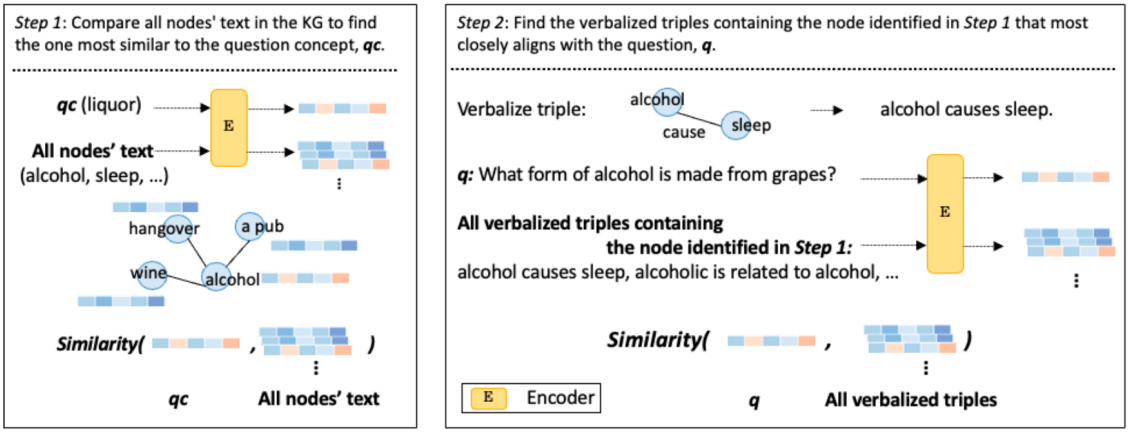



Figure 2: Detailed process of prompting through the KG-based reasoning.

1. KG 기반 추론은 질문 개념 qc에 가장 유사한 노드를 찾고, 이 노드 주변을 탐색합니다.
2. 연결된 노드 간의 관계를 삼중항(triple)으로 표현하고, 자연어 문장으로 변환해 문맥으로 사용합니다.
3. 질문과 이 문장들의 유사도를 비교하여 가장 관련 있는 문장을 선택합니다.
4. 이후에는 유사도가 아니라 **무작위 방식(random-walk)**으로 연결된 노드들을 따라가며 추론 문맥을 확장합니다.
5. 이 과정은 LLM이 단순한 정보 검색을 넘어서 인과적 관계를 고려한 추론을 수행하도록 돕습니다.

# 3. Experiment

## 3.1 Experimental Setup

1. 이 연구는 LLM의 학습 없이 제안된 **지식 그래프 기반 추론 방식**이 실제로 효과적인지를 **Llama 2-Chat** 모델을 활용해 **제로샷(zero-shot)** 환경에서 검증합니다.
2. 평가에는 **CommonsenseQA** 데이터셋의 검증용 데이터 1,221개가 사용되었으며, 검색과 추론 모두에 **ConceptNet** 지식 그래프가 활용되었습니다.
3. 검색에 사용할 삼중항(triple)은 자연어 문장으로 변환되어 총 **3,423,004개의 문장 데이터**로 구성되었고, **Wikipedia** 데이터셋도 비교 실험에 활용되었습니다.
4. 모든 문장과 질문은 **e5-base** 텍스트 임베딩 모델을 이용해 벡터로 변환되어 유사도 검색이나 노드 선택에 사용되었습니다.

5. 평가 기준은 **정확도(accuracy)**이며, 다양한 응답 형식이나 표현 변형을 고려해 정답 판단 시 유연한 기준을 적용했습니다.

## 3.2 Result

1. 실험 결과, KG 기반 랜덤 워크 추론 방식은 RAG보다 **일관되게 더 높은 정확도**를 보이며 성능 향상에 효과적임이 입증되었습니다.
2. 특히, 질문과 직접 관련 없는 정보라도 **인과 구조에 따라 연결된 문장들**을 제공하면 성능이 유의미하게 개선되는 결과가 나타났습니다.
3. 정보의 양이 많다고 항상 좋은 것이 아니며, **2개의 문장을 포함한 프롬프트가 가장 높은 성능을 보인 경우가 많았습니다.**
4. Wikipedia 기반 RAG는 ConceptNet 기반 방식보다 효과가 낮았고, 프롬프트에서 **문서가 질문보다 앞에 배치될 때 성능이 더 우수**했습니다.
5. 전반적으로, 정보의 연관성뿐만 아니라 **추론 경로의 방향성과 프롬프트 구성 순서**가 성능에 중요한 영향을 미친다는 사실이 확인되었습니다.

Table 1: Performance comparison of RAG and KG-based reasoning. For a clear explanation of indicating node location, we assume node 1 is the most similar to the question concept and form the graph sequence as 5 -> 4 -> 1 -> 2 -> 3 ($k$: the number of sentences combined with a question to generate an answer). The highest performance is denoted in bold and the second best results are underlined.

| Type | $k$ | Node Location | Acc. |
|---|---|---|---|
| Baseline | 0 | - | 0.5684 |
| Relevant Information Only | 1 | top-1 triple | 0.5864 |
| | 2 | top-2 triples | 0.5782 |
| | 3 | top-3 triples | 0.5790 |
| Keyword Relevant Information + Graph Inference (KGI) | 1 | 1 -> 2 | 0.5897 |
| | | 4 -> 1 | 0.5766 |
| | 2 | (1 -> 2, 2 -> 3) | <u>0.5913</u> |
| | | (5 -> 4, 4 -> 1) | 0.5577 |
| | | (4 -> 1, 1 -> 2) | <u>0.5913</u> |
| | 3 | (5 -> 4, 4 -> 1, 1 -> 2) | 0.5741 |
| | | (4 -> 1, 1 -> 2, 2 -> 3) | 0.5741 |
| Query Relevant Information + Graph Inference (QGI) | 1 + 2 | top-1 + (4 -> 1, 1 -> 2) | **0.5979** |

Table 2: Performance in situations where the provided information has lower relevance to the question. (R: relevance of information; if "Y," we remain node 1 as the most similar node and randomly select triples from node 1; otherwise, we opt for an unrelated node randomly). The highest performance is denoted in bold and the second best results are underlined.

| Type | $k$ | R | Node Location | Acc. |
|---|---|---|---|---|
| Baseline | 0 | - | - | 0.5684 |
| Irrelevant Information Only | 1 | N | 1 irrelevant triple | 0.5356 |
| | 2 | N | 2 irrelevant triples | 0.5324 |
| | 3 | N | 3 irrelevant triples | 0.5397 |
| Graph Inference Only | 1 | N | 1 -> 2 | 0.5602 |
| | | N | 4 -> 1 | 0.5602 |
| | | Y | 1 -> 2 | 0.5479 |
| | | Y | 4 -> 1 | 0.5659 |
| | 2 | N | (1 -> 2, 2 -> 3) | <u>0.5717</u> |
| | | N | (5 -> 4, 4 -> 1) | 0.5635 |
| | | N | (4 -> 1, 1 -> 2) | 0.5561 |
| | 3 | N | (5 -> 4, 4 -> 1, 1 -> 2) | 0.5667 |
| | | N | (4 -> 1, 1 -> 2, 2 -> 3) | **0.5758** |

Table 3: Evaluating performance variations across various prompt configurations (Prompt Engineering: order of prompt, direction of reasoning information).

| Type | $k$ | Prompt Engineering | Acc. |
|---|---|---|---|
| Relevant Information Only with the ConceptNet Graph | 1 | documents -> question | 0.5864 |
| | | question -> documents | 0.5455 |
| | 4 | documents -> question | 0.5635 |
| Relevant Information Only with Wikipedia | 1 | documents -> question | 0.5455 |
| | 2 | documents -> question | 0.5504 |
| | 3 | documents -> question | 0.5463 |
| Graph | 2 | irregular direction (1 -> 2, 4 -> 1) | 0.5807 |
| | | regular direction (4 -> 1, 1 -> 2) | 0.5913 |
| | 4 | (5 -> 4, 4 -> 1, 1 -> 2, 2 -> 3) | 0.5717 |

# 4. Conclusion

1. 본 연구는 LLM 성능 향상을 위해 **의미 기반 정보 검색(RAG)**과 **지식 그래프 기반 인과 추론**의 상대적 효과를 비교하였고, **인과 추론 방식이 더 우수한 성능을 보임**을 실험으로 입증했습니다.
2. 특히, **직접 관련이 없는 정보**라도 **인과 관계에 기반한 추론 문맥**으로 제공하면 성능이 개선된다는 예상 밖의 결과를 확인했습니다.
3. 이 결과는 상식 기반 질의응답에서 **추론 능력 강화가 단순 정보 제공보다 더 중요한 요소**일 수 있음을 시사하며, 향후 관련 작업에도 유의미한 방향성을 제공합니다.