결정론적 계획 알고리즘을 위한 신뢰성 있는 LLM 설명 생성: Trace-정렬 기반 접근

이수린, SM Wahidur Rahman, 이흥노
광주과학기술원

leesurin@gm.gist.ac.kr, sm.wahidur@gm.gist.ac.kr, heungno@gist.ac.kr

Faithful LLM-based Explanation Generation for Deterministic Planning Algorithms: A Trace-Aligned Approach

Surin Lee, SM Wahidur Rahman, Heung-No Lee
GIST

요 약

본 논문은 결정론적 계획 알고리즘을 대상으로 한 LLM 기반 설명 생성의 한계를 분석하고, 실행 Trace 에 정렬된 설명 생성 접근을 제안한다. 기존 LLM 설명은 실제 연산 과정과 일치하지 않는 사후적 서술을 생성한다는 문제가 있다. 이를 해결하기 위해 알고리즘 실행 과정에서 생성되는 Trace 를 기반으로 설명을 생성하고, 설명과 연산 단계 간의 정렬(Alignment)을 강제하는 제약 조건을 도입한다. 제안한 접근 방식은 특정 알고리즘에 종속되지 않으며, 다양한 결정론적 계획 문제에 범용적으로 적용 가능하다.

Ⅰ. 서 론

최근 대규모 언어 모델(LLM, Large Language Model)의 발전으로 복잡한 의사결정 과정을 자연어로 설명하는 시스템에 대한 관심이 급증하고 있다[1]. 특히 생산 계획, 자원 할당, 스케줄링과 같은 계획(Planning) 문제에서는 단순한 연산 결과뿐만 아니라, 해당 결과가 어떤 논리적 과정에 의해 도출되었는지를 사용자가 이해할 수 있는 형태로 설명하는 기능이 필수적이다. 이러한 설명 기능은 의사결정의 신뢰성을 높이고, 실무자가 결과를 검증하거나 수정하는 데 중요한 역할을 수행한다[2][3].

그러나 기존의 LLM 기반 설명 생성 방식은 주로 통계적 예측 모델이나 분류 모델을 위해 설계되어, 결정론적(deterministic) 구조를 갖는 계획 알고리즘에 적용하기에는 근본적인 한계가 있다. 결정론적 계획 알고리즘은 입력에 따라 출력이 유일하게 결정되며, 명시적인 상태 전이와 연산 단계의 연쇄를 통해 결과가 도출된다. 반면, LLM 은 이러한 내부 연산 과정을 직접 참조하지 않고 최종 출력값만을 바탕으로 사후적(post-hoc) 설명을 생성하는 경향이 있다. 이로 인해 실제 연산 과정과 괴리된 설명이나 검증 불가능한 인과 관계가 서술되는 문제가 발생한다[1].

기존 설명가능 인공지능(eXplainable AI, XAI) 연구인 SHAP 이나 LIME 등은 통계적 모델의 예측 결과에 대한 특징(Feature) 기여도를 해석하는 데 초점을 맞추고 있어, 명시적인 연산 규칙과 상태 전이를 핵심으로 하는 결정론적 알고리즘에는 부적합하다[4][5]. 계획 문제에서 설명의 대상은 모델의 가중치가 아니라, 결과를 산출한 연산 과정 그 자체여야 하기 때문이다.

본 논문은 이러한 문제의식을 바탕으로, 결정론적 계획 알고리즘을 위한 신뢰 가능한 LLM 설명 생성의 핵심 요건으로 Trace 정렬(Trace alignment) 개념을 제안한다. 여기서 Trace 정렬이란 LLM 이 생성한 자연어 설명의 각 구성 요소가 실제 알고리즘의 실행 Trace 단계와 정확히 대응되는 정도를 의미한다. 이는 설명이 실제 수행된 연산, 상태 전이, 의사결정 과정을 얼마나 충실히 반영하는지를 측정하는 척도가 된다. 본 연구에서는 설명을 단순한 사후적 해석이 아닌, **Trace 와 정렬된 자연어 변환**으로 재정의함으로써 설명의 신뢰성과 검증 가능성을 확보하고자 한다.

본 논문의 기여는 다음과 같다. (1) 결정론적 계획 알고리즘에서 발생하는 LLM 설명의 불일치 문제를 Trace 관점에서 정식화하고, (2) Trace 정렬에 기반한 신뢰성 있는 설명 생성 방법론을 제안하며, (3) 해당 접근법이 다양한 계획 문제에 적용 가능함을 논의한다.

Ⅱ. 본론

2.1 결정론적 계획 알고리즘과 설명 불일치 문제

결정론적 계획 알고리즘은 입력에 대해 유일한 출력을 보장하는 특성을 갖는다[2][3]. 이러한 알고리즘은 명시적인 규칙과 상태 전이 과정을 따르며, 전체 연산 과정은 하나의 인과적 사슬(Causal Chain)을 형성하여 최종 결과를 도출한다.

그러나 기존 LLM 기반 설명 생성 방식은 이러한 연산 구조를 충분히 반영하지 못한다. 대부분의 접근법은 최종 출력이나 일부 요약 정보만을 입력으로 받아 내부 연산 과정과 무관한 설명을 생성한다. 그 결과, 설명에 포함된 시간적 순서나 인과 관계가 실제 연산 흐름과 불일치하는 환각(Hallucination) 문제가 발생한다. 이러한 설명은 표면적으로는 설득력 있어 보일 수 있으나, 실제 의사결정의 검증이나 수정에는 실질적인 도움을 주지 못한다[1].

특히 계획 알고리즘이 핵심적인 의사결정을 지원하는 산업 현장에서는, 연산 과정과 일치하지 않는 설명이 결과에 대한 오해를 초래할 위험이 크다. 따라서

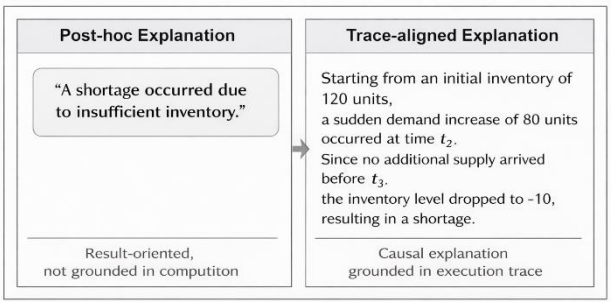
결정론적 계획 문제에서 설명 생성의 핵심은 표현의 유창성이 아니라, 설명이 실제 연산 과정과 얼마나 정확히 일치하는가, 즉 충실성(faithfulness)에 있다.

2.2 Trace 개념과 설명의 정렬 문제

본 연구에서는 설명 불일치의 근본 원인을 Trace 미정렬(Misalignment)로 정의한다. Trace 란 계획 알고리즘 실행 시 발생하는 모든 상태 변화와 연산 단계를 순차적으로 기록한 데이터로, 알고리즘의 실행 과정을 온전히 재현할 수 있는 정보를 담고 있다. 본 연구에서 Trace 는 연산 단계의 시퀀스로 구성되며, 각 단계는 상태 값, 적용된 연산 규칙, 그리고 그 결과를 포함한다.

기존 LLM 방식은 Trace 를 명시적으로 활용하지 않아 설명과 실제 연산 과정 간의 대응 관계를 보장하지 못한다[6]. 반면, 신뢰성 있는 설명을 제공하기 위해서는 설명 내 각 정보가 Trace 의 특정 연산 단계에 명확히 매핑되어야 한다.

이에 본 연구는 설명을 Trace 에 대한 사후적 해석이 아닌, **Trace 의 자연어 변환(Translation)**으로 정의한다. 즉, 설명은 새로운 정보를 추론해내는 것이 아니라, 이미 존재하는 Trace 정보를 언어적으로 정확히 서술하는 결과물이어야 한다. [그림 1]은 기존의 사후적 설명 방식과 본 연구가 제안하는 Trace 정렬 기반 방식의 차이를 보여준다.



[그림 1] 기존 사후적 설명 방식과 Trace-정렬 기반 설명 방식의 비교.

2.3 Trace-정렬 기반 신뢰 가능한 설명 생성

제안하는 Trace 정렬 기반 접근법은 계획 알고리즘의 실행 과정과 LLM 기반 설명 생성을 구조적으로 결합함으로써, 설명의 신뢰성과 검증 가능성을 확보하는 것을 목표로 한다. 본 연구에서 Trace 정렬은 설명 생성의 목표이자 동시에 생성 과정에 적용되는 제약 조건으로 작용한다. 즉, LLM 이 생성하는 모든 설명은 실행 Trace 에 포함된 연산 단계와 논리적으로 정렬되어야 하며, 이를 벗어나는 정보는 허용되지 않는다. 설명 생성 프로세스는 다음의 세 단계로 구성된다.

1) 결정론적 Trace 생성 및 추출

알고리즘의 최종 결과뿐만 아니라, 연산 과정에서 발생하는 모든 중간 상태와 단계를 포함하는 실행 Trace T를 명시적으로 생성한다. 이 Trace 는 알고리즘의 계산 흐름을 완전하게 반영하며, 설명 생성의 객관적 근거(Ground Truth)로 활용된다.

2) 구조화된 근거 추출

생성된 Trace T로부터 설명에 필요한 핵심 상태 변화와 결정 요인을 추출하여, 구조화된 근거 집합(Evidence Set)으로 변환한다. 이 집합은 특정 시점의 상태 값, 연산 결과, 인과적 전이 관계 등을 포함하며, 설명에 사용될 정보의 범위를 한정한다.

3) 정렬 및 충실성 제약

설명 생성 시 추출된 근거 집합을 벗어나는 정보가 포함되지 않도록 엄격한 제약(Constraint)을 적용한다. 설명에 포함된 모든 정보 (날짜, 수치, 인과 관계 등)는 근거 집합에 대해 충실성(Faithfulness)을 유지해야 하며, 인과적 서술은 Trace 상의 실제 연산 단계와 논리적으로 정렬(Alignment)되어야 한다. 이러한 제약을 위반하는 정보는 생성 과정에서 배제된다.

이러한 메커니즘을 통해 LLM 은 임의의 추론이나 환각 없이, 결정론적 계획 알고리즘의 연산 논리를 실무자에게 정확하게 전달한다.

Ⅲ. 결론

본 논문은 결정론적 계획 알고리즘을 대상으로 한 LLM 기반 설명 생성의 한계를 분석하고, 신뢰성 있는 설명을 위한 Trace 정렬 기반 접근을 제안하였다. 기존 방식은 최종 출력에 의존한 사후적 설명에 머물러, 실제 연산 과정과의 정합성을 보장하지 못하는 구조적 한계를 지닌다. 이에 본 연구에서는 설명을 알고리즘 실행 Trace 의 자연어 표현으로 재정의하고, 설명 생성의 핵심을 연산 과정과의 정렬 문제로 전환하였다.

제안한 방법은 LLM 의 환각을 효과적으로 억제함으로써 설명의 신뢰성과 검증 가능성을 동시에 확보하며, 특정 도메인에 국한되지 않고 다양한 결정론적 계획 문제에 적용 가능하다. 향후 연구에서는 실제 계획 문제에 대한 적용 및 정량적 평가를 수행하고, Trace 정렬 과정의 자동화와 확장성 확보 방안을 모색할 예정이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터지원사업의 연구결과로 수행되었음 (IITP-2026-RS-2021-II211835) 그리고 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (RS-2025-22932973)

참 고 문 헌

[1] Wei, J., X. Wang, D. Schuurmans, et al., "Chain-of-thought prompting elicits reasoning in large language models," Advances in Neural Information Processing Systems, vol. 35, pp. 24824-24837, 2022.

[2] Hillier, F. S., and G. J. Lieberman, "Introduction to Operations Research,"McGraw-Hill Science Engineering, 9th ed., Feb. 2009.

[3] Pinedo, M., "Scheduling: Theory, Algorithms, and Systems," Springer, 5th ed., 2016.

[4] Lundberg, S. M., and S.-I. Lee, "A unified approach to interpreting model predictions,"Advances in Neural Information Processing Systems, vol. 30, 2017.

[5] Ribeiro, M. T., S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016.

[6] Jacovi, A., and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" arXiv preprint, arXiv:2004.03685, 2020.